

The Word made Welsh

The semiotics of code-switching and borrowing
in a minority language

Bethan Siân Thomas Tovey-Walsh

Submitted to Swansea University in fulfilment of the requirements
for the Degree of Doctor of Philosophy

Swansea University

2024

Abstract

This study examines the previously little-studied phenomenon of orthographic adaptation of English-origin lexical items in Welsh-language discourse. Using the semiotic framework of enregisterment (Agha, [2005](#)), orthographic practices themselves, as well as orthographically adapted lexemes and texts, are situated as highly salient ways of performing Welshness. Welshness is examined as a cultural, political, and geographical identity as well as a linguistic one, and the interaction of language skills with cultural authenticity is explored, leading to the conclusion that prescriptivist ideals and fears of language loss drive the construction of artificial and emotionally unsatisfying registers of Welsh. An examination of Welsh-English bilingualism on Twitter through the lens of enregistered orthography shows Welsh linguistic identity to be a complex and expansive semiotic construct. In conclusion, I propose that both scholarly and pedagogical approaches to Welsh would benefit from a deeper understanding of the enregisterment of Welsh orthography, viewing multiple layers of language-mixing as a variable feature of Welsh stylistic richness, rather than treating it as an exclusively transgressive practice.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed Bethan Siân Tovey-Walsh..... (candidate)

Date 11 January 2024.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed Bethan Siân Tovey-Walsh..... (candidate)

Date 11 January 2024.....

Contents

Acknowledgements	xiii
List of Figures	xv
Terminology and conventions	xix
Citation and transcription conventions	xix
Other-language items	xix
Languages and their users	xx
Lexical sets	xxii
English vowels	xxii
Welsh vowels	xxiii
Glossary	xxv
Abbreviations	xxv
Initialisms and acronyms	xxv
Welsh-language terms	xxvi
Language(s), language varieties, and their users	xxix
Linguistic units, systems, processes, and products	xxxiv
Introduction: “English words, either pure or Cymricised”	1
“A great and growing ambition for the Welsh language”	5
Language as a medium for education	7

The dangers of monolingualism	9
Seeking an alternative: plurilingual education	11
Establishing research questions	13
Research questions	14
Primary question	14
Secondary questions	15
Tertiary questions	15
Chapter outlines	15
1. “When the foreign words are camouflaged in plain type”: Code-switching, borrowing, and other mixed languagings	15
2. “What’s ‘englyn’ when it’s english?”: A semiotic account of languageness	15
3. “Do I not love my country because I haven’t properly learnt mutations?”: The Welsh language and social constructions of Welshness	16
4. “Rough as Bwlchardudwy”: Orthography and the foreignness of Cymraeg	16
5. “For the benefit of Welsh people who had never seen a letter X before”: Designing a study of orthographic variation in the adaptation of English-origin lexical units	17
6. “Ai hop ddat yew can ryd ddys”: Data analysis and discussion	17
7. “TIREZ LE FFYCIN RIDEAU”: Conclusions	17
1 “When the foreign words are camouflaged in plain type”:	
Code-switching, borrowing, and other mixed languagings	18
1.i When is a lexical unit used by Cymrophones “of English origin”?	18
1.ii What is code-switching?	22
1.ii.a Frameworks and formulae: The mechanics of code-switching	22

1.ii.b	Processing and proficiency: psycholinguistic pressures	26
1.ii.c	Sentiments and swearwords: Emotional pressures	27
1.ii.d	Society and status: Sociocultural pressures	29
1.ii.e	Conversation and communication: pragmatic and discursive pressures	30
1.iii	What is borrowing?	31
1.iii.a	Alteration and assimilation: mitigating the (in)congruity of borrowed words	31
1.iii.b	Superfluity and supplementation: the (in)dispensability of borrowed items	35
1.iii.c	Purism and power: the sociolinguistics of borrowing	36
1.iv	Code-switching and borrowing: a comparison	38
1.iv.a	Flagging and fluency: individual and sociocultural distinctions	38
1.iv.b	Adaptation and acculturation: differences in integration	39
1.iv.c	Semantics and synonyms: functional distinctions	40
1.iv.d	Documentation and dissemination: frequency, widespreadness, and listedness as distinguishing criteria	41
1.v	“New language practices”?: Language, languaging, and the translanguaging turn	42
1.vi	Code-switching and borrowing: the Welsh perspective	47
1.vi.a	Provincialism and pilfering: attitudes from outside the Cymrophone community	49
1.vi.b	Servility and snobbery: attitudes within the Cymrophone community	52
1.vi.c	Saesneg and subordination: the denigration of Welsh English	56
1.vi.d	Innocence and insularity: the moral superiority of Cymrophony	57
1.vi.e	Partition and proscription: the (im)possibility of bilingualism	59

1.vii	Why does the status of ELUs in Welsh matter?	60
2	“What’s ‘englyn’ when it’s english?”: A semiotic account of languageness	64
2.i	Languageness in the context of enregisterment	64
2.i.a	Enregisterment: ascribing social meaning to linguistic features .	66
2.ii	Language as a connotative semiotic	68
2.ii.a	Layers and layouts: The stratified nature of the linguistic sign .	69
2.ii.b	Content and construal: how substantial is substance?	74
2.ii.c	Context and connotation: enregistering languages	77
2.ii.d	Reciting and writing: phonology, orthography, and the double articulation	78
2.iii	Language-mixing and enregisterment in Welsh	84
2.iii.a	Formality and function: enregistering Welsh for education . . .	85
2.iii.b	Pedantry and plausibility: technical vocabulary and terms of art	86
2.iii.b.1	Protocol and perplexity: the enregisterment of English-origin lexical units as “bad Welsh”	93
3	“Do I not love my country because I haven’t properly learnt mutations?”: The uneasy relationship of Welshness and Cymrophony	101
3.i	“Nehemiah of the Vale of Clwyd”: the <i>Fro Gymraeg</i> and the languageness of the land	102
3.i.a	Blue Books and barbarism: constructing the gwerin as a Cymrophone archetype	104
3.i.b	Clans and councillors: the ethnic/civic divide in modern Welsh political thought	106
3.ii	Who speaks Welsh? Sociolinguistic and raciolinguistic ideologies of Welshness	109

3.ii.a	Welshness and whiteness: raciolinguistic constructions of Cymrophony	110
3.ii.b	Chauvinism and sharing: the competition for authentic Welshness	113
3.iii	“I dunno if she was Welsh in the sense of a Welshwoman”: the porous semantics of Welshness and Cymrophony	116
3.iii.a	Cymraeg and Cymreig: Welsh semantics and the semantics of “Welsh”	116
3.iii.b	Lack and legitimacy: the semantics of (non-)Cymrophony	121
3.iv	“You’ll never be a member of our exclusive club, you second-class, linguistic failure”: purism, prestige, and power in the Welsh sociolinguistic context	125
3.iv.a	Register and remonstrance: linguistic prescriptivism and language confidence	125
3.iv.b	Hierarchy and heritage: Cymrophone notions of cultural prestige	130
3.v	Implications: “Bratiaith Guilt” and language ideologies in the development of Welsh	134
4	“Rough as Bwlchardudwy”: Orthography and the foreignness of Cymraeg	138
4.i	Welsh orthography and its relationship to Welsh and English phonology	144
4.i.a	Syllables and stress: Diacritics in Welsh orthography	146
4.i.b	Print and pronunciation: Welsh grapheme-phoneme correspondences	149
4.i.c	Script and strangeness: orthographic conventions for ELUs in Welsh	150
4.i.c.1	Consonants in Welsh ILUs	151

4.i.c.2	Comparing the vowels of Welsh and English	152
4.i.d	Schwa and «siwgar»: cowel reduction and vowel epenthesis in Welsh and Welsh English	155
4.ii	Orthographic adaptation of ELUs: a review of the literature	157
4.iii	Orthography as a locus of cultural expression	159
4.iii.a	Ghoti and foreignness: orthography and its relationship to phonology	159
4.iii.b	Haitch and humour: orthography as cultural practice	162
4.iii.c	Words and Way-uls: The orthographic enregisterment of Welsh English	169
4.iii.d	Gaelic and gobbledigook: the Celtic languages under the Anglophone gaze	170
5	“For the benefit of Welsh people who had never seen a letter X before”: Designing a study of variation in the orthographic adaptation of English-origin lexical units	175
5.i	Introduction	175
5.i.a	Sources of data for orthographic adaptation	175
5.ii	The orthographic adaptation of established loanwords: building the GPC loanword dataset	176
5.ii.a	Identifying loanwords	177
5.ii.b	Adding phonetic information	177
5.iii	A survey of orthographic adaptation practices	179
5.iii.a	Survey respondents’ language use and background	179
5.iii.a.1	Survey loanword prompts	182
5.iii.a.2	Data collection and filtering	183
5.iv	Orthographic adaptation in naturalistic data: building the Twitter corpus	185
5.iv.a	Ethical considerations	185

5.iv.a.1	Access to the dataset	186
5.iv.b	Twitter corpus collection procedure	186
5.iv.b.1	Pilot study: setup	187
5.iv.b.2	Cleaning the dataset	189
5.iv.b.3	Selecting a balanced list of Twitter accounts	192
5.iv.b.4	Selecting lexemes	196
5.v	Limitations in the data	200
5.vi	Data analysis procedure	201
5.vi.a	Quantitative methods	202
5.vii	Analytic categories	204
5.vii.a	Objects of analysis: lexemes, variants, graphemes, and phonemes	204
5.vii.b	Quantitative categories of metadata	206
5.vii.b.1	Listedness: background	206
5.vii.b.2	Listedness of variants	211
5.vii.b.3	Age of attestation, frequency, and widespreadness of lexemes and variants	212
5.vii.b.4	Standardness of orthographic adaptations	213
6	“Ai hop ddat yew can ryd ddys”: Data analysis and discussion	215
6.i	Dataset descriptions	215
6.i.a	Distributional patterns of variants and graphemes across datasets	215
6.i.b	The GPC dataset	220
6.i.c	The survey of orthographic adaptation	221
6.i.d	The Twitter corpus	223
6.i.d.1	The Twitter subcorpora and register variation	223
6.i.d.2	Twitter lexemes: summary statistics	229
6.ii	Quantitative analysis of diacritic use	231
6.iii	Quantitative analysis of lexemes and variants	238

6.iii.a	Listedness, semantics and synonymy	238
6.iii.a.1	Discussion	240
6.iii.b	Listedness and formality markers: treiglâd and morphological inflection	242
6.iii.b.1	Discussion	245
6.iii.c	Conclusions	246
6.iv	Qualitative analysis	247
6.iv.a	Overview	248
6.iv.b	Unadapted variants	250
6.iv.c	Adaptation and code-switching	252
6.iv.d	Performative adaptation	255
6.iv.d.1	Proper nouns	255
6.iv.d.2	Performing Welshness: Welsh and Welsh-English accents in orthographic adaptations	260
7	“TIREZ LE FFYCIN RIDEAU”: Conclusions	264
7.i	Regina progenitrix: the contextual power of languageness	266
7.ii	Primary research question: orthographic variation and semiotic meaning	272
7.iii	Secondary research questions: orthography, enregisterment and languageness	273
7.iv	Tertiary research questions: scholarly and pedagogical implications . .	273
7.iv.a	Research Question 4	273
7.iv.b	Research Question 5	275
7.iv.b.1	Language policy: from monolingualism to plurilingualism	277
7.iv.b.2	Plurilingual strategies for educational policy and practice	279
7.iv.b.3	Plurilingual parents, teachers, and institutions	281

7.iv.b.4	Plurilingualism in academic language	284
7.iv.b.5	Plurilingualism in standardized assessment	287
7.iv.b.6	Plurilingual learning materials	290
7.iv.b.7	Plurilingual curricula and classrooms	292
7.iv.b.8	Plurilingualism for grown-ups: adult education	300
7.v	Conclusions	301
7.v.a	Education and the right to a Welsh identity	302
7.v.b	Iaith Gymreig, iaith Gymraeg: the Welsh-languageness of the English language	304
Bibliography		306
Primary sources		306
Fiction and poetry		306
Popular non-fiction		306
Journalism, essays, and opinion pieces		306
Websites		311
Official reports and records		312
Source code and electronic corpora		315
Language and reference		316
Reference works		316
Grammar and language-teaching materials		317
Secondary sources		317
Appendices		355
Appendix A Transcription conventions		355
A.i	The vowels of English	355
A.ii	The vowels of Welsh	360
A.iii	Lexical sets	363

A.iii.a English vowels	363
A.iii.b Welsh vowels	364
Appendix B Additional postcards	366
Appendix C Comparison of linguistic features in GPC, Survey, and	
Twitter datasets	379
C.i Distribution of orthographic variants for vowel and consonant	
phonemes	379
C.ii Patterns of diacritic use in the representation of vowel sounds	384
C.iii Phoneme–grapheme distributions	387
Appendix D Loanwords in Geiriadur Prifysgol Cymru	487
D.i GPC: summary statistics	487
Appendix E Survey of orthographic adaptation	490
E.i Survey: text	490
E.ii Survey: text (English translation)	511
E.iii Survey: prompt list	518
E.iv Survey: variant forms	520
E.v Survey: demographic and Welsh usage graphs	522
E.vi Survey: summary statistics	529
E.vii Survey: variant forms	532
Appendix F Corpus of Welsh-language tweets	560
F.i Twitter: summary statistics	560
F.ii Twitter: lexeme frequency graphs	569
F.iii Twitter: comparing linguistic features of ELUs and ILUs in the	
personal and official subcorpora	574
F.iv Twitter: list of lexemes selected for close study	581

F.v	Twitter: variant forms	585
F.vi	Twitter: variants of interest after PCA	630



For my siblings.

गौरव के लिए, जो मुझे दुनिया की सबसे बुद्धिमान बड़ी बहन जैसा महसूस कराता है |

Aan Fientje, die me steeds het gevoel geeft dat ik
de meest wijze grote zus ter wereld ben.



And for Michael.
Olórin, Namárië.



Acknowledgements

I am indebted to the *CorCenCC* project for funding this study, and to Swansea University. Without the consistent support of the *CorCenCC* Team and University staff, this project could never have been completed.

My supervisors, Tess Fitzpatrick, Rhys Jones, and Steve Morris, have been endlessly encouraging and patient (and I have given them endless cause to be so). My profound gratitude to them.

Without the thought-provoking questions and suggestions offered by examiners Cornelia Tschichold and David Willis, this thesis would have been much poorer. I am thankful to them for their guidance, and (along with the chair, Julian Preece) for an enjoyable and stimulating viva.

Many people helped me with a variety of questions. I am particularly indebted to Catrin Alun and Heini Gruffudd for their help with some cloddy Victorian texts, and to Meinir Jones for information about current practice in the teaching of Welsh. Lillebeth Larun's perceptive advice was comforting and encouraging in equal measure; *tusen takk!*

Naturally, all errors and infelicities in this work remain my own.

Dawn Knight, Laura Arman, Jenny Needs, Lawrence Antony, and others involved in the *CorCenCC* project have been stimulating and invigorating companions on the PhD journey. The Applied Linguists of Swansea have also provided refreshing companionship on many occasions.

Further afield, but no less important, the linguistics community on social media has helped track down references, answered strange and silly questions, and been a source of much-needed community. Maggie Glass, my fellow Celticist, taught me that sometimes online friends are the realest friends.

My other intellectual community, the XMLers, are due boundless thanks. There are too many of them to name individually, but I must offer special thanks to a few people: Allen Renear and Steve de Rose, for taking my flights of intellectual whimsy as seriously as I take theirs; Debbie Lockett, Mary Holstege, Lauren Wood, Debbie

Lapeyre, and Tommie Usdin, not *just* for being my tech sisters (but I *do* like having tech sisters); Syd and all the Baumen for being a constant loveliness; Jo Rabin, Peter Flynn, and Nic Gibson for being equally punk in very different ways; Claus Huitfeldt, for his generosity and good humour; David Birnbaum, for his mastery of so many things.

Michael Sperberg-McQueen has not only been a remarkable rubber duck, but also a mentor, an intellectual inspiration, a pain in the semiotic ass, a light in the darkness, and a wonderful friend.

My former colleagues in Oxford must take a share of the blame/praise for all of the *gestures wildly*. Particular gratitudes to Edmund Weiner, Anthony Esposito, Emma Lenz and Mark Dunn, Fiona MacPherson, Alois Heuboeck, Jon Dent, Tressy Arts, and Simon Horobin, for encouraging the linguisticness.

So many friends have helped me through these strange years. My lifelong bonus family, the Fabriciuses (and Rowlands), without whom so many things would not have been; my more recent bonus family, the Barnett-Neefsjes; my beast fiend, Julia Cordey; Robin Hall, who has plied me with interesting Swiss orthographies; Paco and Peter, Helen and Steve, who have always made Swansea feel like home; Gary and Pietro, who did the same for Oxford; Haf Elgar, Rhianon Darwall, and Caradog Williams, who could probably make Slough feel like home if I had them with me.

I have been blessed throughout my life with good teachers, whose wisdom stays with me: Gwen and Eric Jones; Linden Evans; Jeffrey Connick; Siôn Woods; Heddyr Gregory; Anna-Vivian Jones; Mike Stevens; Fiona Stafford; Phil West; Malcolm Godden; Mishtooni Bose; Benjamin Thompson; and Bernard Silverman.

The NHS in Swansea looked after me when I did an unexpected neurological; particular thanks to Dr. Hrastelj for his patience with my inevitable orthographic fascination. They also looked after my mother and father at different times since I started the PhD. Without the care, professionalism, and compassion of the staff at Morriston Hospital, I might have given up on this project at multiple points.

My family. Mami, without whom I too might have been one of the “Welshless Welsh”. My dad, who really didn’t have to be quite that dramatic just to get me a PhD extension. My aunts, uncles, and cousins, who are basically a big ball of embodied love.

My husband. “Yr pann yth weleis gyntaf a gereis yn vwyhaf gwr”. Norm may be a theoretical abstraction for Hjelmslev (1961), but he’s everything substantial for Tovey-Walsh (2024).

List of Figures

Figure 1.1	Tweet from the official account of The Spectator	51
Figure 2.1	: A visualization of Hjelmslev’s sign.	70
Figure 2.2	: Grille cypher as a metaphor for Hjelmslev’s “purport”	71
Figure 2.3	: Grille cypher: schema, and substance.	72
Figure 2.4	: Grille cypher: alternative schemata and substances	73
Figure 2.5	Relationship between Saussurean and systemic–functional strata	76
Figure 2.6	The structure of the sign	76
Figure 2.7	Language as a connotative semiotic	83
Figure 2.8	: <i>GPC Online</i> entry for the word “bold”.	93
Figure 3.1	: Plaque above the doorway of Bethel Baptist Chapel in Maesteg.	124
Figure 4.1	: Postcard, postmarked 1905.	139
Figure 4.2	: Postcard, postmarked ?1926.	140
Figure 4.3	: Postcard, postmarked 1908.	141
Figure 4.4	: Postcard, no postmark.	142
Figure 4.5	: The monophthongs of Welsh and English.	153
Figure 4.6	: The diphthongs of Welsh and English.	154
Figure 4.7	Animal Farm playscript	166
Figure 4.8	Animal Farm (TellTale Theatre) poster	167
Figure 4.9	Animal Farm (Blake School) poster	168
Figure 4.10	Corcoran’s Irish Pub sign	172
Figure 4.11	Irish Times bar sign	172
Figure 4.12	Scholars Lounge Irish Pub sign	173
Figure 4.13	Irish Pub sign	173
Figure 5.1	: Token distributions in official and personal Twitter subcorpora	197
Figure 6.1	Proportion of lexemes with variant counts	216

Figure 7.1	“God Sêf Ddy Chwyn” (Hedd, 2023).	267
Figure 7.2	“God Save the Queen” (Reid, 1977)	270
Figure C.1	Number of orth. variants for vowels	380
Figure C.2	Std. dev. of orth. variants for vowels	381
Figure C.3	Number of orth. variants for consonants	382
Figure C.4	Std. dev. of orth. variants for vowels	383
Figure C.5	: Stacked barcharts showing the proportion of accented characters representing different vowel types in this study’s datasets.	385
Figure C.5	: Stacked barcharts showing the proportion of accented characters (cont.).	386
Figure C.6	: Distribution of graphemes for phoneme /j/	389
Figure C.7	: Distribution of graphemes for phoneme /ɪ/ (KIT)	391
Figure C.8	: Distribution of graphemes for phoneme /ɛ/ (DRESS)	393
Figure C.9	: Distribution of graphemes for phoneme /æ/ (TRAP)	395
Figure C.10	: Distribution of graphemes for phoneme /ɔ/ (LOT~WANT)	397
Figure C.11	: Distribution of graphemes for phoneme /ʌ/ (STRUT)	399
Figure C.12	: Distribution of graphemes for phoneme /ʊ/ (FOOT)	401
Figure C.13	: Distribution of graphemes for phoneme /i:/ (FLEECE)	403
Figure C.14	: Distribution of graphemes for phoneme /eɪ/ (FACE~BAIT)	405
Figure C.15	: Distribution of graphemes for phoneme /aɪ/ (PRICE)	407
Figure C.16	: Distribution of graphemes for phoneme /ɔɪ/ (CHOICE)	409
Figure C.17	: Distribution of graphemes for phoneme /aʊ/ (MOUTH)	411
Figure C.18	: Distribution of graphemes for phoneme /əʊ/ (GOAT~KNOW)	413
Figure C.19	: Distribution of graphemes for phoneme /ʊʊ/ (GOOSE~FLEW)	415
Figure C.20	: Distribution of graphemes for phoneme /w/	417
Figure C.21	: Distribution of graphemes for phoneme /jʊʊ/ (TUNE)	419
Figure C.22	: Distribution of graphemes for phoneme /jʊə/ (CURE)	421
Figure C.23	: Distribution of graphemes for phoneme /ɪə/ (NEAR~BEER)	423
Figure C.24	: Distribution of graphemes for phoneme /ɛə/ (SQUARE~LAIR)	425
Figure C.25	: Distribution of graphemes for phoneme /aə/ (FIRE)	427
Figure C.26	: Distribution of graphemes for phoneme /tə/ (TOWER)	429
Figure C.27	: Distribution of graphemes for phoneme /ʊə/ (POOR~BREWER)	431
Figure C.28	: Distribution of graphemes for phoneme /ɑ:/ (BATH~PALM~START)	433

Figure C.29 : Distribution of graphemes for phoneme /o:/ (THOUGHT~NORTH~FORCE)	435
Figure C.30 : Distribution of graphemes for phoneme /ɜ:/ (NURSE)	437
Figure C.31 : Distribution of graphemes for phoneme /ə/ (COMMA)	439
Figure C.32 : Distribution of graphemes for phoneme /ɜ:/ (LETTER)	441
Figure C.33 : Distribution of graphemes for phoneme /p/	443
Figure C.34 : Distribution of graphemes for phoneme /b/	445
Figure C.35 : Distribution of graphemes for phoneme /m/	447
Figure C.36 : Distribution of graphemes for phoneme /f/	449
Figure C.37 : Distribution of graphemes for phoneme /v/	451
Figure C.38 : Distribution of graphemes for phoneme /t/	453
Figure C.39 : Distribution of graphemes for phoneme /d/	455
Figure C.40 : Distribution of graphemes for phoneme /θ/	457
Figure C.41 : Distribution of graphemes for phoneme /ð/	459
Figure C.42 : Distribution of graphemes for phoneme /l/	461
Figure C.43 : Distribution of graphemes for phoneme /n/	463
Figure C.44 : Distribution of graphemes for phoneme /s/	465
Figure C.45 : Distribution of graphemes for phoneme /z/	467
Figure C.46 : Distribution of graphemes for phoneme /tʃ/	469
Figure C.47 : Distribution of graphemes for phoneme /dʒ/	471
Figure C.48 : Distribution of graphemes for phoneme /ʃ/	473
Figure C.49 : Distribution of graphemes for phoneme /ʒ/	475
Figure C.50 : Distribution of graphemes for phoneme /ɹ/	477
Figure C.51 : Distribution of graphemes for phoneme /k/	479
Figure C.52 : Distribution of graphemes for phoneme /g/	481
Figure C.53 : Distribution of graphemes for phoneme /ŋ/	483
Figure C.54 : Distribution of graphemes for phoneme /h/	485
Figure E.1 : Geographic distribution of survey respondents.	522
Figure E.2 : Survey respondents' age at time of survey.	523
Figure E.3 : Survey respondents' age of acquisition of Welsh.	523
Figure E.4 : Survey respondents' Welsh-medium primary and secondary education.	524
Figure E.5 : Survey respondents' reported home language(s) in childhood.	524
Figure E.6 : Survey respondents' reported current home language(s).	525
Figure E.7 : Survey respondents' reported use of Welsh on social media.	525

Figure E.8 : Survey respondents' reported social use of spoken Welsh. . . .	526
Figure E.9 : Survey respondents' reported professional use of spoken Welsh.	526
Figure E.10 : Survey respondents' reported informal use of written Welsh. .	527
Figure E.11 : Survey respondents' reported formal use of written Welsh. . .	527
Figure E.12 : Survey respondents' reported total use of Welsh.	528
Figure F.1 : Lexeme frequency by semantic category in the Twitter corpus and subcorpora	570
Figure F.2 : Lexeme listedness in GPC by semantic category in the Twitter corpus and subcorpora	571
Figure F.3 : Lexeme listedness in six dictionaries by semantic category in the Twitter corpus and subcorpora	572
Figure F.4 : ELU frequency and listedness by synonym status.	573
Figure F.5 : Average number of emoji per tweet in the official and personal Twitter subcorpora.	574
Figure F.6 : Rates of English and Welsh inflectional morphemes on ELUs and ILUs by listedness status.	575
Figure F.7 : Rates of inflection on lexemes by listedness status in six dictionaries.	576
Figure F.8 : Rates of inflection on variants by listedness in GPC.	577
Figure F.9 : Rates of inflection on variants by listedness in seven dictionaries.	578
Figure F.10 : Rates of treigladd on lexemes by listedness status.	579
Figure F.11 : Rates of treigladd on variants by listedness status.	579
Figure F.12 : Comparisons of inflection and treigladd of ELUs and ILUs, separated by Twitter subcorpus.	580

Terminology and conventions

Citation and transcription conventions

Conventions for representing the sounds of English and Welsh are explained in Appendix A, including lists of vowel set labels (based on those of Wells (1982a, pp. xviii–xiv)) which will be used to refer to vowel phonemes in Welsh and in English.

Quotations from Twitter will not be attributed to specific accounts, except in cases where the author is a public figure or institution, in line with guidelines for ethical use of social media data (discussed in Chapter 5.iv.a).

Although it generally seems preferable to use terms which will be familiar to readers, in a few important areas I have chosen to adopt non-conventional terms because the standard options are in some sense inadequate or inappropriate in the context of this study. Welsh-language terms are also used throughout the study. Those which figure prominently in the main text are defined in the Glossary.

Other-language items

The main focus of this study will be on lexical units of English origin in Cymrophone discourse, and it will be necessary to discuss these units in as neutral a way as possible. This means avoiding making any explicit or implicit claims as to their integration into the Welsh lexicon: a single term is needed which covers both English-origin words which are accepted as “Welsh” in Cymrophone discourse and words which stand out as “English”. This term should be applicable to any word whose root morpheme is not derived from a Welsh (including proto-Welsh or Brythonic) etymon. Words whose origin is in some language other than Welsh will therefore be called “extrinsic lexical units” (ELUs). These are lexical units whose root element(s) developed from etymological resources extrinsic to the language itself. Words which can be traced back to Welsh or one of its ancestral languages will be

called “intrinsic lexical units” (ILUs).

This study is concerned in part with how and why extrinsic lexical units come to be included in the Welsh lexicon, or excluded from it. As we shall see, this implicates a variety of linguistic, sociocultural, and individual factors. A given word with an English etymon may be considered “Welsh” by some users of the language, and “English” by others. Whatever its status, however, the word remains an ELU because its etymon is extrinsic to the Welsh language.

Having coined these terms, the simplest way to designate the language which is the medium of communication in a given linguistic act is to call it the “intrinsic language”, or IL. In this study, Welsh will usually be the IL under discussion, but the term could be applied to any language when its contact with other languages is examined.

Unlike terms such as “foreign language” and “non-native language”, the term “extrinsic language” makes no assumptions about the cultural status of its referent. It refers only to the status of a language as taxonomically separate from the IL. Of course, what constitutes a taxonomically distinct language is not always easy to decide (Tamburelli, 2021a, 2021b; Tosco, 2021); however Welsh and English have a documented history of separate historical development before being brought into close contact, and remain mutually unintelligible. The status of English as an EL relative to Welsh, and vice versa, will therefore be taken for granted for the purposes of this study.

Languages and their users

I have opted for the term *user* rather than *speaker* of a language because it is a more natural term to use when discussing written forms of language. Since much of the discussion here will centre on a corpus of Twitter messages, it will at times be necessary to speak of the behaviour patterns of writers and speakers of a language distinctly. Selecting *user* as the neutral term frees the term *speaker* specifically to discuss oral use of a language.

Terms traditionally used to describe an individual’s earliest-learned or most-used language, such as *native language* and *first language* are somewhat unsatisfactory in the context of Welsh-English bilingualism. They assume monolingualism as the norm, eliding questions such as which language is the “first” of a concurrent bilingual. To avoid this, I will use the terms *primary language* and *additional language*. In adults, a primary language is a language whose acquisition was begun in early childhood, and

in which the user has developed an age-appropriate level of communicative competence. An additional language is one learned in later childhood or adulthood (whether naturalistically or through formal education); or a language whose acquisition was begun in earlier childhood, but which was either then abandoned (perhaps because the family moved to another country, or because of loss of access to speakers of the language), or was never acquired to a standard adult level (for example a heritage language, heard occasionally from family or community members, but without the consistent exposure required for developing full competence).

These categories are intentionally very broad, and somewhat subjective in their criteria. Primary languages are not exclusive; an individual who is raised multilingually may acquire multiple primary languages. Users would generally be expected to develop an advanced level of communicative competence in a primary language, but no assumptions are made about which variety of the language they acquire: it may be a standard variety, a regional dialect, or both. The acquisition of literacy is not an essential criterion for considering a language to be primary. A child brought up speaking one language at home, but schooled in a different language (particularly one with a different writing system), need not learn to read the home language for it to be considered one of her primary languages.

Throughout this study, I will generally refer to bilingualism, and to the use of two languages together. Nonetheless, much of what is said may also be true of multilingual switching between three or more languages. I will explicitly distinguish bilinguals from multilinguals only when differences between the two are relevant. The discussion applies specifically to bilingualism in spoken languages; cross-modal bilingualism (in a spoken and a sign language) and bilingualism in two sign languages are beyond the scope of this study.

Lexical sets

Lexical set labels are used throughout this study (following Wells (1982a, pp. xviii–xiv)) to refer to the vowels of English and of Welsh. A full discussion of these sets, with motivations for changes to Wells’ original English list, is provided in Appendix A. Here, the lists are given in simple form for ease of reference. English sets are listed with IPA symbols labelling standard pronunciations in both Prestige Southern-England English (PSEE) and Welsh English (WE).

English vowels

Lexical label	PSEE phoneme	WE phoneme
Short Vowels		
KIT	/ɪ/	/ɪ/
DRESS	/ɛ/	/ɛ/
TRAP	/æ/	/ɐ/
LOT	/ɔ/	/ɔ/
WANT	/ɔ/	/ʌ/
STRUT	/ʌ/	/ə/
FOOT	/ʊ/	/ʊ/
Reduced Vowels		
LETTER	/ɜ-/	/ɜ-/
COMMA	/ə/	/ə/
R-linking Monophthongs		
BATH	/ɑ:/	/ɐ/
PALM	/ɑ:/	/a:/
START	/ɑ:/	/a:/
THOUGHT	/o:/	/ɒ:/

NORTH	/ɒ:/	/ɒ:/
FORCE	/ɒ:/	/ɒ:/
NURSE	/ɜ:/	/æ:/
R-linking Diphthongs		
NEAR	/ɪə/	/jœ:/
BEER	/ɪə/	/ijə/
SQUARE	/ɛə/	/e:/
LAIR	/ɛə/	/ejə/
FIRE	/aə/	/əjə/
TOWER	/ɑə/	/əwə/
POOR	/ʊə/	/uwə/
BREWER	/ʊə/	/ɪwə/
Yod-initial Vowels		
TUNE	/jʊu/	/ɪu/
CURE	/jʊə/	/ɪwə/
Diphthongs > /i/		
FLEECE	/i:/	/i:/
FACE	/ei/	/e:/
BAIT	/ei/	/ei/
PRICE	/ai/	/əi/
CHOICE	/oi/	/ɒi/
Diphthongs > /u/		
MOUTH	/au/	/əu/
GOAT	/əu/	/ɒ:/
KNOW	/əu/	/ou/
GOOSE	/ʊu/	/u:/
FLEW	/ʊu/	/ɪu/

Welsh vowels

Short Vowels	
PRIN	/ɪ/
LLYN	/ɛ/
NEN	/ɛ/

RHAN	/ɐ/
LLON	/ɔ/
CRWN	/ʊ/
YN	/ə/

Long Vowels

GWIN	/i:/
CLUN	/ɪ:/
GÊN	/e:/
GLÂN	/a:/
BÔN	/o:/
CŴN	/u:/

Diphthongs

SAIN	/eɪ/
GWAUN	/aɪ/
HAEN	/a:ɪ/
TROI	/ɔɪ/
POEN	/o:ɪ/
CWYN	/ʊɪ/
EIN	/əɪ/
CREU	/əɪ/
LLIW	/ɪu/
DRYW	/ɪu/
RHEW	/ɛu/
GLAW	/ɛu/
BYWYD	/əu/

Glossary

Abbreviations

f. *feminine*

lit. *literally*

m. *masculine*

pl. *plural*

sg. *singular*

Initialisms and acronyms

(Where necessary, initialisms are defined later in this Glossary and/or on their first appearance in the main text.)

EL *extrinsic language*

ELU *extrinsic lexical unit*

GPC *Geiriadur Prifysgol Cymru*

IL *intrinsic language*

ILU *intrinsic lexical unit*

NP *noun phrase*

MLF *Matrix Language Frame*

MS *Member of the Senedd*

OED *Oxford English Dictionary*

OT *Optimality Theory*

PCA *Principal Components Analysis*

PSEE *Prestige Southern England English*

RP *Received Pronunciation*

S4C *Sianel 4 Cymry* (“Channel 4 Wales”)

SFL *systemic functional linguistics*

SSBE *Standard Southern British English*

WE *Welsh English*

Welsh-language terms

bratiaith NOUN

(lit. *language of rags*)

Pejorative term for a language variety judged to be corrupted or degraded. In particular, Welsh which has a high proportion of English-origin loanwords, or which is characterized by code-switching into English. (See Chapter?? of the main text for further discussion.)

canu caeth NOUN

(lit. *bound composition*)

Poetry which adheres to the rules of *cynghanedd* (see below). (Sometimes translated as *strict-meter poetry* (Constantine, 2017)).

canu rhydd NOUN

(lit. *free composition*)

Poetry written without following the rules of *cynghanedd*. Unlike English free verse, *canu rhydd* may feature regular metre and rhyme; it is distinguished entirely by its lack of *cynghanedd*.

cynghanedd NOUN

(lit. *concord*)

A Welsh poetic technique, in which a line is split into parts whose phonetic structure must match in some way. Consonance, alliteration, and rhyme may

all feature as techniques for matching the parts of a line according to intricate rules.

Cymdeithas yr Iaith Gymraeg NOUN

(*The Welsh Language Society*)

Often *Cymdeithas yr Iaith* or simply *Cymdeithas*.

Welsh-language pressure group, formed in 1962 by Plaid Cymru activists who felt that existing efforts to increase the legal status of the Welsh language were proving ineffective (Merriman & Jones, 2009).

Cymraeg NOUN

The Welsh language.

Cymraeg ADJECTIVE

Relating to the Welsh language. Welsh-medium or Welsh-speaking.

Cymreig ADJECTIVE

Welsh by virtue of some non-linguistic characteristic, such as nationality, citizenship, location, culture, or origin.

Cymry NOUN, PL., **Cymraes** NOUN, SG. F., **Cymro** NOUN, SG. M.

Cymry are Welsh people. A Cymraes is a Welsh woman or girl, and a Cymro is a Welsh man or boy. Sometimes these terms refer specifically to Cymrophone people.

Cymry Cymraeg NOUN, PL.

Cymrophone Welsh people.

Cymry di-Gymraeg NOUN, PL.

Non-Cymrophone Welsh people.

eisteddfod NOUN

(**eisteddfodau** PL.) In modern Wales, cultural festivals, usually held annually, featuring competitions in literature, music, and other arts. The modern eisteddfod grew from the revival by Victorian antiquarians of a medieval event, at which bards demonstrated their skills in front of their patrons (E. Edwards, 2019).

gwerin NOUN

(*folk, peasantry, or proletariat*)

In some political discourse, the rural populace of Wales, imagined as Welsh-speaking and as guardians of Welsh tradition. (See section of the main text for further discussion.)

llediaith NOUN

(lit. *semi-language*)

In neutral use: dialect; provincialism. Pejoratively: barbarism; a language variety judged to be corrupted or degraded (cf. *bratiaith*, above).

Plaid Cymru NOUN

(lit. *The Party of Wales*)

Often simply *Plaid*.

Welsh political party. Formed in 1925, Plaid Cymru originally focused its efforts on protecting Welsh cultural and linguistic identity, located particularly in the Fro Gymraeg (see above) (P. Gruffudd, 1995). Over time, however, the party's policies have developed towards the promotion of civic, rather than ethnic, nationalism (R. Jones & Fowler, 2007).

Saeson NOUN, PL., **Saesnes** NOUN, SG. F., **Sais** NOUN, SG. M.

Saeson are English people. A Saesnes is an English woman or girl, and a Sais is an English man or boy. Sometimes (now rarely and usually pejoratively) specifically refers to people who are either monoglot Anglophones, Cymrophones who choose to use English, or Cymrophones who have a mastery of English.

Senedd NOUN

(*Parliament, Senate*) The devolved legislature of Wales, commonly called the Senedd, and officially named *Senedd Cymru* in Welsh and *Welsh Parliament* in English.

treigladd NOUN

The system of changes to word-initial Welsh consonants in particular syntactic positions. Often called *initial consonant mutation* or simply *mutation*. The Welsh term is preferred in this study, since it is an unambiguous technical term referring specifically to the Welsh system.

trawsieithu NOUN

Coined by C. Williams (1994), trawsieithu is a strategy for use with bilingual pupils, where materials in one language are used as the basis for pedagogical tasks in another. The word is composed from Welsh “traws-”, “iaith”, and “-u”: *trans-*, *language*, and a morpheme which indicates that the word is a verbnoun. Although the idea was adopted by Anglophone linguists as “translanguaging”, subsequent theoretical developments have separated translanguaging significantly from its original conception. When referring to the strategies proposed by C. Williams (1994), therefore, this study retains the original Welsh term.

Y Fro Gymraeg NOUN

(lit. *The Welsh-language Region*)

Often translated into English as “the Welsh-speaking heartland”, or simply “the heartland” (R. Jones & Fowler, 2007).

An area made up of the western and north-western parts of Wales, which has had, and continues to have, a high proportion of Welsh speakers, and whose inhabitants tend to identify strongly with a Welsh ethnic and civic identity. The term was popularized by Balsom (1985). (See ?? of the main text for further discussion.)

Language(s), language varieties, and their users

additional language NOUN

A language whose acquisition was begun in later childhood or in adulthood; or a language whose acquisition was begun in childhood, but in which a standard adult level of spoken communicative competence was not achieved before adulthood (for example because use of the language was discontinued, or because there was never sufficient exposure for full acquisition).

Anglophone NOUN and ADJECTIVE

A person who can use the Welsh language, at any level of competence or fluency. Using English as a medium of communication; that uses, or has the ability to use, the English language.

Anglophony NOUN

The ability to use English as a medium of communication; the state or status of being Anglophone.

asymmetric bilingual NOUN

A bilingual whose competence in one language is greater than their competence in the other.

balanced bilingual NOUN

A bilingual whose competence in both languages is essentially equal.

Cymrophone NOUN and ADJECTIVE

A person who can use the Welsh language, at any level of competence or fluency. Using Welsh as a medium of communication; that uses, or has the ability to use, the Welsh language.

Cymrophony NOUN

The ability to use Welsh as a medium of communication; the state or status of being Cymrophone.

embedded language NOUN

The subordinate language in a mixed-language sentence. The matrix language controls the sentence's syntactic frame, and the embedded language provides items to fill positions within that frame.

etymon NOUN

The extrinsic language lexeme from which an ELU is derived. The English {sugar} is the etymon of the Welsh {siwgr}.

Extrinsic Language (*EL*) NOUN

An Extrinsic Language is a language other than the one which is the primary medium of communication of a given instance of language in use.

Extrinsic Lexical Unit (*ELU*) NOUN

A lexical unit whose root morpheme comes from an Extrinsic Language.

Hispanophone NOUN and ADJECTIVE

A person who can use the Spanish language, at any level of competence or fluency. Using Spanish as a medium of communication; that uses, or has the ability to use, the Spanish language.

Intrinsic Language (IL) NOUN

The Intrinsic Language is the language which is the primary medium of communication of a given instance of language in use.

Intrinsic Lexical Unit (ILU) NOUN

A lexical unit whose root morpheme comes from the Intrinsic Language.

language VERB, languaging NOUN

In contemporary linguistics, *language* may be treated as a verb (along with the associated verbal noun *languaging*), highlighting the active nature of language in use, rather than treating languages as objects separate from the actions, actors, and contexts which produce them. Languaging “is how people make sense of a negotiated meaning in interactions with other people” (Babino & Stewart, 2020, p. 76).

linguageness NOUN

A property of a linguistic artefact (from a single phoneme or grapheme to an entire language) which associates it with a particular language. For example, the Cyrillic graphs «Я» and «И» are sometimes used to replace Latin «R» and «N» in English-language texts when the author wishes to create an association with Russianness or some aspect of Soviet history or culture. The Russian linguageness of these characters is strong for Anglophone readers who have seen examples of written Russian, even if they cannot read the Cyrillic alphabet. When these characters are used to spell English words (for example by spelling “Animal Farm” as «АИИМАЛ FA ЯМ»), the resulting text has both English and Russian linguageness, despite having solely English lexical content, and even though the Cyrillic graphemes are not functioning to represent the phonemes as in Russian. The fact that these graphemes are understood as Cyrillic is enough to assign Russian linguageness to the shapes divorced from the phonological units they would usually represent, and their Russian linguageness, in turn, evokes associations with Russian culture and history.

learner NOUN

Someone who is in the process of acquiring a language. The term “learner” will be used specifically to designate those whose process of acquisition is still active, whether because they are receiving instruction, because they are instructing themselves, or because they are acquiring the language naturalistically through repeated exposure. Additional-language users are not always learners: a user

may already have acquired advanced competence, comparable to that of a primary-language user. Alternatively, she may have abandoned active efforts to learn the language, and have no incidental exposure to it.

listedness NOUN

For a lexeme, the status of being present in a dictionary or dictionaries.

matrix language NOUN

The principal language of a mixed-language sentence, which controls the overall structure of the sentence, providing a syntactic frame into which extrinsic-language items are embedded.

mixed-language ADJECTIVE

The term “mixed-language” is used in this study as a general descriptor for any behaviour or text in which more than one language plays a part. This avoids having to be specific about the exact nature of the linguistic act involved, for example by referring specifically to code-switching, code-mixing, borrowing, translation, and so on.

monolingualism NOUN

The act of speaking a single language; the state of having a single language in one’s linguistic repertoire.

As a theoretical term, monolingualism refers to ideologies seeking to enforce or encourage the use of a single language (often limited to a standardized variety of that language). Monolingualism may affect anything from individual interactions (for example, if one member of a conversation expresses disapproval of an interlocutor’s code-switching between two languages) to the national (for example, if a country’s judicial system only permits the use of one language for official legal proceedings) and the international (for example if a global organization enforces the use of a single working language as a *lingua franca* for participants from multiple countries). People subject to monolingualist ideologies may not themselves be monolinguals, but are expected to operate as though they are monolinguals under certain circumstances.

plurilingualism NOUN

An approach to language arising from the work of the Council of Europe in establishing the Common European Framework of Reference for Languages

(Beacco et al., 2016; *CoE: LPD*, 2007; Lenz & Berthele, 2010; Little et al., 2011; Piccardo, 2018), plurilingualism encourages awareness of connections between languages and language varieties. Used particularly as a framework for pedagogy, plurilingualism values an individual's entire linguistic repertoire, including languages known only partially, imagining linguistic competence as a continually evolving body of integrated skills (Coste & Simon, 2009).

Prestige Southern England English (PSEE) NOUN

A diachronic term for the socially prestigious variety of British English pronunciation, which includes both Received Pronunciation (RP) and its more recent descendant, referred to by some scholars as Standard Southern British English (SSBE) (Lindsey, 2019; McMahon, 2020).

primary language NOUN

Any language whose acquisition was begun in early childhood, and at least one variety of which has been acquired to a standard adult level of spoken communicative competence.

sequential bilingual NOUN

A bilingual with two primary languages, one of which was acquired earlier than the other.

simultaneous bilingual NOUN

A bilingual with two primary languages acquired contemporaneously.

target language NOUN

A language which is the object of additional-language acquisition. A learner's language is more or less target-like, depending on how closely it conforms to the rules and norms of (usually a standard variety of) the target language.

(un)marked NOUN, **markedness** NOUN

A linguistic object (including anything from a single unit to an entire language) is unmarked if a proficient user of the language in question would consider its use to be neutral in the given context. For example, in a Welsh-medium classroom, Welsh would usually be the unmarked code, and English would be marked. Users may not agree on the markedness of a given object in context, particularly if the varieties they use are different. For a user of U.S. English, «colour» is a marked spelling; it is unmarked for a user of British English, for whom «color» is the marked spelling.

user NOUN

A human being, considered from the point of view of their linguistic competences and behaviours. *User* is a modality-neutral counterpart to “speaker”: “users of French” or “French users”, rather than “French speakers”.

variety NOUN

A variety is a particular form of a language used by a coherent group of people or in a specific context. In this study, “variety” is most commonly used to refer to dialects and sociolects, which are forms of language used by particular communities defined by geography or by shared social identity. However, the term may also refer to forms of language defined by formality or context of use, such as registers and styles. Varieties can belong to large or fairly small groups: British English, Welsh English, Southern Welsh English, and Cardiff English are all possible varieties, and someone from Cardiff might be described as using any of these varieties, depending on breadth of focus. A language is the sum of the varieties which are recognized by users as belonging together.

Wenglish NOUN

1. A colloquial variety of Welsh characterized by substantial use of English-origin lexis, code-switching to English, and showing the influence of English in its syntax (L. Evans et al., 2019; Madoc-Jones & Parry, 2012; Rymes & Smail, 2021; Selleck, 2013).
2. A variety of English spoken in Wales (Selleck, 2013), particularly one with “substantial lexical borrowing and phonetic and phonological influence from Welsh” (A. James, 2018, p. 28); specifically: the colloquial English of the Valleys region of southern Wales (A. James, 2011, 2018). (Sometimes considered to be derogatory: a “highly trivialising name” for “a monstrous parody of Welsh English” (Penhallurick, 1993, p. 32).)

Linguistic units, systems, processes, and products

emic ADJECTIVE, **etic** ADJECTIVE

Terms introduced by Pike (1967), on the pattern of “phonemic” and “phonetic”, to describe different ways of describing the same object of analysis. An etic approach considers linguistic objects in a discrete, decontextualized manner, using criteria imposed by the analyst. An emic approach considers them in their cultural context, using criteria discovered through studying the

language system itself. Phonetic units, for example, are described in terms of language-agnostic articulatory and acoustic criteria, whereas phonemic units are artefacts of a specific language system, abstracted from the properties of groups of phonetic units which can fulfil the same function in the system.

graph NOUN

Graphs are represented between guillemets: «g». A graph is a representation of a character with specific physical characteristics, as used in an instance of language. Two graphs which are considered to be instances of the same grapheme, like «g» and «G», are called allographs. Like phones, graphs are representations of the actual forms used in an instance of language, but do not necessarily represent the exact physical details of those forms when it is not necessary to do so.

grapheme NOUN

Graphemes are represented between angle-brackets: <g>. A grapheme is an abstract, minimal unit of written language. In alphabetic systems, the basic grapheme corresponds to a single written character, or alphabetic letter. In systems which distinguish upper-case and lower-case characters, a grapheme is usually represented as lower-case, and understood to comprehend both case-forms of the character.

graphemic ADJECTIVE, **graphetic** ADJECTIVE

The graphemic level of linguistic description considers the abstract graphemes which represent broad categories of written marks. The graphetic level of description considers the specific graphs, rather than the graphemes to which they are assigned. In graphemic terms, {cup} is written <cup>. In graphetic terms, {cup} might be written «Cup» at the start of a sentence, «cup» at the end of a sentence, «CUP» on a digital seven-segment display, and so on. These instances differ in how they realize the graphemes <c>, <u>, and <p>: they are different graphetic realizations of the graphemic sequence <cup>. Graphetic transcription, like phonetic transcription, may be broad or narrow (Hayes, 2009). Given the following instance of written language: *graphetic*, a broad graphetic transcription could represent the first graph as «g», with a looped tail, where a narrower transcription would seek to match the open-tailed variant used in the original: «g». An even narrower transcription would wish to capture finer detail, for example representing the direction in which handwritten

graphs slope, or where they sit in relation to lines on ruled paper. In this study, all graphetic transcriptions are broad.

lexeme NOUN

Lexemes are represented inside curly braces: {lexeme}. A lexeme is made up of one or more morphemes, and “is the set of forms that minimally contain the same lexeme-stem, or one of its suppletive counterparts, and that may only contain inflectional affixes in addition” (Haspelmath, 2023, p. 27). A lexeme is labelled for convenience with its stem form; that is, the form to which inflectional morphemes may be added, but the lexeme is not just the root form. It comprises the abstract lexical item, including all its possible variation. The wordforms «love», «loving», «loved», and «loves» are all instances of the lexeme {love}. Lexemes are abstract: the lexeme itself does not have a fixed form, and can be realized either in writing or in speech. For the purposes of this study, lexical items with the same root morpheme but different parts of speech are considered to belong to a single lexeme. The wordform «loves» might therefore either represent a noun with plural inflection, or a verb with third-person singular present-tense inflection, but in either case it is an instance of the same lexeme: {love}. Forms with added derivational morphemes are considered to be separate lexemes: {lover} is a separate lexeme from {love}.

morpheme NOUN

Morphemes are represented inside curly braces: {morpheme}. A morpheme is an abstract unit: specifically, it is the smallest linguistic unit which is capable of expressing a denotational meaning. This meaning may be lexical or grammatical (such as “plural” or “past-tense”).

morph NOUN

A morph is a sequence of zero or more phonemes which realizes a morpheme. (See section 2.ii.d for further discussion of the morph as a sequence of phonemes). Morphs which realize the same morpheme are called allomorphs. The English plural morpheme, for example, has the allomorphs /z/ (as in {kids}), /s/ (as in {kits}), /ɪz/ (as in {kisses}), and /∅/ (the zero-morph, as in {sheep}).

orthography NOUN, **orthographic** ADJECTIVE

Orthography is used in this study as a counterpart to *phonology*, for written language. It therefore refers to the system governing the organization of

written symbols, including how various written shapes are grouped into graphemes, and the orthotactic rules which decide how graphemes can be combined in sequence.

phone NOUN

Phones are represented between square brackets: [ə]. A phone is a representation of an actual speech sound, with specific articulatory features, as used in an instance of spoken language. Phones are realizations of phonemes, which act as abstract categories for the different potential ways to pronounce the same speech sound. Phones which can realize the same phoneme in a particular variety of a language are called allophones. Two phones are allophones of a phoneme if the difference between them is linguistically redundant: “in phonological terms, this means that the difference arises automatically in different contexts but does not convey any new or different information” (McMahon, 2020, p. 18).

phoneme NOUN

Phonemes are represented between forward-slashes: /ʌ/. A phoneme is an abstract, minimal unit of speech. Specific speech sounds (phones) are grouped together as belonging to the same phoneme, on the basis of a number of tests which show how these sounds are distributed in the language in question. For example, the phoneme /ʌ/, which represents the vowel of the English word {cut}, might be realized with the phone [ə] by a speaker of a Welsh variety of English, [ɐ] by a speaker of Received Pronunciation, and [ʊ] by a speaker from Yorkshire in England. If we were to record these speakers’ pronunciation of words such as {strut}, {slump}, and {sun}), we would be likely to find that each speaker uses the same phone as they used in {cut} to realize the vowel in each of these words. We could then say that the phoneme /ʌ/ has each of these phones (as well as a variety of others, for speakers with different accents) as a potential realization.

phonemic ADJECTIVE, **phonetic** ADJECTIVE

The phonemic level of linguistic description considers the abstract phonemes which represent broad categories of speech sounds. The phonetic level of description considers the specific phones, rather than the phonemes to which they are assigned. In phonemic terms, {cut} is pronounced /kʌt/. In phonetic terms, {cut} might be pronounced [kət] by a speaker from Aberystwyth in Wales, and [kʊt] by a speaker from Yorkshire in England. These speakers differ

in the realization of the phoneme /ʌ/: they are therefore producing different phonetic realizations of the phonemic sequence /kʌt/.

phonology NOUN

The system which organizes phones into phonemes, and uses phonotactic rules to govern how phonemes can be combined in sequence, both as the expression of individual lexemes, and in stretches of speech composed of numerous lexemes.

scoligraphy NOUN, **scoligraphic** ADJECTIVE

Scoligraphy refers to spellings in alphabetic scripts which obscure the phonological form of a word. The famous <ghoti> as a spelling of {fish} is an example of deliberate scoligraphy, but scoligraphy need not be intentional. Unadapted loanwords are often scoligraphic in their new language, if the originating and borrowing languages have very different conventions for representing phonology with written forms. (See chapter 4 for a full discussion of this term.)

sublexical ADJECTIVE

A term used particularly in discussing reading strategies, to refer to information about a word's graphemes and phonemes. Sublexical information about a word does not rely on recognizing it as a meaningful lexeme; it is exclusively derived from the individual units which make up the word's form.

submeaningful ADJECTIVE

Submeaningful units are linguistic units which do not carry denotational meaning. Morphemes are the smallest meaningful units in language. Phonemes and graphemes are considered submeaningful, since it is only as syntagms which realize morphemes that they convey meaning. In English, for example, the grapheme <s> at the end of a word will often be a realization of the plural morpheme, but that does not mean that every instance of the grapheme <s> conveys a plural meaning: the word <lens>, for example, is a singular noun. Despite being commonly associated with the plural morpheme, therefore, it is not the grapheme <s> itself that carries plural meaning.

syntagm NOUN, **syntagmatic** ADJECTIVE

A syntagm is a sequence of linguistic units combined to make up a larger unit: a spoken word is a syntagm of phones; a paragraph is a syntagm of sentences.

The lexeme {bookcase} is a syntagm made up of the morphemes {book} and {case}

text NOUN

Following the use popularized in Systemic Functional Linguistics, a text is defined in this study as a coherent unit of language in context (Matthiessen et al., 2010, pp. 218–19). Texts can be produced in any modality, including speech, although the term *utterance* will often be used to refer specifically to a spoken text, since it is less prone to misinterpretation.

utterance NOUN

A coherent unit of spoken language in context.

verbnoun NOUN

(Welsh *berfenw*)

A special type of noun which can participate in light-verb constructions (P. Willis, 1988), although it is often treated as a verb in dictionaries.

English-origin verbs are usually converted to verbnouns by adding the most common verbnoun suffix {-o} (which may be realized as <-o> or as <-io>): {facebooko} – *facebooking*; *to facebook*; {cario} *carrying*; *to carry*.

variant NOUN

Variants are represented inside angle brackets: <variant>. In this study, the term is used to describe sequences of graphemes representing a lexeme stem: <color> and <colour> are variants of a single abstract lexeme, {colour} (the form chosen to label the lexeme is arbitrary; it might equally be labelled {color}). Variants therefore have a specific orthographic form, but any inflectional morphemes, as well as changes due to treiglad, have been stripped from them.

wordform NOUN

Wordforms are represented between guillemets: «wordform». They are the actual sequences of graphs produced in a given instance of written language in use.

writing NOUN, **written** ADJECTIVE

Language represented in visual or tactile form, for example by making marks on a surface, by causing a representation of marks to be displayed and/or stored in digital form, or by forming shapes from a physical material. “Writing”

therefore includes language which is handwritten, painted, drawn, carved, printed, typed, embossed, molded, sculpted, and so on.

Introduction: “English words, either pure or Cymricised”

A year ago, I listened to a clergyman from Wales, who pretended to speak Welsh. But [...] he used, in each sentence, an average of two English words, either pure or Cymricised.

(D. J. Evans, 1889, p. 142)

Welsh and English are related, both being members of the Indo-European language family (MacAulay, 1992). This distant relationship is not particularly evident, however; Welsh and English are not mutually intelligible. In modern Wales, Welsh has *de jure* equal status with English in public life: both languages are official languages of the Senedd (the devolved parliament of Wales), for example (Senedd, 2022). However, the history of Welsh is one of official repression, derision, and neglect, from the 1536 Acts of Union, which formalized the English conquest of Wales, up until the modern era. Welsh was removed from all official use by the Acts of Union, and the Welsh people slowly began to turn to English as the only practical way to access education and social advancement (A. R. Thomas, 1992).

Nonetheless, Welsh remained a living language, albeit amongst a dwindling minority of the population (A. R. Thomas, 1992), until concerted political activism in the mid-twentieth-century led to official changes in its status. The Welsh Language Act of 1967 made it possible to use Welsh in legal proceedings, and allowed for the provision of official documents in Welsh (*Welsh Language Act*, 1967), although it was not until 1993 that Welsh was formally placed on an equal footing with English in Wales (*Welsh Language Act*, 1993). Further legislative action during the twentieth and early-twenty-first centuries has strengthened the legal position of Welsh. The devolved National Assembly for Wales (later renamed the Senedd) passed the *Welsh Language Measure* (2011), which established the office of the Welsh Language Commissioner with the brief to promote and protect Welsh and to ensure that there is no discrimination against the language in public life (Comisiynydd y Gymraeg,

n.d.).

Although the official status of Welsh in Wales has been established in law for decades, however, it remains a minority language. Only 19% of Census respondents in 2011 indicated that they were able to speak Welsh, a number that falls to less than 15% when ability to read and write Welsh are also included (ONS, n.d.). At present, English is a (and, often, the only) primary language of the majority of the population of the United Kingdom, including Wales (ONS, n.d.; National Records of Scotland, 2013; Northern Ireland Statistics and Research Agency, 2012)¹.

As a result of the historical oppression and repression faced by Welsh people in Wales, some scholars of the late-twentieth- and twenty-first-century have theorized modern Welsh identity in postcolonial terms (e.g. Aaron & Williams, 2005; Bohata, 2004; Mac Giolla Chríost, 2005). Wales has been described as “England’s first colony” (M. Johnes, 2019, p. 2), situating Welsh people—Cymrophone or not—as postcolonial subjects. However, the later complicity of Welsh people in the British imperial project places modern Wales in a far more complex relationship to England and to Britain than that of simply an oppressed colony (M. Johnes, 2019). As ap Gareth (2009, p. 35) penetratingly notes, Wales is both at the centre of Empire and peripheral to that centre. Even as Cymrophones faced the denigration of their language and constant official pressure to abandon it in favour of English, they were also enthusiastic participants in colonizing the Americas, seeking to transplant their language into lands made available by the violent dispossession of indigenous peoples (Iyengar, 2015).

¹ Census data for 2011 are the most recent available at the time of writing. It is not possible to give precise numbers for English as a primary language in the U.K., for two reasons. Firstly, the 2011 England and Wales census question about “main language” recorded use of English as a main language everywhere except in Wales, where the question recorded use of English or Welsh as a main language (ONS, n.d.). It is therefore impossible to say exactly how many respondents consider English, rather than (or in addition to) Welsh, to be their “main language”. Secondly, the 2011 census for Scotland did not ask a question comparable to the “main language” questions asked in the England and Wales census and the Northern Ireland census (Office for National Statistics, 2015). The claim that a majority of the population use English as a primary language is therefore based on the following sources: actual census responses for England and Northern Ireland (where 92% and 96.8% of respondents, respectively, report English as their main language) (ONS, n.d.; Northern Ireland Statistics and Research Agency, 2012); an estimate for Wales, where 97% of the population report that their main language is either English or Welsh, and 14.6% report being able to speak, read, and write Welsh (ONS, n.d.), implying that an absolute minimum of 82.4% of the population have English, not Welsh, as their main language; and the high numbers reporting ability to speak, read, and write English and/or Scots in the Scottish census (National Records of Scotland, 2013, Table QS213SC).

This chapter's epigraph comes from an 1889 article in the Welsh periodical *Cymru Fydd*, written by D. J. Evans, a professor of classical languages at the University of Ohio. The settler-colonial project in North America was attractive to Welsh people who felt themselves and their communities to be under threat from rapacious capitalism and industrialization (Van Vugt, 2006, p. 70), and Ohio had a substantial Welsh population by the mid-nineteenth century. Although in absolute terms the number of Welsh immigrants to North America was not particularly large (Tyler, 2018), they were disproportionately represented amongst British immigrants in Ohio. In 1850, 18% of British immigrants to Ohio were Welsh, whereas Welsh people made up only 5% of the population of Britain itself at that time (Van Vugt, 2006, p. 49). The early Welsh communities of Ohio were notable for their insularity and adherence to Welsh linguistic and cultural identity, to the extent that Welsh immigrants who did not speak Welsh tended to settle amongst English immigrants or American-born settlers rather than in the established Welsh communities (Van Vugt, 2006).

The Welsh language would ultimately be lost as a community language amongst U.S. settlers. Immigration from Wales declined, and the American-born Welsh increasingly married outside their community. Welsh-American communities were perceived as pious, sober, and hard-working, with a conspicuous Protestant identity which made them more acceptable to mainstream American culture than, for example, their Irish Catholic counterparts (Tyler, 2018). The Welsh were enthusiastically welcomed as ideal citizens of the white American state, and their gradual adoption of the English language removed the final significant barrier to assimilation (Tyler, 2018).

Despite this, some Welsh-American communities were able to sustain a distinct ethnolinguistic identity until well into the twentieth century (Tyler, 2018). The maintenance of Welsh culture was encouraged by the Cymrophone American press: the popular newspaper *Y Drych*, for example, was essentially a monolingual Welsh publication from its first appearance in 1852 until the 1920s (Coupland et al., 2003). At the time D. J. Evans (1889) told his anecdote, many of Ohio's Welsh citizens were Cymrophones, some of them monoglots and some Welsh-English bilinguals (Tyler, 2018).

The linguistic situation of the Ohio Welsh of the late 1800s reflects and parallels the situation in the country they had left behind them. As in Ohio, the Welsh language in Wales of the nineteenth and twentieth centuries faced major demographic and cultural changes, and pressures to assimilate into majority,

Anglophone culture. Welsh was, and still is, deeply implicated in discourses of ethnolinguistic belonging. It is central to identity construction at the level of both the individual and the community. As we shall see in Chapter 3.i.a of this study, Cymrophone communities in Wales, like the early Welsh communities of Ohio, have been perceived as insular and unwelcoming to non-Cymrophones on the one hand, and idealized as pious and culturally pure on the other.

When D. J. Evans (1889, p. 142) describes a clergyman who “used, in each sentence, an average of two English words”, he is describing a type of linguistic behaviour which, as we shall see, is widespread amongst bilinguals, regardless of which languages they use. Now commonly called “code-switching”, the alternation of languages by bilinguals is a practice which both raises strong emotions in popular linguistic discourse, and has preoccupied scholars for decades. Language-mixing behaviours amongst Welsh-English bilinguals are not mere matters of individual linguistic selection, but are performed and interpreted within a heavily ideological understanding of language. Evans’ description of the speaker as one “who pretended to speak Welsh” illustrates the ideological weight carried by code-switching, a practice which is imagined as a sign of moral failing and as a threat to linguistic integrity.

The description of the clergyman’s English words as “either pure or Cymricised” (D. J. Evans, 1889, p. 142) introduces another dimension to the discussion. The “pure” English words are those which retain their English-language form. The “Cymricised” words, on the other hand, are those which have been adapted in some way (perhaps phonologically or morphologically) to the Welsh linguistic system. Such adapted words exist in a liminal position: one of the major questions in the scholarly literature on language-mixing is whether to categorize these words as code-switches or as loanwords. Ideologies of nation and identity permeate the languages of Wales, and every word is subject to scrutiny in the pursuit of linguistic purity.

The processes by which English-origin lexical items come to be labelled as “English” or as “Welsh” are the main focus of this study. I have not come across an established term to describe the perception that a word belongs to one or another language. Accordingly, I will refer to this property as the “*linguageness*” of lexical items. The processes by which linguageness is determined may appear to operate on a limited scale, affecting only the individual lexical units themselves. However, I will show that the ascription of linguageness to a lexical unit reflects the linguageness of much broader constructs influenced by sociocultural pressures, and that approaches to

labelling English-origin items in Welsh are intimately connected to ideologies of Welsh identity.

“A great and growing ambition for the Welsh language”

In 2017, the Welsh Government announced its *Cymraeg 2050* (2017) strategy, declaring a target of a million Welsh speakers in Wales by the year 2050. A white paper was released in the same year, proposing a Welsh Language Bill which would reduce the bureaucratic burden of Welsh language policy, focusing on encouragement and promotion of the language rather than on enforcing compliance with regulations (*Striking the balance*, 2017). There is an emphasis on liberalism and choice in language policy in modern Wales, which emphasizes providing opportunities for learning and using Welsh and English on equal terms (Selleck, 2013).

Nonetheless, this study will show that illiberal and repressive language ideologies have historically been widespread amongst Cymrophones, even as they have also been applied to Cymrophones by Anglophones. The legacy of such ideologies persists, particularly in Welsh-medium pedagogy. Examining the sociocultural constructs of Welshness (and Englishness) which underpin language ideologies in Wales, we shall see that linguistic purism is bound up with complex ideas about cultural survival and the maintenance of tradition which combine to create exclusive and exclusionary images of who can, and who may, speak Welsh. These images are at odds with, and undermine, policy makers' attempts to increase Welsh-language use in modern Wales.

A Welsh Government report in 2013 noted the low standards of attainment in Welsh by students in English-medium education, and recommended a wide-ranging reform of the place of Welsh in the school curriculum (*One language*, 2013). At the time, Welsh was treated as two separate subjects: “Welsh first language”, studied by pupils at Welsh-medium schools, and “Welsh second-language”, studied by pupils at English-medium schools. In 1999, Welsh had been made compulsory for all pupils in Welsh state schools, up to Key Stage 4 (which is usually completed by age 16) (*Welsh in education*, 2017). Students in Welsh-medium education must attempt GCSE qualifications in “Welsh first language”, whereas those in English-medium schools are only required to study the subject, with the option to attempt a GCSE in “Welsh

second language” if they and their teachers so desire. As a result, many learners never achieve a qualification in Welsh, and Welsh second language has lagged behind other curriculum subjects in levels of attainment at all stages of formal education (*One language*, 2013).

In the field of educational policy, a 2015 independent review of the primary and secondary school curriculum (Donaldson, 2015) recommended that Welsh-language provision in schools in Wales be strengthened and expanded. The report cites evidence that “the Welsh language and bilingualism”, along with “the focus on Welsh identity” were perceived as a particular attraction of the education system in Wales (Donaldson, 2015, p. 15).

As part of achieving the *Cymraeg 2050* targets, and with evidence from Donaldson (2015) and *One language* (2013) (amongst others) in mind, the Welsh Government has proposed a strategy to improve standards of attainment in Welsh by pupils in English-medium schools. As recommended by *One language* (2013), in the near future pupils in English medium schools will work towards qualifications in “Welsh”, rather than “Welsh second language” (*Cymraeg 2050: changes to Welsh Language qualifications*, 2022). This change reflects the desire for a “continuum of learning” (*One language*, 2013, p. 2) across all schools in Wales, giving pupils “a fair and equal opportunity to achieve in Cymraeg, regardless of which type of school they attend” (*Qualifications Wales*, 2022). A spokesperson for Qualifications Wales, the regulator for examinations in the state education sector in Wales, stated the long-term aim of developing “one over-arching Cymraeg qualification for all learners in all settings” (Wightwick, n.d.). The overall message implied by these curricular reforms is that users of Welsh are not to be separated into discrete groups based on their level of fluency or communicative competence, but rather are to be understood as a single community of learners, united by a common language.

Nonetheless, as I will argue in the conclusion to this study, the realities of educational practice and policy for Welsh-medium and English-medium schools in Wales falls far short of the ideal of a united community of learners. Welsh-medium pupils’ exposure to Welsh literary and cultural traditions is far greater than that of their English-medium peers. On the other hand, Welsh-medium pupils suffer from difficulties not experienced in English-medium schools, such as a lack of pedagogical resources in their language of schooling (C. Baker, 1995; BBC News, 2016; C. H. Thomas, 1982; C. Williams, 1994). As work by Selleck (2013), amongst others, shows, Welsh-medium pupils also face repressive and punitive attitudes to their natural linguistic behaviours.

While Welsh public policy over recent years reflects “a great and growing ambition for the Welsh language” (*Striking the balance*, 2017, p. 2), there is a perpetual undercurrent of concern about the sociocultural context in which Welsh operates. Responses to the new curriculum’s plans for increasing the use of Welsh in English-medium schools have included concerns about the perception of the language in largely Anglophone communities, “with messaging needing to communicate that Welsh is not a threat to non-Welsh speakers” (N. Morgan et al., 2022). Popular attitudes towards language are, in many cases, proxies for attitudes about other markers of social identity (J. Edwards, 2009, p. 94). These attitudes form the context for language use; any attempt to promote and strengthen Welsh in Wales must reckon with language’s role in shaping national identity and popular discourse about Wales and Welshness.

Understandably, perhaps, the Welsh Government’s concern is largely with gaining support for Welsh from those who may be resistant to promotion of the language. I will argue, however, that attitudes from within the Cymrophone community are equally important in encouraging uptake of the language, and in supporting Welsh as an attractive, dynamic, and modern mode of expression. I will propose plurilingual approaches (Beacco et al., 2016; Lenz & Berthele, 2010) to language in Welsh education, drawing on evidence of skilful linguistic creativity in Cymrophones’ deployment of English and Welsh combined.

Language as a medium for education

Strict monolingual policies are not unusual in educational contexts. Whether the medium of education is a dominant language or a minoritized one, monolingualism is often felt by policy makers, teachers, and pupils to be the most appropriate way to ensure full acquisition of the language of schooling (Iannacci, 2008; I. U. Khan et al., 2023; Willans, 2010). Teachers may feel constrained by institutional or governmental policy to encourage or enforce a separation of languages in the classroom, even if their personal preference is to be more permissive (Stevenson, 2015). Code-switching in writing is particularly censured in many educational contexts, as it is at odds with ideals of formal language use (Canagarajah, 2011).

When pupils struggle with schooling, it is common to blame linguistic factors and to insist on “un renforcement des «fondations», un retour aux «bases»”², with an

² reinforcing the “foundations”, a return to the “basics”

approach “centrée sur le seul fonctionnement du système et la dimension normative de la langue et au détriment de l’expression personnelle spontanée et de l’usage de toute langue autre”³ (Castellotti et al., 2008, p. 11). The acquisition of a standardized variety of the language of schooling is understood as the foundation for all educational success, while a pupil’s other languages and varieties are deliberately sidelined in order to focus on the standard variety.

When the medium of education is a minoritized language, additional factors come into play. Justifications for monolingualism in these contexts may include ensuring that students gain sufficient exposure to a language they may not encounter outside school (Ballinger et al., 2017; Juan-Garau & Lyster, 2019), as well as symbolizing resistance to the dominant language and culture (W. M. Collins, 2003). Ballinger et al. (2017, p. 26) recommend caution in relaxing monolingual policies in immersion education, suggesting that attention must be paid to the status of the languages in question to ensure that allowing majority-language use is truly pedagogically useful, taking care “to avoid the very societal language imbalance that immersion programs are often designed to redress”.

Such arguments in favour of monolingualism in minority-language education are often framed in terms which place the maintenance of the language, rather than the academic, linguistic, and cultural development of the pupil, as the primary goal of immersion education. This elides the importance of affirming students’ complex cultural and linguistic identities, and the positive effect that such affirmation may have on academic success (Reyes, 2008). A child in a bilingual school in Wales told Selleck (2013, p. 31) that teachers “always say the language is dying and stuff like that” when enforcing Welsh-only policies in the classroom, implying that the protection of Welsh as a living language is the primary justification for monolingualism in the school.

Public discourse relating to Welsh and its position in Wales has tended to emphasize its status as a minority language and, as I will show, discourse within the Cymrophone community generally reflects this perception. English is positioned as a dominant, colonizing force, indexing Englishness, Britishness, and Anglo-American imperialism. Welsh is defined in opposition to English, as an enclave in the linguistic landscape, continually fighting to maintain the integrity of its borders. As evidence

³ *focused exclusively on the functions of the system and the normative dimension of language, and to the detriment of spontaneous personal expression and of the use of all other languages*

from Selleck's (2013) informants shows, this construction of English and Welsh may alienate not only non-Cymrophones from outside Wales, but also Welsh non-Cymrophones, and even Cymrophones themselves.

The dangers of monolingualism

There is ongoing concern about the actual amount of Welsh used by Cymrophones in their everyday lives (L. Evans et al., 2019). The individual's sense that she can use the language skillfully is vital to developing confidence (L. Evans et al., 2019), and confidence in one's command of a language will often translate into increased actual language use (Starks, 2005; E. M. Thomas et al., 2014a). Welsh-medium pupils from English-speaking homes report significantly less confidence in using Welsh, along with significantly less use of Welsh in general than those who have Welsh as a home language (E. M. Thomas & Roberts, 2011).

At school, confidence affects learning strategies, with pupils educated in a language in which they do not feel confident favouring rote-learning rather than attempting to understand the material (Watkins et al., 1991). According to Starks's (2005) work in New Zealand, although speaking a majority language at home can reduce confidence in minority language skills, receiving primary education in the minority language can go a long way towards mitigating this effect. Nonetheless, E. M. Thomas and Roberts (2011) suggest that low confidence in the use of Welsh may arise when young people perceive it as the language of education alone, rather than as a general-use medium for interaction.

Willans's (2010), working with Vanuatan pupils in Anglophone education whose primary language is Bislama, found the English-only policy of the school to be a significant barrier to effective learning and communication, describing an interaction where students felt unable to ask their teacher for the information they needed because the question was too difficult for them to formulate in English. She concludes that "the students switched to English for the sake of obeying the school rule, rather than because this is the best language in which to convey meaning effectively" (Willans, 2010, p. 13). When teachers were not present, students created bilingual spaces in which communication was generally more successful and authentic. When the English-only policy was enforced, in contrast, a simulacrum of communication was often the result, using strategies such as choral responses, repetition of the teacher's words, and reading text from the blackboard, to create the outward impression that learning was taking place (Willans, 2010).

The implications of monolingualism in education go beyond its effects on the mechanics of the learning process, and touch on issues of personal identity formation. Despite the vast differences in historical and cultural context between Wales and South Africa, the rationale offered by Dampier (2012) for supporting a bilingual pedagogy for Black South African pupils has much to offer in discussing the Welsh approach to language in education. Dampier criticizes the official South African policy of additive bilingualism, which advocates introducing English as a language of instruction only once the home or community language has been acquired to a certain level of competence. This, he argues, treats English in an entirely instrumental manner, as “a tool to be used solely for social, economic, educational and political gain” (Dampier, 2012, p. 75), and assumes that only the students’ home language(s) can have a role in identity formation and cultural identification. In attempting to redress the harms done to African languages during apartheid, South African policy now deprives students with non-English home languages of the opportunity to develop a truly bilingual identity, and to define themselves as belonging to two (or more) different linguistic cultures:

Students will enter the language classroom knowing that they are Zulu, for instance, and that they are being taught a language that is not actually South African (viz. English), even though the constitution says it is. They are consequently never given the opportunity in official policy to develop, formulate and articulate a hybrid cultural-linguistic classroom identity that would be distinctly South African (Dampier, 2012, p. 75).

Just as South African educational policy positions English as “a language that is not actually South African”, Welsh-medium education (and Cymrophone culture more generally) often positions English as a language that is not actually Welsh, as we shall see in later chapters. For children whose home language is English, the monolingual-Welsh approach of Welsh-medium education risks denigrating an important aspect of their identity. It is not uncommon for Cymrophones who acquired Welsh through immersion education rather than at home to feel insecure about the standard of their Welsh (Morris, 2014). School policies that prohibit and denigrate English and mixed-language behaviours such as code-switching may therefore fail to provide these pupils with any linguistic identity that they can claim with confidence, and certainly fail to encourage the development of a healthy, integrated bilingual identity. As Chapter 3 will argue, Cymrophony can offer benefits in terms of cultural identity, community, and a sense of authentic Welshness.

However, purist, monolingualist constructs of Cymrophony mean that many Cymrophones consider themselves unworthy of claiming these benefits.

Ironically, Welsh-medium education risks reproducing the linguistic purism and monolingualism of hegemonic English education policies found throughout the Anglophone world, which insist on English-only classroom interaction, devaluing students' home, community, or heritage languages as irrelevant at best, and barriers at worst, to educational success. Although English is evidently the dominant language in the wider social context of Wales, Welsh-medium schools' language policies may create microcultures which minoritize students whose home language is English. This experience could reasonably be predicted to make such students feel more comfortable outside Cymrophone spaces, where they do not feel themselves to have a minoritized identity.

Supporters of Welsh-medium education emphasize the benefits of bilingualism, and in particular claim that it is an advantage in learning further languages (Coslett, 2019). Their opponents in the debate are positioned as colonizers (Coslett, 2019) who would actively deny their children the right to learn an invaluable skill and to acquire cultural and emotional capital as citizens of Wales (Ffransis, 2015). Support for bilingualism is also reflected in two recent Welsh Government reports (LinguaNi, 2020; *Welsh in education*, 2017), which emphasize the need to raise teachers', pupils', and parents' awareness and appreciation of bilingualism. However, this is framed within a broader aim of recognizing the value of learning Welsh.

To a large extent, the Welsh Government's emphasis is on increasing the status of Welsh, improving the teaching of Welsh in English-medium schools, and promoting Welsh-medium education to non-Cymrophone parents. There is far less focus on nurturing the bilingualism of Cymrophone children, or on recognizing the bilingual as something other than a person with two separate monolingual competences. Awareness of what is needed to create confident bilinguals seems lacking amongst many policy-makers in Wales, and the monolingualism of pedagogic discourse fails to respect bilingualism of modern Cymrophone identities. The need to plan for truly bilingual education fitting the realities of Cymrophone language use is pressing.

Seeking an alternative: plurilingual education

Recent decades have seen the development of plurilingualism as an approach to language and languages in education (e.g. Beacco et al., 2016; *CoE: LPD*, 2007; Lenz

& Berthele, 2010; Little et al., 2011), arising from the work of the Council of Europe in establishing the Common European Framework of Reference for Languages (Piccardo, 2018). Plurilingualism is a reaction to the monolingualist assumptions which dominated scholarship and pedagogy of the nineteenth and twentieth centuries; it emphasizes connections between languages and varieties, and values even partial or oblique knowledge of a language (Piccardo, 2018). The plurilingual individual's linguistic competence is not imagined as the sum of discrete competences in separate languages and varieties, but as an integrated repertoire, which is continually evolving (Coste & Simon, 2009).

European educational institutions are, on the whole, not well positioned to foster plurilingualism, since mainstream pedagogy tends towards instrumentalist understandings of language(s) in terms of future educational and occupational prospects (Coste & Simon, 2009). The educational contract which exists between schools and society in modern European states requires that pupils be guided towards academic success, provided with the skills and knowledge expected of a member of the wider community, and prepared to exercise their democratic citizenship responsibly (Castellotti et al., 2008). In most modern states “le postulat a aussi été que la poursuite de ces visées dépendait largement de la maîtrise d’une langue commune, unificatrice”⁴ (Castellotti et al., 2008, p. 10). Chapter 3.i.b will discuss how such ideologies of language as a unifying, civic force are manifested in U.K. and Welsh politics.

Like Dampier, Coste and Simon (2009) reject an instrumentalist view of language(s) in education, advising that “languages in the curriculum in school ought not to be viewed merely as a profitable investment”, but rather that schools should recognize the diverse linguistic resources of their pupils. This, they argue, can help to raise the self-image of pupils who use them (Coste & Simon, 2009). Proponents of plurilingual education do not ignore the civic and societal role of language. Unlike the nationalist language ideologies which will be discussed further in Chapter 3, however, plurilingualism does not confine this role within national borders. Instead, they view plurilingualism as promoting “positive acceptance, a prerequisite for maintaining linguistic diversity”, and “respect for each language” (CoE: LPD, 2007, p. 10). Accordingly, plurilingual education is geared towards finding commonalities

⁴ *the assumption has also been that the pursuit of these aims depends to a great extent on the mastery of a communal, unifying language*

between languages and cultures, developing skills of communication across language boundaries, and celebrating the full linguistic repertoire of each individual. The implications of plurilingualism for Welsh education, and the possibilities it provides for integrating code-switching and language mixing into the curriculum, will be explored in detail in the concluding chapter of this study.

Establishing research questions

As we will see in Chapter 1, work on code-switching and borrowing in Welsh has focused particularly on patterns of English-origin words in Cymrophone discourse, seeking triggers for their use, and asking how we might distinguish code-switching from borrowing. The primacy of Myers-Scotton's (1993) Matrix Language Frame Model of code-switching in research on Cymrophone bilingualism encourages an understanding of bilinguals as using two linguistic systems, with each word in an utterance belonging squarely to one language or the other. This begs the very basic question of what we mean when we identify a word as belonging to a single language, and elides bilingual experience by demanding a neat separation of languages rather than allowing for a more complex, mixed language system. This theoretical model has its counterpart in a language purism which insists on eradicating English words from the Welsh lexicon, yet fails to establish exactly what the criteria for identifying an "English" word are.

The sociolinguistic dimension of the use of English-origin lexical items by Cymrophones has not yet been adequately addressed in the literature. I will therefore examine the tensions between metalinguistic categorization of lexical items as "English" or as "Welsh" words, considering evidence of the attitudes of Welsh-English bilinguals towards these words. I will suggest that binary categorization of English-origin words is harmful to Welsh-English bilinguals' construction of a confident Cymrophone identity, as their real-world experience as bilingual (as opposed to monolingual twice over) is destabilized by monolingualist insistence on the separation of languages.

A plurilingual way of thinking about languages in modern Wales will emerge from this discussion, one which argues for seeing English and Welsh, for all their sociocultural differences, as equal partners in the construction of Welsh linguistic identity. The potential of such a model as a contributor to Welsh Government policy goals will be outlined, to show the potential real-world impact of this linguistic approach. In conclusion, I will offer suggestions for how the fuller understanding of

Cymrophones' language use can inform both scholarly approaches to code-switching and borrowing, pedagogical approaches to the teaching of Welsh, and policy approaches to encouraging Welsh-language use.

Research questions

The full motivation for this study's primary research question will emerge from the discussion in its first four chapters. These chapters review prior literature on language-mixing, situate languageness as an outcome of the semiotic processes of enregisterment, and describe attitudes to language mixing in modern Welsh. Drawing on recent research into dialect respellings of Welsh-English on Twitter (Durham, 2022), as well as on patterns in Anglophone mockery of Welsh, orthography is identified as a site of semiotic meaning.

The subsequent chapters describe the methodology used to design a study of orthographic adaptation of ELUs by Cymrophones, and analyse the data subsequently collected. Patterns of orthographic adaptation of ELUs are first quantitatively analysed, to identify lexemes whose variant forms are of particular interest. These lexemes are then subject to qualitative analysis in context, with a discussion of how differences in orthographic adaptation express meaning beyond the denotational content of the lexeme itself.

The study is driven by a primary research question, with four questions which arise from it. The primary question is the broadest one, and will provide shape to the data collection and analysis, as well as guiding the theoretical discussion. The secondary questions emerge directly from the primary one, and can only be answered in the course of answering the primary question. The tertiary questions, in turn, synthesize the responses to the primary and secondary questions, in order to address the significance of this study to linguistic scholarship and to Welsh-language pedagogy and policy. As the focus on education and policy in this introduction implies, the tertiary questions are by no means less important than the primary or the secondary ones; the arrangement merely reflects the fact that some questions need to be answered first in order to provide a basis for answering others. Indeed, RQ5 will ultimately emerge as the most consequential of the questions to be considered here.

Primary question

Q1: Does variation in the orthographic adaptation of English-origin lexical items in modern colloquial Welsh function as a vehicle for

semiotic meaning?

Secondary questions

Q2: To what extent can the observed variation in orthographic adaptation be ascribed to enregisterment?

Q3: To what extent can the observed variation in orthographic adaptation be characterized as variation in languageness?

Tertiary questions

Q4: Do the observed patterns of variation have implications for scholarly approaches to code-switching and borrowing?

Q5: Do the observed patterns of variation have implications for Welsh-language pedagogy?

Chapter outlines

1. “When the foreign words are camouflaged in plain type”: Code-switching, borrowing, and other mixed languagings

The first chapter of this study will outline research on code-switching and borrowing, and the possibility of distinguishing between the two phenomena. It will also briefly review attitudes towards language-mixing phenomena in the Welsh context. I will suggest that asking whether a lexical item is a code-switch or a borrowing can, to a large extent, be understood as asking about the language to which it belongs, or its “languageness”. Establishing the answer to this question requires a holistic approach to bilingualism and language-mixing. I will argue that previous work which seeks to categorize code-switches and borrowings in Welsh-English bilingualism has paid insufficient attention to the semiotic potential of language-mixing phenomena.

2. “What’s ‘englyn’ when it’s english?”: A semiotic account of languageness

I introduce the question of how English-origin lexical units are ascribed the property of belonging to English or to Welsh, addressing Research Question 1, and relate this

to the theoretical notion of enregisterment (Agha, 2005, 2009). A selection of Welsh-language educational resources is analysed, including guides to “good Welsh” and reports on school pupils’ examination performance. This analysis shows that the construct “English words” is often invoked without being clearly operationalized. I discuss the implications of this tendency for language confidence and educational success, suggesting that it is part of a reductive conception of Welsh which ignores the diverse styles and registers which the language must exhibit in modern use. Finally, this chapter presents a model of language as a connotative semiotic system, with particular attention to the relationship between orthography, phonology, and the other levels of linguistic meaning and expression.

3. “Do I not love my country because I haven’t properly learnt mutations?”: The Welsh language and social constructions of Welshness

Attitudes towards language-mixing in modern Welsh must be understood in context. The third chapter therefore offers an overview of the Welsh language’s recent history, and discusses the sociocultural construction of Welshness, exploring what is enregistered when we enregister Welsh languageness. It considers the semantics of terms for Welsh language, nationality, and geography in both Welsh and English, showing that constructs of Welshness encoded in the Welsh lexicon are often inseparable from Cymrophony

4. “Rough as Bwlchardudwy”: Orthography and the foreignness of Cymraeg

This chapter gives an overview of the Welsh orthographic system and its differences from English orthography. There is a brief review of the literature on orthographic adaptation of loanwords, before an exploration of the potential of orthography as a locus of social and cultural meaning.

5. “For the benefit of Welsh people who had never seen a letter X before”: Designing a study of orthographic variation in the adaptation of English-origin lexical units

This chapter outlines the methodology behind the data collection and analysis to be presented in the following chapters. Three complementary datasets are described: a set of orthographically adapted loanwords from *Geiriadur Prifysgol Cymru*; elicited loanword adaptations from a survey of Cymrophones; and a corpus of Cymrophone tweets collected from Twitter. Finally, the choice of a mixed-methods approach to data analysis is explained.

6. “Ai hop ddat yew can ryd ddys”: Data analysis and discussion

An analysis of the corpora described in the previous chapter, illustrating how Welsh-English bilinguals use the semiotic resources of both languages dynamically and creatively, enregistering various linguistic features as markers of different kinds of Welshness and non-Welshness. A quantitative approach is first taken, showing patterns of use in different contexts which illuminate prior literature on Welsh-English bilingualism, and support a new approach to using dictionary data for measuring loanword integration. Salient lexemes are identified using Principal Component Analysis, and these are subjected to a qualitative analysis, in which contextualized examples of ELU adaptations can be interpreted as performing a variety of semiotic functions.

7. “TIREZ LE FFYCIN RIDEAU”: Conclusions

The final chapter begins by drawing together the data analysis with the theoretical framework established in earlier chapters, showing how the semiotic relationship of orthography to phonology, and of phonology to the full linguistic system, play a part in the enregisterment of Welsh languageness. I then move on to address the research questions, proposing avenues for future research in code-switching and borrowing, and outlining the implications of this study for Cymrophone pedagogy and Welsh-language policy.

Chapter 1

“When the foreign words are camouflaged in plain type”: Code-switching, borrowing, and other mixed languagings

The publishing convention of italicising words from other languages clarifies that some words are imported: it ensures readers can tell the difference between a foreign language and the language of home.

[...]

When the foreign words are camouflaged in plain type you can forget how they came to be there, out of place, in the first place.

I have been thinking about this advice and I have decided to follow it.

Now all of my readers will be able to remember which words truly belong in Aotearoa and which do not.

(Somerville, [2021](#)), italics in original.

1.i When is a lexical unit used by Cymrophones “of English origin”?

This study is concerned with the distinction between “English” words and “Welsh” words, as they are used by Cymrophones, and how these attributes are reflected in Cymrophone orthographic practices. In particular, it is concerned with ELUs of

English origin, and how it is that some of these are perceived as “Welsh” words while others are not. There are a number of questions to be asked about what exactly constitutes an English-origin ELU, taking us back to the early history of the Welsh language.

Contact with other languages has always been characteristic of Welsh. Even before the birth of the Welsh language as distinct from Cornish, Breton, or Cumbric, its Brythonic ancestor coexisted with Vulgar Latin (Schrijver, 2007). Although Welsh was the primary language of the territory inhabited by the Welsh people from the time of Anglo-Saxon invasions until well after the medieval period, communities where English, Flemish, French, or Irish was the common vernacular also came and went (R. R. Davies, 1997; Fulton, 2011). French and Latin were, at various times, spoken languages amongst the nobility and the clergy of Wales (Putter, 2010; Russell, 2013). The monoglot-Welsh peasantry in the medieval period, like their monoglot-English counterparts, may have been numerous, but they lived alongside—and subordinate to—a multilingual elite. Some of these, such as the Norman lords of *Marchia Wallia*, generally had little grasp of Welsh; many others, the nobles and princes of *Pura Wallia*, were fluent Welsh speakers. All would have needed some degree of multilingual competence to gain access to the complex power structures of medieval Britain (Fulton, 2011). The lexical development of Welsh has therefore progressed throughout in close proximity to a socially prestigious and legally dominant cohabitant or cohabitants, as well as to one or more neighbouring vernaculars.

The multilingual evolution of Welsh is reflected in many of the etymologies given in *Geiriadur Prifysgol Cymru* (hereafter *GPC*) for some of the earliest attested words in the language. For *saim* (“fat”, “lard”, “grease”), for example, the source is either direct borrowing from Old French *saim*, or borrowing of the same word via Middle English *seime* (*GPC Ar-lein*, 2024, s.v. *saim*). Relationships between Welsh and other languages in the earlier stage are too complex for English to be treated as the obvious default source of borrowing, although it seems likely that borrowing from French waned along with the fortunes of the vernacular language itself in later medieval Britain (Johnston, 2000).

The complexity of identifying the source of Welsh words with close English counterparts is illustrated by Durkin’s (2014) discussion of {baby}. The English lexeme has been explained as a reduction of Middle English {baban} (“baby” or “infant”), itself an accepted borrowing of the Welsh lexeme {baban} (*OED3*, n.d., s.v. *baby*). The Welsh {babi} is listed by *GPC* as a borrowing from English {baby}, in

turn. However, basic familial lexis of this kind with reduplicated phonemes such as *ba*, *ma*, and *da* is difficult to pin down etymologically because of the high likelihood of spontaneous development of such forms from early infant vocalizations (Durkin, 2014).

The earliest evidence cited by GPC for {babi} in Welsh comes from the sixteenth century, whereas the earliest known use of English {baby} dates to the late fourteenth century (OED3, n.d., s.v.). This fact alone is not sufficient to say that the English word precedes the Welsh, however. In a language with a relatively sparse written record, any given lexeme is less likely to appear swiftly in writing or print once it has been introduced in the spoken language. The *English Short Title Catalogue* (n.d.) lists only 21 printed works in Welsh before 1600, most of them religious. E. Rees's (1987) *Libri Walliae* lists a somewhat larger number in the same period: 32 printed works in Welsh, most of them religious or didactic (along with a further 13 works with a Welsh connection, but written either in English or in Latin). In contrast, there are over 12,000 ESTC records for works in English before 1600, covering a wide variety of genres. Welsh was represented in print orders of magnitude less frequently than English in this period, as it would be in later centuries. The massive asymmetry that exists between sources for historical Welsh and English means that establishing the history of a Welsh word and its relationship to English cognates and counterparts must be approached cautiously.

At the opposite end of the historical development of Welsh, a new problem arises in identifying ELUs and assigning them an unambiguously English origin: the presence of so-called *internationalisms* in the vocabulary.¹ Leaving aside the exclusionary, Eurocentric way in which the term is often applied in practice (Piirainen, 2008), the concept of the internationalism is useful for exploring the modern history of English loanwords in Welsh. Görlach (2003) explains the dilemma posed for the lexicographer by neo-classical internationalisms, in particular, which use Latinate or neo-Greek elements. These morphological resources are held in common across much of Europe, so a lexeme made of neo-classical elements, like {telephone} or {platypus}, might reasonably be coined in any of a number of

¹ *Internationalism* is probably a rendering of German *Internationalismus*, a linguistic term for a word "das in gleicher Bedeutung und gleicher oder ähnlicher Form in verschiedenen Kultursprachen vorkommt" (which occurs in different Standardized Languages with the same meaning and the same or similar form) (Duden online, n.d., s.v. *Internationalismus*). *Kultursprache* has no direct English equivalent. It connotes a language with a tradition of sophisticated literacy, connected to a defined (perhaps, though not necessarily, national) culture.

European languages. Such lexemes cannot be firmly traced to an originating country unless there is accompanying historical evidence of their origin and spread (Görlach, 2001).

Welsh is nonetheless somewhat different from any of the sixteen Indo-European languages surveyed by Görlach. Modern Welsh shares geographic and political space with English, and is what we might call a bilingual language: there are now extremely few adult Cymrophones who are not also fluent Anglophones (with the possible exception of a few thousand Welsh-Spanish bilinguals in Patagonia). Modern Welsh exists not hand-in-hand but thought-by-thought with English, within the linguistic systems of its speakers. Welsh also has a subordinate political and cultural status to English. As a result, Welsh writers who have commented on the phenomenon of code-switching and borrowing in Welsh have tended to take for granted that the words polluting their language come from English, as we shall see later in this chapter.

In 2007, a reviewer in the popular literary journal *Y Traethodydd* offered the opinion that Welsh has made English loanwords so much its own “nes gwneud inni deimlo ein bod yn siarad Cymraeg glân, gloyw”² when using them (E. L. Jones, 2007, p. 59). In contrast, in his autobiography of 2010, the broadcaster John Meredith expresses his fury over the use of English words when Welsh synonyms are available (J. Meredith, 2010). He singles out the noun {sialens} (from English {challenge}) and its Welsh synonym {her}, claiming that the former has almost entirely replaced the latter in use. Despite the presence of {sialens} in Welsh since at least the sixteenth century (*GPC Ar-lein*, 2024), Meredith cannot accept it as anything other than an English word. In contrast, {her} is, in his view, unquestionably an ILU, although *GPC Ar-lein* (2024, s.v. heriaf) states the possibility that it is derived from the English verb {harry}. Etymology and/or longevity of use are clearly not always sufficient to determine an item’s status; users of any language may be quite unaware of the lexicographic niceties of the words they use and may therefore use long-established loanwords with no awareness that they are not ILUs (Poplack, 2017).

² that we feel that we are speaking clean, clear Welsh”.

1.ii What is code-switching?

There have been suggestions for finer-grained terminology than simply “code-switching” to describe language-mixing phenomena in bilingual speakers. In particular, some authors prefer to distinguish code-switching from code-mixing: the former refers to intersentential switching, or the use of different languages for alternating sentences; the latter to intrasentential switching, in which lexical units of both languages are used in the same sentence (Bokamba, 1988; Di Sciullo et al., 1986; Stell & Yakpo, 2015). The motivations for making this distinction vary: some definitions rely on sociolinguistic considerations, others purely on structural differences (Boeschoten, 2001). This fact suggests the difficulty of making an objective distinction between various types of language-mixing phenomena. In what follows, therefore, I have chosen to use *code-switching* as an umbrella term to refer to any phenomena in which there is alternation between languages within a single act of discourse, without attempting to make sentence boundaries a specific definitional criterion.

1.ii.a Frameworks and formulae: The mechanics of code-switching

Code-switching found one of its earliest academic champions in Gumperz (1977), who was a proponent of the theory that code-switching has specific communicative functions and conveys metalinguistic information about the utterance being made. Counter to the then-prevailing notion of code-switching as a marginal phenomenon of linguistic interference, Gumperz argued that code-switching was not idiosyncratic, but followed certain rules relating to the speakers’ relationship and the social context, and performed specific discourse functions (Gumperz, 1977). Following Gumperz, there has been widespread agreement that code-switching is rule-governed, well-formed language (Poplack, 1980; Stadthagen-González et al., 2019) and that it performs functions beyond simply presenting propositional meaning (Amuzu, 2012; Pfaff, 1979; Poplack, 1980).

There is a vast literature dealing with grammatical models of code-switching, as well as with experimental approaches to testing those models. Since theories about the structural rules governing code-switching are not of direct interest to this study, they will not be discussed in detail here. A few key authors and themes of relevance to the literature on Welsh-English code-switching and to this study’s focus on

language as a semiotic property will be briefly highlighted.

Work on typological and structural aspects of code-switching in Welsh-speaking bilinguals was pioneered by Deuchar (2005a, 2005c, 2006), based on small corpora of conversations between Welsh-English bilinguals. Subsequent research has largely used the *Siarad* (n.d.) (Deuchar et al., 2014) spoken corpora of Welsh-English and Welsh-Spanish bilinguals to break new ground in our understanding of Welsh in contact with other languages (Couto et al., 2015; P. Davies & Deuchar, 2014; P. G. C. Davies, 2010; Deuchar & Stammers, 2012, 2016; Deuchar et al., 2007, 2018; Stammers, 2010; Stammers & Deuchar, 2012). These studies, whose findings will be discussed in this and later chapters, respond particularly to the theoretical approaches established by Poplack (1980) and Myers-Scotton (1993).

One of the earliest theorists to suggest universal constraints on code-switching was Poplack (1980), who posited that switching could not occur below the level of the lexeme. Her Free Morpheme Constraint asserts that forms displaying a root morpheme from one language combined with a bound morpheme or morphemes from another must be classed as (morphologically-integrated) borrowings. The category of *nonce-borrowing* is therefore associated with the work of Poplack (Poplack, 2012; Poplack et al., 1988; Sankoff et al., 1990), accounting for single ELUs which occur very rarely in a corpus, but which (unlike code-switches, according to this typology) are morphosyntactically integrated. The status of these items is, inevitably, a matter of particular contention in attempts to distinguish code-switching from borrowing.

The concept that the languages involved in code-switching make an asymmetrical contribution gained prominence in the work of Joshi (1985). According to his conception, a sentence has a single main language, the *matrix language*, and any other languages which contribute to the sentence are subsidiary, or *embedded languages*. The notion of asymmetry, as well as Joshi's terms *matrix language* and *embedded language*, would be taken up by Myers-Scotton (1993) in her Matrix Language Frame (MLF) model, the most developed and most influential asymmetrical model of intersentential code-switching (Bentahila & Davies, 2001). It is based on the premise that, in the mixed-language productions of bilinguals, there is always a matrix language (Myers-Scotton, 1993). This language is exclusively responsible for the morphosyntactic frame of a sentence. Content morphemes are provided by both the matrix language and the embedded language(s), and are combined with syntactic morphemes from the matrix language to form full clauses

(Myers-Scotton, 1993). This approach differs markedly from that of Poplack (1980), whose constraints apply equally to all languages involved in a code-switched utterance (Bentahila & Davies, 2001).

According to Myers-Scotton, the matrix language is identifiable by two characteristics: it is the source of particular types of functional morphemes; and it controls the word order of the clause (Myers-Scotton, 2002). Other theorists have suggested different criteria by which the primary language of a sentence can be identified. (Di Sciullo et al., 1986) identify the matrix language of any maximal projection as the language of the lexical item which asymmetrically c-commands all other lexical items. In contrast, Treffers-Daller (1991) defines the primary language of the sentence far more simply, as the language of the sentence's finite verb. There exists no single, accepted definition of the primary or matrix language in code-switched utterances (Wasserscheidt, 2020).

Nonetheless, bilinguals often seem able to recognise instinctively how their languages contribute to mixed-language discourse. Kamwangamalu and Lee (1991), for example, found that English-Chinese bilinguals could intuitively assign a matrix language to code-switched sentences. The researchers expressed the concept of a matrix language in two ways: participants were asked "to determine the language of each of the sentences they heard", or to indicate whether "they felt that the sentence was English-based or Chinese-based" (Kamwangamalu & Lee, 1991, pp. 250–251). The concept of "matrix language" is therefore implicitly defined as the assignment of a "language" property somewhere around the level of the clause. Participants appeared to find higher-level syntactic features such as word order and clause structure more salient than lexical features in the process of assigning a language to a sentence (Kamwangamalu & Lee, 1991).

The idea that languageness is a property of linguistic units or structures recurs in code-switching research. Di Sciullo et al. (1986) propose to understand code-switching constraints from the perspective of generative grammar, as constraints on switching between two elements related by government. In basic terms, they argue that, where X governs Y, Y must have the same language property as X. Joshi (1985) argues that sentences as a whole have a language property (determining the matrix language of the sentence), and that the acceptability of switching the language of a particular item depends on its position in the parse tree structure.

An alternative to the models discussed thus far are approaches which assume no structural constraints on code-switching other than those belonging to general

human linguistic competence. Wasserscheidt (2014, 2016, 2020) and MacSwan (1999), for example, both understand code-switching as the result of general linguistic competence. Their theoretical frameworks differ: Wasserscheidt uses the principles of Construction Grammar (Fillmore, 1988; Fillmore et al., 1988), whereas MacSwan bases his model on Chomsky's (1993) Minimalist program. However, both argue that linguistic units, whether conceptualized as constructions or as lexical items, contain all the necessary constraints to determine how they may be combined into larger syntagms. Ungrammaticality in code-switching, just like ungrammaticality in monolingual use, is a result of incongruence between the parameters of linguistic items in contact (MacSwan, 1999).

Language users “do not produce languages or codes, they produce linguistic units” (Wasserscheidt, 2016, p. 2). The properties of linguistic units in the individual's mental lexicon is what determines their ability to fill particular structural slots. These properties may vary from user to user, from dialect to dialect, from register to register, and so on. The language does not impose rules upon its units; rather, the properties of units, taken in aggregate, characterize the language. Accordingly, “language and grammar are nothing other than generalizations over usage” (Wasserscheidt, 2016, p. 3), a conception which fits well with the connotative semiotic model which will be established in the following chapter. Language emerges when a linguistic community consistently agrees to use a coherent set of signs for communication.

A comprehensive theory of code-switching must be capable of accounting for its observed individual and social patterns as well as its structural properties (Pfaff, 1979; Poplack, 1980). Theoretical models of code-switching syntax may delineate the bounds of what is technically possible in code-switching, but the kinds of code-switching that will actually occur in a given context are also influenced by sociolinguistic and individual variables (Sebba, 1998).

This section will explore some of the evidence for psychological, personal, and sociocultural factors which exert pressure for or against code-switching. Sociocultural constraints operate at a large scale on communities of language users. Code-switching is observed to be the unmarked linguistic mode for young Cymrophones, used frequently in informal conversation (Evas et al., 2017), particularly by those who are balanced bilinguals (Musk, 2010). As we shall see later in this chapter, however, Cymrophone culture has often responded negatively to the prevalence of code-switching.

Language users are not units of a linguistic swarm intelligence, but rational

actors operating in specific contexts. They are therefore capable of personal linguistic choices to deploy or suppress language-mixing (Myers-Scotton & Bolonyai, 2001), whether those choices are automatic responses to psycholinguistic stress (Myers-Scotton, 1993) or conscious decisions. It is necessary to consider these various kinds of pressure holistically, in order fully to understand why and how particular instances of code-switching occur.

1.ii.b Processing and proficiency: psycholinguistic pressures

Robust theories of code-switching must have the flexibility to account for individual differences in code-switching behaviour, even amongst members of the same speech community (Sebba, 1998). Some grammatical models of code-switching have been criticized for failing to account for asymmetric bilingualism, in which users have markedly different proficiency in their languages (Assenza, 2016). Jake's (1998) analysis of a language learner's interlanguage as a kind of code-switched discourse offers a useful perspective. Jake (1998) positions the target language as an *intended* Matrix Language, with the interlanguage as a *de facto* Matrix Language into which ELUs from the learner's other language(s) are embedded to fill gaps (Jake, 1998).

Poplack (1980) found that intra-sentential code-switching, the most stigmatized kind of code-switching, was the kind that was most produced by balanced bilinguals, and seemed to require the most skill in both languages. Supporting Poplack's (1980) findings, relative age of acquisition of the two languages seem to affect Welsh-English code-switching patterns. Welsh-English simultaneous bilinguals have been found to produce more multi-word and single-word code-switching than either English-dominant or Welsh-dominant bilinguals, even when the acquisition of the later-acquired language in the latter two groups had begun as early as four years of age (Deuchar et al., 2016).

Although code-switching may not necessarily denote a lack of skill or an imbalance in linguistic resources, this does not rule out the possibility that it is motivated by lexical lacunae or the greater availability of particular vocabulary in one language than in the other, as Amuzu (2005) argues for Ewe-English code-switching in Ghana. Huang and Milroy (1995) found that Cantonese-English bilinguals' use of different code-switching styles and strategies was correlated with language preference and bilingual competence. The need to counter folk-linguistic ideologies of code-switching as laziness or lack of proficiency should not mean disregarding the potential of code-switching as a valid strategy to maintain fluent linguistic

production.

The activation of languages is argued to differ between balanced bilinguals, who are capable of activating a “monolingual mode” in either language, and asymmetrical bilinguals, who can activate monolingual mode only in their dominant language (Assenza, 2016). For balanced bilinguals, there are minor neural costs to processing spoken code-switching, but with no measurable effect on recall (Pérez & Duñabeitia, 2019), and with exceedingly brief delays that may be eliminated entirely if the switch site is a syntactic boundary (Grosjean, 2008). Comprehension of code-switching can be challenging, with increased potential for ambiguity. Bilinguals apparently use the distributional properties of words as an aid to processing, with patterns that are more common in the language community also being easier to process (Tamargo et al., 2016). Nonetheless, regular code-switching experience seems to be necessary in order to take full advantage of this processing aid (Beatty-Martínez & Dussias, 2017).

Greater processing costs are incurred in both production and comprehension when switching from a subsidiary language into a dominant language than when switching in the reverse direction, suggesting that the more strongly activated language is subject to accordingly stronger inhibition (Declerck & Grainger, 2017). Total deactivation of a language may not be possible for any bilinguals, even in a supposedly monolingual mode; this entails that both languages are co-activated to some extent in language production, even for asymmetric bilinguals using their dominant language (Friesen et al., 2016; Rossi et al., 2018).

Bilinguals converge on each others’ choice of code in co-operative discourse, introducing code-switching in a language when another speaker has spoken in that language (Scheu, 2000). Users are disposed to reuse the structure of recently processed sentences, a priming effect which includes a tendency to reproduce the matrix language used in the immediately preceding utterance (Fricke & Kootstra, 2016). This supports the finding that words already produced by the user in a given language contribute to the activation of the same language for subsequent words, although without entirely inhibiting the non-target language (Lauro & Schwartz, 2017). Priming effects on code-switching have been found to be weaker in low-proficiency than in high-proficiency bilinguals (Kootstra et al., 2012).

1.ii.c Sentiments and swearwords: Emotional pressures

Lying somewhere between psycholinguistic and sociolinguistic approaches to code-switching is the question of the emotive force and functions of code-switched

utterances. Code-switching is regularly employed in the expression of emotive language (Agarwal et al., 2017; Lantto, 2014). Bilinguals often report different emotional reactions to speaking different languages (Dewaele, 2016b), so it is reasonable to imagine that certain types of sentiment may seem easier or harder to express in a given language. Code-switching can function as a mediator of emotional engagement, for example by indicating whether or not the speaker identifies with the message, or whether they are citing opinion or fact (Gumperz, 1977).

Code-switching has also been documented as an important part of humour (D. C. S. Li, 2000), storytelling (Auckle, 2017) and, particularly in the case of children, play (de León, 2019; Paugh, 2019). The qualitative analysis of data in Chapter 6 will show the significant potential of language-mixing for constructing humour in Cymrophone discourse.

Code-switched taboo words in an additional language may contribute to humorous discourses (Fägersten, 2017), and are sometimes considered more permissible than non code-switched swearing (Auckle, 2017). Code-switching from a majority language to swear in a minority language may be used emblematically, to resist discourses of minoritized languages as fossilized cultural artefacts, to resist purism in language revitalization efforts, and to challenge common myths that swearwords do not exist in indigenous or minoritized languages (Muehlmann, 2008). The perceived emotional impact of taboo words is strongest in a user's primary language, and decreases in additional languages even for highly proficient users (Dewaele, 2004b, 2010). As a result, swearing in an additional language may distance the user somewhat from the taboo intent or impact of the swearwords (Adaros & Tironi, 2017). Swearing tends to use languages that were acquired early and are used regularly, and people tend to use their dominant language for swearing (Dewaele, 2004a, 2010; Rudra et al., 2016).

The importance of emotional expression to individuals' communicative competence has been advanced as a reason why taboo language should be taught in additional-language programs (Finn, 2017; Horan, 2013). Since additional-language users often fail to perceive the emotional force of taboo words accurately, their presence in classroom instruction provides opportunities to give explicit instruction in the pragmatics of such words (Adaros & Tironi, 2017; Dewaele, 2016a). Dewaele's (2018) account of additional-language speakers of English and their use of the English taboo word *cunt*, in particular, is a reminder that sociopragmatic competence can be very hard to acquire, even for otherwise highly proficient language users.

1.ii.d Society and status: Sociocultural pressures

Some bilinguals stigmatize code-switching as a sign of lack of attention, poor education, bad manners, or lack of linguistic skill; perhaps as a result, they often underestimate the extent of code-switching in their own speech (Gumperz, 1977). Those who have purist attitudes against code-switching are, according to some studies, less likely than others to employ code-switching (Treffers-Daller, 1991). Ratings of code-switching examples in grammaticality judgement tasks correlate with the raters' attitudes to code-switching; those who disapprove of code-switching give lower grammaticality ratings than those who approve of it (Badiola et al., 2018).

Code-switching may become integral to a linguistic construction of both individual and group identities, in a process which the next chapter will define in semiotic terms as enregisterment (Agha, 2005). Code-switching into a language associated with high-prestige contexts such as pedagogy, government, or international institutions may demonstrate the user's alignment with social values such as education, political power, social class, and/or wealth (Amuzu, 2013). Code-switching into English from South Asian languages has been identified as a means of marking one's high socioeconomic status and alignment with modernity, while code-switching into Sanskrit is a marker of religion and caste (Kachru, 1978; Sridhar, 2008). Code-switching from Kannada employing Perso-Arabic lexical items may mark the speaker as unrefined (Sridhar, 1978). In parallel with the successful deployment of English as a sign of status, there is also a tradition in Indian and Sri Lankan cinema of using unsuccessful code-switching into English as a source of humour (Dissanayake, 2008), making those who do not have the elite's access to "good" English the butt of jokes.

In situations where one language involved in code-switching is a minority language, that language often symbolically becomes the in-group, informal code for bilinguals (Pérez-Sabater & Moffo, 2019), with the majority language characterised as an out-group, impersonal, formal code (Gumperz, 1977). Nonetheless, as Chapter 2.iii.a will show, a minority language acquired and used primarily in education may instead become enregistered as the formal, impersonal code, especially if use of the majority language is punished, making it attractive as an in-group code symbolizing rebellion against authority.

The direction of switching is potentially important: a user who chooses to switch from in-group to out-group code may be doing something different from one who switches in the opposite direction (Gumperz, 1977). Even code-switching involving the same languages with the same direction of switching may produce

different outcomes, depending on factors including proficiency. Demonstrating proficiency in the EL through code-switching can be a means of structuring an in-group of code-switching users, while producing code-switching that demonstrates incomplete acquisition of the EL (i.e. code-switching that sounds “foreign”) may exclude someone from that in-group, however much they try to signal alignment through code-switching (Zenner & Van De Mierop, 2017).

1.ii.e Conversation and communication: pragmatic and discursive pressures

Because pragmatic aspects of language are generally culture-specific and easily misunderstood by outsiders, bilinguals will often avoid code-switching with other bilinguals until they have come to know them (Gumperz, 1977). Community norms also affect whether code-switching is perceived as the unmarked communicative mode, or as a marked practice with significant pragmatic weight (Gardner-Chloros et al., 2013). While code-switching can create distrust by excluding outsiders, it can also be used to negotiate knowledge sharing in globalized environments such as multinational companies (Ahmad & Barner-Rasmussen, 2019). Even within individual institutions, nonetheless, a variety of different conventions for multilingual practices may obtain (Wodak et al., 2012).

Although it may be a source of misunderstanding or tension, code-switching may also be used to reduce ambiguities, by introducing translations or elucidations (Gumperz, 1977; Kachru, 1978). The shape of developing discourse may also be clarified by code-switching, with information structure and turn control functions such as focussing (Karrebæk, 2003), addressee specification, and interjection (Gumperz, 1977). Emblematic switching, mainly of easily-switched items such as fillers and tags, allows less-proficient users to participate in the language community (Poplack, 1980).

In postcolonial contexts, users can employ code-switching to perform hybrid local and global identities, as a form of resistance to monolingualism and cultural homogeneity (Bautista, 2004; de León, 2019; Sarkar & Winer, 2006).

Code-switching by Welsh-English bilinguals is largely a phenomenon of Cymrophone discourse, with scant presence of Welsh as an embedded language in utterances where English is the matrix language (Carter et al., 2011; Deuchar & Davies, 2009; Deuchar et al., 2016). Nonetheless, the literary representation of code-switching into Welsh in predominantly English-language texts has received

attention as a postcolonial strategy to alienate the non-Cymrophone reader from full participation in a novel's bilingual milieu (Bohata, 2004; Sheppard, 2018).

The choice to represent the realities of bilingual language use, including code-switching, may be seen as political; yet, as Bohata (2004) argues, the choice to elide linguistic differences and write bilinguals as though their language use is naturally monolingual is no less political. Code-switching in Welsh literary texts has traditionally been frowned upon, reflecting wider cultural attitudes to code-switching in everyday speech (Sheppard, 2018). Its use in modern Welsh writing is therefore an interesting point of tension between tradition and linguistic reality, signifying resistance from at least some quarters to the artificial monolingualism of literary Welsh. Code-switching has the potential to be multilayered, allowing bilinguals to deploy their linguistic resources strategically and to create new meaning in the dialectic between languages.

1.iii What is borrowing?

A review of the research shows code-switching to be a fundamentally contingent and contextual phenomenon. Code-switching, however it is defined, involves decisions made by individuals to deploy resources belonging to different languages together in a coherent discourse. Borrowing, in contrast, is a phenomenon of autonomous linguistic units and their place in a language system, rather than one of individual choices. As a consequence, approaches to studying borrowing are generally less interested in factors such as emotion and cognitive state, focusing instead on the systemic factors which affect the incorporation of ELUs into a language.

1.iii.a Alteration and assimilation: mitigating the (in)congruity of borrowed words

Processes by which loanwords may be assimilated into a language are at the heart of this study's interest in the languageness of English-origin items in Welsh discourse. Degree of assimilation is often one of the key indicators of an ELU's role or status in the IL. In modern Japanese, for example, unassimilated loanwords, assimilated loanwords, and established loans form lexical strata which relate to different registers and styles of the language (Itô & Mester, 1999). Once the process of assimilation is fairly far advanced, it is possible for ELUs to be indistinguishable from ILUs even for

those who speak the IL as a primary language (Poplack, 2017). A discussion of prior research on orthographic adaptation is postponed until Chapter 4, since it is the main focus of this study. An brief overview of other types of adaptation will be given here.

Morphosyntactic integration includes integration of nouns and adjectives into the target language's gender and number systems, and conjugation of verbs according to the target language system (Poplack, 2017). This generally involves changes to the wordform itself, but morphosyntactic integration can also involve pairing the loanword with elements which are appropriate for the IL syntax, for example the addition of obligatory IL determiners with EL nouns (Poplack et al., 1988).

Poplack and Dion (2012) assert that ELUs that retain EL affixes are not morphosyntactically integrated, illustrating their argument with English ELUs that retain English plural marking when used in Quebec French. In Welsh, some established loanwords are commonly made plural in Welsh using a borrowed form of the English plural morpheme. This is usually written in Welsh as <s>, or as <ys> in cases where the English morpheme would be pronounced [ɪz]. The ELU {bocs} (from English {box}), first attested between the sixteenth and seventeenth centuries, has the plural {bocsys} (*GPC Ar-lein*, 2024, s.v. “bocs(2), bocys(2)”), and {tomato}, first attested in 1851, has {tomatos} as its most usual plural (*GPC Ar-lein*, 2024, s.v. “tomato”). The fact that English-origin plural morphemes <-s> and <-ys> are listed in *GPC* and treated there as loans into Welsh complicates Poplack and Dion's (2012) claim: if these EL inflectional morphemes have themselves been absorbed into the IL, then it is not clear whether their use with EL roots should be considered as a mark of a lack of integration. In the data analysis in Chapter 6, however, the English plural morpheme is treated as an unambiguously English inflection, in line with prior research.

Borrowed verbs pose particular difficulties for morphosyntactic assimilation, since verbs often have a complex range of syntactic functions and are often marked for features such as person and tense (Muysken, 2000). Observed accommodation strategies for verbs are varied, and include the direct assimilation of the EL verb using IL morphemes, the use of IL auxiliaries, light-verb strategies, and nominalization of the EL verb (Muysken, 2000; Wohlgemuth, 2009). The ultimate aim of all these strategies is to make the lexeme congruent with the EL system; or, at least, to make it congruent enough that it no longer stands out as alien.

Phonological assimilation also depends on the congruity of EL and IL systems. It may involve minimal changes to the ELU, or changes so radical that the ELU becomes essentially unrecognizable to a monolingual speaker of the EL. Loanwords

which are less fully assimilated often inhabit more peripheral areas of the lexicon, sometimes alongside items such as onomatopoeias, swearwords, and exclamations, all of which are subject to a reduced set of phonological constraints compared to items in the core of the lexicon (Itô & Mester, 1995).

The form taken by phonological constraints has been theorized in a variety of ways. Broersma and Hamann (2009) argue that the production of an ELU based on an underlying EL phonological form involves phonotactic rules, auditory-phonetic cues, and evaluation of the surface form's faithfulness to the EL. Work in the field of Optimality Theory (OT) also highlights faithfulness, positioning loanword phonology as the result of faithfulness and markedness constraints in interaction (Itô & Mester, 2001; Prince & Smolensky, 2002). Faithfulness constraints are those which attempt to ensure that the IL form of a loanword's phonology matches the EL form as closely as possible. The OT framework sets these against IL markedness constraints, which attempt to ensure that none of the IL's phonological or phonotactic rules are broken. The strength of faithfulness relative to phonological markedness is what determines differences in assimilation between ELUs in different strata (Itô & Mester, 1999, 2001). The higher-ranked faithfulness is in relation to the markedness constraints, the fewer phonological adaptations will be made to the ELU (Kiparsky, 1982). In Chapter 2.ii, the notion of faithfulness will recur in Agha's (2009) model of bilingual language mixing, which he imagines as a process of assessing linguistic units' fidelity to the various language systems in play.

In some cases, ELUs may be produced according to the constraints of more than one stratum by different speakers, or in different linguistic registers (Itô & Mester, 2001), suggesting that the level of loanword assimilation can have pragmatic functions rather than simply being the result of fixed and immutable processes in a speech community or in a single user's idiolect (Itô & Mester, 2001). Relative faithfulness to language-specific norms emerges as a resource for meaning-making in itself. I will argue in Chapter 6.iii.b.1 that research on code-switching and borrowing should pay greater attention to the precise forms taken by ELUs, rather than conducting analyses primarily at the level of the lexeme. This allows adapted, unadapted, and partially adapted variants of the same lexeme to be treated as distinct objects, rather than as undifferentiated instances of the same object of analysis.

When ELUs in unadapted form do not match the phonotactic rules of the IL, hearers who are unfamiliar with the EL may experience auditory illusions which mean that they never perceive the actual EL phonological form (Durvasula et al., 2018). Apparent phonological integration may therefore actually be the unavoidable

result of the user having little experience in using the EL phonology (Poplack & Sankoff, 1984). Faithfulness constraints are at their strongest in bilinguals who have both IL and EL as primary languages, since they are the most able to perceive the EL phonology accurately and evaluate the IL form against it.

Parry-Williams (1923) outlines sound changes in pre-twentieth-century English-origin ELUs in Welsh that result from differences in phonetic realization of phonemes shared by the two languages. He notes that Tudor portrayals of Welsh English have the Welsh systematically realizing /b/ in English words as [p], and suggests that English misperception of the “more forcibly pronounced” Welsh /b/ is responsible (Parry-Williams, 1923, p. 15). Indeed, devoicing of consonants is a characteristic ascribed to Welsh English in a number of Early Modern satires. ap Morgan (1642), for example, begins with the words “Cood People, tis written in the Pook of Kot” (i.e. *Good People, Tis written in the Book of God*), showing devoicing of English /g/, /b/, and /d/.

A number of older English ELUs that have been assimilated into Welsh show precisely the sound changes being lampooned here: *potel* (“bottle”), *coblyn* (“goblin”), *tropyn* (“drop”) are all examples offered by H. Lewis (1931). These changes may be the result of perceptual errors, but might instead result from the root forms being interpreted as having undergone soft treigladd, so that the established loanword forms are those with the “treigladd” reversed. It is unclear, therefore, whether such forms should be considered as phonologically or morphologically adapted to Welsh.

When the phonology of the EL and the IL are similar, phonological adaptation becomes harder to measure, since constraints ensuring faithfulness to the EL form and those reducing markedness in the IL form may overlap. Hickey (2009) argues that distinguishing loans and switches in Modern Irish according to phonological integration criteria is extremely difficult, since Irish English accents, even in monolinguals, developed under the influence of Irish phonology. Speakers therefore tend to have convergent phoneme inventories in the two languages. The same is true for the majority of Cymrophones (Fowkes, 1959; Mayr et al., 2017), meaning that English-origin ELUs will often require few phonological adaptations by speakers with Welsh English accents. Nonetheless, Chapter 6 will show that English-origin ELUs in Welsh may reflect a Prestige Southern-England English (PSEE) pronunciation, rather than a Welsh-English one, a fact which is reflected in the standard orthographic forms of many ELUs.

1.iii.b Superfluity and supplementation: the (in)dispensability of borrowed items

Borrowing processes are governed by a number of factors, some inherent to the structural typology of the IL and EL, others relating to culture and textuality. As I have already argued, borrowing as a theoretical construct differs from code-switching in that it focuses on the language system, rather than on individual acts of language in use. Motivations for borrowing are, in consequence, most often imagined at level of language as a communal property. Nonetheless, individual motivations may also be influential if they are shared widely across a language community, since these may well translate into system-wide changes.

Some taxonomies make a distinction between different types of loanword based on whether the lexeme's referent is wholly new to the IL or whether a close synonym already exists. Myers-Scotton (1993), for example, describes "cultural" borrowings, which enter the IL lexicon as labels for objects and concepts that are new to the IL culture, as opposed to "core" borrowings, which enter as additional labels for objects and concepts that have pre-existing IL labels. In general, bilinguals have fairly equal access to core synonyms in their languages: a Welsh-English bilingual is likely to have easy access to both {pig} and {mochyn}, for example.

The distinction between insertional, additional, and replacement borrowings (Monaghan & Roberts, 2019) is similar to that between core and cultural borrowings, but focuses on the position of the borrowing once it enters the IL rather than on its status as a synonym for an existing ILU. Insertional borrowings are ELUs that fill lexical gaps in the IL. Additional borrowings have an IL counterpart, and come to co-exist with it in the IL as synonyms or near-synonyms. Replacement borrowings replace their IL counterpart, and come to be the accepted IL term in its place (Monaghan & Roberts, 2019).

ILUs differ in their susceptibility to replacement or supplementation with ELUs. Various factors determine how stable an ILU is. Words that are used more frequently in a language community seem to be more stable than infrequently used words, for example (Monaghan & Roberts, 2019). Words that tend to be acquired early by children in a language community are also more stably represented in the community (Monaghan, 2014). Consequently, replacement or additional borrowing is more likely to occur when the ELU corresponds to a less-frequent, later-acquired ILU (Monaghan, 2014; Monaghan & Roberts, 2019). For a given concept, the more synonyms used to express it in a language, the higher its lexical replacement rate,

meaning that any lexeme with a close synonym is less stable than one without (Vejdemo & Hörberg, 2016). The length of an ILU may also affect how likely it is to be the target of replacement or supplementation by an ELU, although the precise effects vary between languages (Monaghan & Roberts, 2019).

1.iii.c Purism and power: the sociolinguistics of borrowing

The extent to which ELUs are accepted by a language community is highly context-dependent. As we shall see, the use of ELUs in Welsh is bound up with ideals of identity and authenticity, which go far beyond the kinds of linguistic constraints on borrowing described above. Indeed, ELUs are a matter of contention in many minoritized languages. In some cases, users of minoritized languages may develop practices of hypercorrection, and prefer to neologise using IL morphological resources rather than accept ELUs into their language (Kroskrity, 2014). On the other hand, neologism in minority languages may also be a feature of standardized ‘new speaker’ varieties, which are largely transmitted through institutions such as schools and adult learning centres. In such cases, borrowings will be characteristic of the “traditional” varieties, transmitted through family use, but often not supported by formal education and literacy in the minoritized language (Hornsby, 2015).

Borrowing from prestige languages, like code-switching into those languages, can be a way for a speaker to project characteristics such as education, modernity, and high social class (Zibin, 2019). Unlike code-switching, which requires bilingual skill at some level, borrowing can fulfil this purpose even for language users who have no proficiency in the prestige language. One reason for the unreliability of phonological integration as a test for the loanword/code-switch distinction is that users may consciously choose to retain EL phonology, even on established loans, if doing so is a marker of social prestige (Scotton & Okeju, 1972).

Although the prestige of the EL can mean its easier acceptance by IL users, loanwords from a language that enjoys social prestige will not automatically be accepted. The language itself may be associated with status and advancement, yet simultaneously resented for its influence on the IL, for example, (Rüdiger, 2018). Attitudes towards loanwords may also vary according to demographic characteristics such as age, gender, and education level, or according to the domains in which the loanword is used (Michnowicz et al., 2018). This can lead to a cline of acceptability of ELUs, which encourages their enregisterment (Agha, 2005) as markers of particular linguistic styles. In such contexts, ELUs represent a meaning-making

Table 1.1 Etymologically differentiated lexical strata in English

Latin	French (incl. Anglo-Norman)	Germanic (incl. Old English, Old Norse, Dutch, etc.)
select	decide	choose
query	question	ask
fallacy	error	mistake
perimeter	border	outskirts
devoted	loyal	steadfast
conclude	judge	deem
prior	in advance	before
feces	ordure	shit
pudenda	vagina	cunt

resource which can be deployed by users of the IL to signal a change in linguistic context.

The lexical strata of English illustrate the semiotic possibilities offered by the languageness of lexical units. Latin-origin words tend to mark a legalistic, ecclesiastical, scientific, or other very formal register. French-origin words may also be legalistic and rather formal, as well as characterising artistic, culinary, and political genres. Germanic-origin words tend to be informal, everyday words, and may even be considered rude or obscene (Görlach, 1997). Such connotations may vary diachronically, as the relative status of the source languages changes, as well as synchronically between communities. Nonetheless, the relationships between synonyms or near-synonyms in English drawn from different sources are often strikingly uniform in their stratification, as Table 1.1 shows.

Not all of these words could be classed as ELUs in English. The words of Germanic origin, in particular, are often ILUs which developed during the Old English period, when English was still a comprehensively Germanic language in both lexicon and grammar. Nonetheless, these examples of origin-based lexical stratification in English show that denotative meaning is not the only aspect of ELUs which determines their semiotic potential.

(Parry-Williams, 1923) surveys historical attitudes towards loanwords in Welsh,

concluding that earlier writers were largely concerned to maintain the homogeneity of Welsh vocabulary, even as words from English were pouring into colloquial speech. H. Lewis (1931) describes the influx of English borrowings into Welsh as a natural result of close cultural contact with a more powerful neighbour, and criticizes those who call it a sign of linguistic weakness and national degradation. He notes also that, while ELUs may abound in spoken language, they are nowhere near as frequent in the literary language. This implies that in Welsh, as in other languages, ELUs can be characteristic of register variation.

1.iv Code-switching and borrowing: a comparison

We turn now to the problem of distinguishing code-switching from borrowing, which is a recurring issue in the literature (Adamou, 2016; Clyne, 1987). While code-switching can involve strings of arbitrary length, only lexical units with a discrete meaning (from bound morphemes up to fixed phrases) are traditionally considered as borrowings (F. W. Field, 2002; Gumperz, 1977; Macalister, 2006). Since long strings of ELUs are not the focus of this study, however, this criterion is not of significant interest here, and the focus will be on drawing distinctions between the borrowing and code-switching of individual lexico-semantic units.

Recent work has acknowledged the difficulty inherent in separating code-switching from loanwords, and highlights the fact that different criteria for these categories may produce different experimental results for apparently identical word forms (Nguyen, 2018). Code-switching and borrowing may be understood as a continuum of phenomena, without clear boundaries distinguishing them (Myers-Scotton, 1992). Alternatively, they may be seen as clearly distinct categories, operating under different constraints (Sankoff et al., 1990). In either case, various criteria have been advanced to distinguish the unambiguous code-switch from the unambiguous borrowing.

1.iv.a Flagging and fluency: individual and sociocultural distinctions

Pfaff (1979) is one of a few theorists who consider the ELU's status for the individual user, rather than its position in the IL system as a whole. This approach recommends a focus on the speaker's internal state as she uses an ELU, considering particularly

which language she imagines the ELU belongs to. Pfaff's (1979) work provides one source of inspiration for this study's theoretical focus: the concept here called "linguageness" reflects, in part, her idea that a user has an internal sense of the language to which the lexemes she uses belong. As Haspelmath (2009) notes, however, a full focus on a user's internal state would require access to the mental lexicon of individuals at the time they use the ELU, which is not generally possible when looking at large-scale linguistic data.

Poplack (1980) interprets the linguistic context of ELUs as a sign of their status for the individual user, arguing that code-switching is characterized by fluency and a lack of hesitation phenomena. An opposing claim is that code-switches are *more* likely than loanwords to be accompanied by certain flagging phenomena, such as correction of the form (M. C. Jones, 2005). Where flagging can be interpreted as an apologetic or face-saving discourse move, the increased use of flagging with code-switches would fit with the tendency for code-switching to be deprecated by the community.

1.iv.b Adaptation and acculturation: differences in integration

According to a common formulation, as already mentioned, code-switches maintain the separate morphophonemic features of the two linguistic systems, whereas loanwords tend to become morphosyntactically and phonologically integrated into the target language (Di Sciullo et al., 1986; Grosjean, 2008; Poplack & Sankoff, 1984; Quintanilla, 2018). Phonological integration has been shown to be less stable than morphosyntactic integration (Sankoff et al., 1990), with even long-established loanwords showing variable patterns of pronunciation, particularly in situations where it is socially desirable to retain EL phonology (Myers-Scotton, 1992). As a result, it is not always included as a criterion for linguistic integration of an ELU.

Critiques of this approach have commented on the binary, all-or-nothing approach to integration, in which a word is either a fully integrated borrowing, or an unintegrated code-switch (Deuchar et al., 2018). In describing the analysis of data for this study, I will also argue that the all-or-nothing approach is less useful than one which considers that different instances of the same lexeme may have different statuses. In the most basic terms, I will propose the possibility that one way of spelling a lexeme may best be understood as a borrowing, and a different way of spelling the same lexeme may best be understood as a code-switch.

The Morphological Integration Hypothesis (Myers-Scotton, 1992) understands

morphological integration as a continuum: more-integrated items will be subject to more of the IL's morphological processes, and less open to co-occurring EL morphological processes. Nonetheless, the Matrix Language Frame model predicts that any lone ELU will be syntactically integrated into the IL, as well as somewhat morphologically integrated, regardless of its status as code-switching or borrowing (Myers-Scotton, 1992). If this is so, morphosyntactic integration cannot be considered as a particularly useful distinctive criterion for the kind of single-item ELUs which are the focus of this study.

1.iv.c Semantics and synonyms: functional distinctions

An alternative approach to distinguishing types of ELU argues that borrowing mainly occurs deliberately, in order to fill lexical gaps, whereas code-switching is unconscious and not strongly triggered by the lack of a lexeme in one of the speaker's languages (Bokamba, 1988). Cultural borrowings, which fill lexical gaps, have therefore been described as an extreme end-point of the code-switching/borrowing continuum (Myers-Scotton, 1993). Since they are the only possible lexemes available to encode a particular object or concept, they must enter the IL abruptly as loanwords. Core lexemes, in contrast, are seen as occupying a liminal space between the code-switching/borrowing endpoints (Myers-Scotton, 1993).

While, on the face of it, this may be a useful distinction, in practice it may be difficult to draw a clear line between core and cultural borrowed lexemes, particularly in cases where users of core ELUs are unaware that they are not ILUs. In addition, there is plentiful evidence, some of which has already been discussed in this chapter, showing that bilinguals may use code-switching deliberately and consciously for a variety of purposes. Bokamba's (1988) framing of code-switching as unconscious does not account for this evidence.

The question of whether an ELU is truly synonymous with an ILU at the time of borrowing, or whether instead it fills a lexical gap that is not readily apparent from considering only the denotational semantics of the lexemes in question, is sometimes very difficult to answer. For older loans in particular, it requires detailed information about when exactly it entered the language, and the state of the IL vocabulary at that time, including the historical precise semantics of both the ELU and any semantically similar ILUs. Some of the problems with exploring historical varieties of Welsh have already been discussed above.

Myers-Scotton (1993) points out that there may be psychosocial differences

between core lexemes from two different languages, even if they encode the same basic concept. The fact that two words might be assigned as translations of each other in a bilingual dictionary does not necessarily imply that they are interchangeable to speakers, particularly when one considers their connotations and metalinguistic functions rather than their simple propositional content. As we shall see in the analysis of this study's data in Chapter 6, ELUs and ILUs which are, to all intents and purposes, synonyms do not behave identically in modern Cymrophone texts, a fact which may not in all cases be attributable to their etymological origin.

1.iv.d Documentation and dissemination: frequency, widespreadness, and listedness as distinguishing criteria

The final set of distinctive criteria for categorizing ELUs relate to their visibility in the language community, and these criteria are prominent in the literature examining ELUs in Welsh. An overview of the criteria will therefore be offered here, with a fuller discussion and critique deferred until this study's methodology is described (Chapter 5.vii.b.1).

Increased frequency of use is one of the primary criteria used to classify ELUs as borrowings in the Poplackian tradition of code-switching research (M. C. Jones, 2005; Macalister, 2006; Poplack & Sankoff, 1984). This increased frequency is interpreted as indicating social integration into the language of the community (Poplack & Sankoff, 1984). There is some question as to what exactly constitutes "frequency". Both Deuchar et al. (2018) and Poplack (2012) criticize as arbitrary Myers-Scotton's (1993) definition of a frequent form as one that occurs more than three times in a fairly large corpus. They are certainly correct that this would be a badly underspecified definition, if it went no further. However, Myers-Scotton's discussion of frequency is more nuanced than these summaries of it would imply. The actual frequency criterion she suggests is that of *relative* frequency, measured against the occurrence in the corpus of ILUs encoding the same concept as the ELU under investigation (Myers-Scotton, 1993). Notably, she argues that the three-occurrence cutoff, taken alone, "has little meaning" (Myers-Scotton, 1993, p. 207).

Whether or not an ELU is considered widespread is similarly a factor in need of operational definition. The clearest definition of widespreadness is that of Poplack et al. (1988), who require that an ELU be used more than ten times by more than ten different speakers in their corpus. (This requirement can, of course, easily be scaled to

account for corpora of different sizes.) Poplack (2012) criticizes Stammers and Deuchar (2012) for failing to add this dimension to their categorization of frequency.

If there are monolingual speakers of the IL (or, at least, speakers who are not also speakers of the EL) using the ELU, it indicates that it has spread outside the community of bilinguals. This is a factor which fairly unambiguously categorizes it as a borrowing (F. W. Field, 2002; Myers-Scotton, 1993). This is, of course, no longer a workable approach to Cymrophone ELU use.

The attestation of an ELU in IL reference works may also be used as a measure of its widespreadness and its integration into the language community (Deuchar et al., 2018; Poplack et al., 1988). Deuchar et al. (2018) link the literal listedness of an ELU in dictionaries to Muysken's (2000) psycholinguistic conception of listedness as the presence of the word in the communal mental lexicon. Since the latter construct is difficult (if not impossible) to access, they argue that "the community dictionary may be expected to be a guide to listedness" (Deuchar et al., 2018).

It is clear that a speech community must have a shared mental lexicon, with broad similarity in how entries in that lexicon are stored by individual members, since communication using language would be impossible if this were not so. It is less obvious, however, that the concrete listing of an ELU in a dictionary implies that it is already abstractly listed in the community lexicon. That is to say, there may be many factors other than actual usage by the language community that drive the inclusion of a word in a dictionary or other lexical reference. This, and other, problems with the construct of listedness in previous work on Welsh-English bilingualism will be discussed in chapter 5.vii.b.1. A more robust measure will be proposed, and subsequently evaluated using data from this study's Twitter corpus in chapter 6.

1.v "New language practices"?: Language, languaging, and the translanguaging turn

The distinction between code-switching and borrowing assumes that bilinguals somehow treat lexical units differently based on the languageness of those units. When a bilingual code-switches, consciously or not, she changes languages, switching from the language of the discourse to another language. When she uses a loanword, there is no such switch; the ELU is treated as an entry in the IL lexicon.

If bilinguals were not able to distinguish the languages they use, and to decide which units belong to which language, there would be no way to explain why, in

many communities, code-switching is stigmatized while loanword usage is accepted. Although community members may disagree about where the boundaries lie which distinguish a code-switch from a borrowing, it is clearly the case that bilingual communities very often recognize different categories of ELU, and have specific norms proscribing or permitting their use in various contexts.

The ascription of languageness to an ELU is evidently a complex process. Over time, if an ELU is often used by the community in ways that suggest it should be treated as part of the IL lexicon, more and more users will come to treat it as such. It is even possible that processes of phonological and morphological adaptation may disguise the lexical unit's origin, so that at least some language users have no awareness that they are using an EL. However fluid the boundaries between languages may be, however, in order to define practices such as code-switching and borrowing there must be some conception of languages as entities distinct from one another. A borrowing is only a borrowing if there are two separate languages between which lexical units are transferred. A code-switch must have separate languages between which to switch. Translanguaging, one of the major theoretical movements in recent academic linguistics, has challenged this notion of discrete languages, however.

The term “translanguaging” is generally credited to the research of C. Williams (1994), who coined the Welsh term *trawsieithu*. This described classroom practices in which Cymrophone pupils were encouraged to engage with learning materials using both their languages, such as “*darllen yn y naill iaith, mewnoli'r wybodaeth a'i hatgynhyrchu yn y llall*”³. Williams argued that this would help pupils to develop their proficiency in reading, synthesising, and reproducing knowledge in both languages. Learning in this way requires “*safon weithredol uchel o ddwyieithrwydd*”⁴ (C. Williams, 1994, p. 262).

Williams presented his argument as relating particularly to “*yr iaith leiafrifol o fewn system ddwyieithog o ddysgu*”⁵ (C. Williams, 1994, p. 263). The minoritized status of Welsh meant that there would often be a serious imbalance in the availability of learning materials in the two languages. Instead of seeing this as a disadvantage,

³ *reading in the one language, internalizing the information and reproducing it in the other*

⁴ *a high operational standard of bilingualism*

⁵ *the minority language within a bilingual system of education* (the final word, “*dysgu*”, may mean either “teaching” or “learning”)

nonetheless, Williams suggested that it could offer pupils an opportunity to develop their academic Welsh-language capabilities as well as encouraging a more thorough understanding of the materials.

The central early theorist of translanguaging as a critical practice, in the form in which the term has become widely accepted, is Ofelia García (e.g. García & Li, 2014; García & Lin, 2017). Her elaboration of the term, in discussion with Williams and his Welsh colleagues (C. Baker, 2019) focused on the bilingual's hybrid linguistic practices as a means of "mak[ing] sense of their bilingual worlds" (García, 2009, p. 42). Although the popularity of translanguaging has introduced Williams' term to a wide audience, the original pedagogical meaning of the term, as well as its specific relevance to minority-language contexts, has largely been lost.

Translanguaging, as distinct from Williams' trawsieithu, is imagined as an inclusive, flexible mode of linguistic existence, which reacts to the needs and requirements of the environment (García & Li, 2014; García & Lin, 2017, e.g.). Through this translanguaging lens, bilingual language use is not assessed using the standards and rules of monolingual language varieties, but rather is understood as a unified competence in which each of the individual's linguistic resources informs and is informed by each of her others (García, 2009). As such, translanguaging is an attempt to shift the object of focus: rather than taking individual languages as the focus (as García (2009) contends the concept of code-switching does), it focuses instead on the language user. Translanguaging understands the behaviours often described as "code-switching" as a concurrent mobilization of linguistic resources, rather than one of switching between discrete competences (García, 2009).

Later developments in translanguaging theory have muddled the waters considerably from the position found in García (2009). In one of the key texts of this more recent theorization, translanguaging is made up of "*new* language practices" (García & Wei, 2014, p. 21) which elide the differences between named languages and instead make use of the bilingual's full range of linguistic resources without differentiation. In contrast, code-switching has been described as "monoglossic" in outlook (García & Lin, 2017), identifying codes as separate systems in order to study how they are combined in bilingual utterances. This is not, on the face of it, much of a departure from García's earlier formulation. However, García and Wei (2014) explicitly reject an understanding of translanguaging as a synthesis or hybrid language practice, and assert that it is a practice that creates entirely new modes of discourse. The distance of this conceptualization from that of García (2009) is clarified when we consider how the earlier work situated translanguaging as an

elaboration of C. Williams's (1994) *trawsieithu* and as encompassing code-switching, while García and Wei (2014) explicitly rejects *trawsieithu*'s conception of individual language competences being mobilized in concert, and situates translanguaging as a separate phenomenon from code-switching.

Like Bhatt and Bolonyai (2019, 2022) and Auer (2022), I question the validity of "translanguaging" as a term distinct from "code-switching". Translanguaging's claim to identify language practices in which distinctions between codes are invisible to the language user has not yet been able to offer any evidence for this supposed invisibility. Indeed, claims about how translanguaging is used as a discursive or sociolinguistic strategy rely on the concept that different codes serve different functions for the speaker and her audience (Bhatt & Bolonyai, 2022). If the translanguaging individual is unaware of the separate codes she is using, it is difficult to see how translanguaging can simultaneously be described as a strategic language practice (Bhatt & Bolonyai, 2019, p. 72).

One recent summary of the Welsh origin of *trawsieithu* describes it as a term "from the Welsh, *trawsieithu*, and coined by Cen Williams" which "offers a new vision and practice of language learning and instruction that does not separate the two languages as independent of each other" (Rodriguez, 2017). This clearly misrepresents the original intention of the term, which was not to elide differences between languages but to emphasize them by encouraging deliberate activation of each language for separate purposes in learning. Because C. Williams's (1994) original work is written in Welsh, *trawsieithu* has become accessible to the majority of readers only in English-language formulations by other authors. Williams' insight into the pedagogical value of deploying a bilingual's languages strategically, so that each language supports the other, has been lost.

The identification of languages as discrete objects is an issue with implications beyond academic linguistics. The establishment of discrete geographic territories is inextricably related to projects of language standardization, often founded upon purist linguistic ideologies (Goddard, 2011, p. 41). Political formulations of language are based, not upon the mutual intelligibility of varieties, but upon the identification of those varieties with particular ethnic, cultural, and social groups (Kordić, 2007), as the case of the language varieties variously called Serbo-Croat, Serbian, Croatian, Bosnian, and Montenegrin illustrates (Kordić, 2007; Phillipson & Skutnabb-Kangas, 2013).

Colonial projects worldwide have used language as a tool of control and expansion, thrusting standard varieties of European majority languages (now, chiefly,

English) to the top of the linguistic hierarchy and devaluing local languages (Phillipson & Skutnabb-Kangas, 2013). Where attention has been paid to indigenous languages, it has usually been in order to impose cultural ideals of literacy and religious morality (Phillipson & Skutnabb-Kangas, 2013). One of the ironies of linguistics as an academic discipline is that the most wide-ranging source of information about endangered languages is *Ethnologue* (2022), produced by the Christian missionary organization SIL International (formerly the Summer Institute of Linguistics) (Dobrin & Good, 2009). SIL's main goal is to facilitate the production of Bible translations in endangered languages, and the organization's presence in indigenous territories challenges traditional linguistic practices and knowledge systems, with a focus on developing literacy for Christian religious purposes rather than as a response to existing community needs (Bodley, 2015; Dobrin & Good, 2009; Stoll, 1982).

Theoretical positions such as those proposed in translanguaging, nonetheless, have little to offer by way of supporting minoritized languages and their users. The logical result of framing languages as intellectual or political fictions is a rejection of policies to protect minoritized languages (Tamburelli, 2021a), invalidating communities whose identity is based, in part, on linguistic affiliation (Phillipson & Skutnabb-Kangas, 2013). If there are no languages, distinct one from another, then there is no object in need of protection, and no shared characteristic by which to define a community. As we shall see in Chapter 3.i.b, the position of language as a marker of group identity is bound up, in modern sociopolitical discourse, with notions of nationalism. The rejection of ethnic nationalism as an illiberal, often racist, ideological framework further threatens to undermine the formation of communal identity for users of minoritized languages.

Resistance to monolingualism in applied linguistics, and the consequent promotion of linguistic diversity and flexibility, is seen by some as merely a new way to frame traditional linguistic imperialism. In practice, "multilingualism" often means making space for one or more (former) colonial languages, and using one of these languages as a lingua franca for convenience (Phipps, 2019). The burden of language learning, and of communicating and working in a second or other language, falls disproportionately on those from linguistic minority communities. Non-dominant languages may become siloed, as the elaboration of function and scope associated with use in science or politics is reserved for the majority languages.

The establishment of English as the language of law, politics, and academia has had, and continues to have, a significant influence on the development of the Welsh

language. One of the motivations for C. Williams's (1994) development of trawsieithu was that Cymrophone pupils would inevitably be expected to use English-language materials, because of the lack of materials in their language of education. Translanguaging, purportedly a development of Williams' work, in fact makes a nonsense of his pedagogical approach, as well as undercutting attempts to support the expanded use of minoritized languages, by denying that the separation of languages is coherent in theory or acceptable in practice.

Languages are, in a fundamental sense, socially constructed. This does not mean that languages have no existence except as imagined constructs, however, or that bilinguals engaged in language-mixing practices are simply deploying resources from an undifferentiated linguistic competence, with no conscious awareness that these resources belong to separate languages. As Auer (2022) argues, languages are social constructs not in the sense of being imaginary, theoretical abstractions, but in the sense that there is some real society of people using that language regularly as a means of communication.

Languages exist for their users embedded in social context. As a result, "the bilingual repertoire is not unstructured but includes cognitively stored information about who to talk to, when and in which 'language', i.e. which feature constellation to select in a given situation" (Auer, 2022, p. 135). The next chapter will examine enregisterment as a theoretical explanation for how such "feature constellations" come to be, and explore how a semiotic model of language as a meaning-making resource reveals all aspects of human linguistic communication, from entire language systems to the phonemes and graphemes which are the systems' minimal units of expression, as imbued with languageness.

1.vi Code-switching and borrowing: the Welsh perspective

Recent scholarship using the *Siarad* (n.d.) spoken corpora of Welsh-English and Welsh-Spanish bilinguals has broken new ground in our understanding of Welsh in contact with other languages (Couto et al., 2015; P. Davies & Deuchar, 2014; P. G. C. Davies, 2010; Deuchar & Stammers, 2012, 2016; Deuchar et al., 2007, 2018; Stammers, 2010; Stammers & Deuchar, 2012). Much of this scholarship has sought to adjudicate between the approaches of Poplack and Meechan (1998) and Myers-Scotton (1993) to distinguishing code-switching and borrowing.

Deuchar et al. (2018) note that these two primary approaches contain symmetrical, mutually-exclusive hypotheses and definitions of the code-switching and borrowing categories: where Poplack and Meechan use integration as a defining criterion, and analyse frequency of use as a testable hypothesis, Myers-Scotton takes frequency as a defining criterion and level of integration as a testable hypothesis. More fundamentally, Poplack and Meechan (1998) consider code-switching and borrowing as separate phenomena, while Myers-Scotton (1993) argues that they fall on a continuum.

Stammers and Deuchar analyze the relationship between listedness in dictionaries, frequency in their corpus, and morphosyntactic integration (measured via soft treiglad) of English-origin verb-nouns in bilingual clauses where Welsh is the matrix language (Stammers, 2010; Stammers & Deuchar, 2012). The overall conclusion of these analyses is that Myers-Scotton's (1993) understanding of code-switching and borrowing as a continuum is unsupported by the evidence, and that Poplack and Meechan (1998) are right to think of code-switching and borrowing as discrete phenomena. However, Poplack and Meechan's use of linguistic integration as the chief criterion for distinguishing ELUs is found to be insufficient. Instead, code-switches are defined by Deuchar and colleagues as ELUs with both low frequency and low morphosyntactic integration, and borrowings as ELUs with either low or high frequency and high morphosyntactic integration. Borrowings may therefore establish themselves from what were originally code-switches, if they become more frequent and more integrated (Deuchar et al., 2018).

Resistance to the incursions of words from foreign languages is a feature of prescriptivist discourse in many, if not most, languages. Language is commonly viewed as a signifier of group identity (E. Williams, 2009), and adherence to this construct of language leads fairly predictably to purist, conservative attitudes (Wohlgemuth, 2009). With the growth of English as a globalized language, borrowing from English has been a particular source of concern for the linguistic authorities of a variety of languages, including Japanese (Schmidt, 2009), French (Ben-Rafael, 2008), and Dutch (De Vries Jr., 2008).

Unlike these languages, however, each of which is the official and majority language of at least one modern nation-state, and each of which has been imposed outside those borders by colonial rule, Welsh must also be understood in comparison to other minoritized languages, where resistance to other-language influences happens in a variety of different ways. Modern-day speakers of Rapa Nui on Easter Island use a deliberately-constructed purist register of their language for formal

political discourse, in contrast to the colloquial everyday form which is highly mixed with the Spanish of their colonizers (Makihara, 2007). Purism on the part of Navajo elders has been seen as a trigger for insecurity in younger speakers of the Navajo language, which may ultimately lead to a refusal to speak the language at all (M. C. Field, 2009).

The Welsh and English languages exist within a complex sociocultural context in contemporary Wales. As we will see in Chapter 3, Welsh is framed by some as a marker of authentic Welshness, and by others as a symbol of ethnocentric insularity and anti-English bigotry. English is understood not simply as a marker of specifically English identity, but as a symbol of hegemonic Britishness. For some, the English language is a metonym for the British state and its history of control and colonialism. If speaking Welsh is a way to access Welshness, speaking English, even if only a few words, is a capitulation, an admission of one's colonized status.

1.vi.a Provincialism and pilfering: attitudes from outside the Cymrophone community

Social media has brought mockery of Welsh into public view (D. Jones, 2018); what might once have been throwaway comments, made in person and intended only to raise a quick laugh, are now immortalised online and may be “overheard” by an unanticipated audience. Such derisive comments might be dismissed as idle badinage. However, both their ubiquity and the repetition of a small set of consistent themes suggest that comments found on social media reflect myths about Welsh which are culturally pervasive beyond the online realm. These myths depict Welsh as deficient and illogical, and Cymrophones as exclusive and calculating.

One of the most widespread tropes encountered in anti-Welsh popular discourse is the claim that a group of Cymrophones, usually in a community location such as a pub or a shop, switched from English to Welsh when they detected the presence of a non-Cymrophone. The former leader of Plaid Cymru, Leanne Wood, tweeted:

I see that old chestnut ‘they all switched to Welsh when I walked into the pub’ has resurfaced again.

To everyone who claims this – you do realise that it has become such a cliché, people laugh out loud now when we hear this?

Please – keep entertaining us!

(@LeanneWood. December 29, 2018.)

Despite all attempts to debunk them, stories of this kind are repeated regularly, both at second hand and by those who claim it really happened to them.

The repetition of this trope displays a profound ignorance about how bilinguals use language. As we have seen, bilingual linguistic behaviour is complex and varied, and does not necessarily follow monolingual norms, yet those who repeat this trope never consider the possibility that the bilinguals they encounter might not have been engaging in a monolingual conversation at all, but might naturally switch languages even without the presence of a monoglot Anglophone. The English language is positioned as an unmarked default code; Cymrophones are imagined to speak it preferentially unless there is some (nefarious) reason to switch to Welsh. Welsh, as a result, is painted as a tool for deliberate, bigoted exclusion of Anglophones, rather than as a natural medium of communication for Cymrophone people in their own communities.

The conception of English as a default language is reflected in a second common trope, which sees Welsh “criticised by pillocks on Twitter for ‘stealing’ words from English” (E. James & Robins, 2018, p.302). One Twitter user asserts that “The thing with the Welsh language, is that half of it is English because there’s no Welsh word for most things. 🤔 pointless language.” These attitudes are not only found amongst the anonymous voices of social media. The mainstream British press, too, regularly gives a platform to opinions such as those of author Roger Lewis, writing for the tabloid publication *The Daily Mail*, who described Welsh as an “appalling and moribund monkey language [...], which hasn’t had a new noun since the Middle Ages—hence *pwdin* is pudding, *snwcer* is snooker, *tacsi* is taxi and *bocsio* is boxing” (R. Lewis, 2011).

The English language is here assumed to be a kind of linguistic *primum mobile*. Terms such as “taxi” and “ambulance” are considered English words, which the Welsh borrowed or stole because they had no words of their own. The fact that both words came into English from French, and are formed on Latin roots, does not make them any less English in this conception; yet the extra step into Welsh (and particularly the orthographic adaptation this requires) is illegitimized. As a result, Welsh is enregistered as impoverished and old-fashioned, with a lexicon whose generative capacity fossilized in ancient times.

A third often-repeated trope in the mockery of Welsh targets the language’s orthographic system, depicting it as random and nonsensical: “Welsh looks like a cat walked over a keyboard”, and “Welsh looks like someones [sic] just got angry and keyboard smashed” are examples collected from Twitter. Perhaps the most common



Figure 1.1 Tweet from the official account of The Spectator

instantiation of this particular jibe is the claim that Welsh has too few vowels. Fig. 1.1 shows a cartoon tweeted by the official account of conservative current-affairs magazine The Spectator, in which a man complains that “half the vowels are missing” from his Welsh Scrabble set.

Welsh uses the graphemes <w> and <y> to represent vowel phonemes, with <w> also used to represent the approximant /w/. English uses <w> only for /w/, while <y> occasionally represents a vowel, but is commonly used for the approximant /j/. As a result, Welsh words such as <crwydrwr> (*wanderer*) or <trwsgl> (*awkward*) appear, according to Anglophone conventions, to represent sequences of consonants without vowels. The monolingual Anglocentric perspective insists that English is the ultimate arbiter of linguistic conventions. Orthography that deviates from English norms is judged as deficient or nonsensical, rather than as the realization of a different (but equally reasonable) set of rules. The foreignness of Welsh orthography for Anglophones will be discussed in greater detail in Chapter 4.

In addition to showing the tendency to apply an Anglophone perspective to interpretations of Cymrophony, the tropes discussed here display two important assumptions: that languages can be hypostatized as identifiably discrete competences;

and that “normal” language use means consistently using a single language rather than switching between two (or more). This replicates political language ideologies, which will be discussed in Chapter 3.i.b, treating the use of a single, common language as essential to maintaining national unity. The tendency towards reductive ideas of bilinguals’ languages as separate competences, according to which it is reasonable to imagine that Cymrophones normally speak English because it is the dominant language, and in any case always conduct conversations monolingually, fails to match with the reality of bilingualism. As we will see, Cymrophone prescriptivist attitudes towards language mixing, while in many ways quite different from the mocking tropes of non-Cymrophones, rest to a large extent on similar monolingualist ideals.

1.vi.b Servility and snobbery: attitudes within the Cymrophone community

Deuchar et al. (2016) note the general disapproval of code-switching in Cymrophone culture, expressing the hope that their work, showing that code-switching is actually more prevalent amongst balanced bilingual speakers, will encourage those who associate code-switching with inadequate Welsh-language ability to rethink their view. Anti-code-switching norms have similarly been observed amongst Francophone Canadians (Poplack, 1987), for example, although the same study found comparatively more relaxed attitudes amongst Spanish-speaking Puerto Ricans in New York. Censure of code-switching is not inevitable, therefore, but neither is it uncommon.

Bilinguals, even those who employ code-switching themselves, are not immune to external pressures to condemn code-switching. In some minority-language contexts, supportive attitudes towards the minority language and its survival, as well as frequent use of the language, correlate with lower frequency of code-switching (M. C. Jones, 2005), suggesting that code-switching is considered misaligned with support for the language. The use of English-origin ELUs in Welsh has long been condemned, both by prominent authors and in popular discourse. More than three centuries ago Thomas Jones, a prominent publisher of Welsh-language works (Johnston, n.d.), explained his policy regarding English-origin loanwords in the introduction to his Welsh-English dictionary. Setting such words inside parentheses

“iw neilltuo hwŷnt oddiwrth y gymraeg gywir”⁶, he complains that, just as the English are quick to borrow fashionable words from French, the Welsh are “ynfydu am lediaith y Saesnaeg, yn gymmaint hyd oni ddaeth iaith y Cymrŷ yr awron mo’r llygredig ag ymmadrodd eu Cymmydogion”⁷.

In this description, Jones brings conceptions of nationality and authenticity to the fore. Loanwords are “Saesnaeg Candrell”⁸, because they have been adapted to suit Welsh phonology and morphology. They remain English despite adaptation, and distinct from “true Welsh words”. The adoption of English-origin words is understood as a matter of national character, a bad habit picked up from the English, who themselves sought to emulate the fashionable French. Borrowing is presented as vain frippery which undermines the purity of a language.

One of the best-known modern Welsh poems about language loss is *Colli Iaith* by Harri Webb, whose first verse reads:

Colli iaith a cholli urddas
Colli awen, colli barddas
Colli coron aur cymdeithas
Ac yn eu lle cael bratiaith fas.⁹ (Webb, 1995)

The pejorative term *bratiaith* (lit. “language of rags”) is often repeated in purist Cymrophone discourse, and enregisters an English-influenced variety of Welsh as fragmentary and insufficient, patched together from scraps, rather than as a rich and coherent linguistic system. The term “bratiaith” has often, in practice, encoded a particular prejudice against the dialects of the industrialized south of Wales (I. W. Rees, 2015). The denigration of southern Cymrophony persists into the present: comic author Dewi Rhys expresses surprise that anyone would count Welsh from the south as “[C]ymraeg posh’ [...], gan fod geirfa’r hwntws yn llawn geiriau Saesneg!”¹⁰ (Rhys, 2009).

⁶ to distinguish them from the true Welsh [words]

⁷ wild for the English dialect, so much so that the language of the Welsh is now as polluted as that of their neighbours

⁸ broken English

⁹ Forfeit language, forfeit eminence, / Forfeit genius, forfeit eloquence, / Forfeit culture’s crowning brilliance, / And in its place get vulgar cants.

¹⁰ ‘posh Welsh’ [...], since the vocabulary of the southerners is full of English words!

In the words of a Twitter user, “[D]ylid wastad ymwrthod rhag diogi a bratiaith fel idiomau Susnag anghyffredin lle mae myrdd o ffyrdd eraill [sic] i eirio yn Gymraeg”¹¹. Even authors who represent themselves as pedagogical authorities occasionally contribute to this discourse of language-mixing as laziness. Cymrophones were characterized in 1933 by the prominent Welsh educationalist Edmund David Jones (Johnston, n.d.) as “rhy ddiog i wneud dim ond defnyddio’r gair cyntaf a ddaw i’n meddwl, i ba iaith bynnag y perthyn”¹² (E. D. Jones, 1933, p. 49). More recently, in the textbook *Seren Iaith!*, which is aimed at helping younger Cymrophones to “[g]wella sgiliau ieithyddol”¹³, the reader is asked “Ydych chi’n ddiog ac yn defnyddio geiriau Saesneg yn lle rhai Cymraeg?”¹⁴ (Breese & Clement, 2017).

Servility and snobbery were also common moral flaws ascribed by nineteenth- and early-twentieth-century authors to Cymrophones who Anglicized their language. In 1888, M. D. Jones, a prominent nationalist, minister, and principal of the Bala Independent College (Johnston, n.d.), wrote to the periodical *Y Celt* ridiculing the “bratiaith Dic Sion Dafyddol”¹⁵ of men of business and fashionable women, who believe “fod anwybodaeth o Gymraeg yn arwydd o’u rhagoriaeth pendefigaidd”¹⁶ (M. D. Jones, 1888, p. 7). A character in an early nineteenth-century ballad by satirist Jac Glan-y-gors (John Jones) (Johnston, n.d.), Dic Siôn Dafydd was an uneducated Welsh countrysider who succeeded in business in London and pretended to have forgotten how to speak Welsh. “Dic Siôn Dafydd” became a byword for Welsh quislings who, as the Anglican priest and academic Griffith Hartwell Jones (Johnston, n.d.) described them, “ymserchant ym mhob gwlad namyn eu gwlad eu hunain, ac a fynnant sathru dan draed y clawdd terfyn rhwng Cymru a Lloegr, a’i wneuthur yn gŷdwastad â’r llawr”¹⁷ (G. H. Jones, 1926, p. 3).

¹¹ One should always resist laziness and bratiaith like unusual English idioms where there’s a variety of other ways to word things in Welsh.

¹² too lazy to do anything but use the first word that comes to our mind, regardless of the language to which it belongs

¹³ improve linguistic skills

¹⁴ Are you lazy, using English words instead of Welsh ones?

¹⁵ Dic-Sion-Dafyddish patois

¹⁶ that ignorance of Welsh is a sign of their aristocratic superiority

¹⁷ are infatuated with all lands but their own, and who wish to tread underfoot the bulwark between Wales and England, and flatten it to the ground

Maintenance of the boundaries between Welsh and English, as much as those between Wales and England, is represented by many Cymrophone authors as essential to Welsh self-respect. To muddy these boundaries would be to accept that Welshness is inferior to Englishness, and that assimilation into English language and customs is the only path to advancement. The eminent scholar and linguist, Sir John Morris-Jones (Johnston, [n.d.](#)), harshly condemned the moral turpitude of early nineteenth-century Welshmen who sought assimilation:

yr oedd llawer o'r dosbarth canol yn ceisio dynwared y bobl fawr; pob Cymro bron a godai uwch bawd sawdl yn mynnu siarad Saesneg â'i blant, a llawer un mor ddirmygedig â chywilyddio o'i iaith a'i genedl. I'm bryd i, ni all dynoliaeth ddisgyn yn is na hyn¹⁸ (Morris-Jones, [1926](#), p. 30).

Those who seek advancement through English are depicted as despicable snobs, whose social and financial gains cannot make up for their debasement as human beings.

Other authors showed greater sympathy, particularly towards Welsh parents who felt torn between the survival of their language and their children's future security. The lack of official support for Welsh, and the pressure to use English in business, meant that English appeared to be the sole linguistic route to success:

Nid oes, o bosibl, unrhyw Gymro, boed ef mor wladgarol ac mor ieithgarol ag y b'o, a feiddia ddweyd heddyw nad yw gwybodaeth ddeallol o'r iaith Saesneg yn un o anhebgorion llwyddiant bydol.¹⁹ (B. G. Evans, [1889](#), p. 71)

The same author wonders whether the disdain for and neglect of Welsh in education is responsible "am waseiddiwch dirmygus a chywilyddus y Cymro cyffredin yn mhresenoldeb y Sais; am y *bowio*, y capio, a'r syrio, i bob coegyn estronol ddaw i'n

¹⁸ *many of the middle class attempted to mimic the gentry; almost every Welshman who gained a little advancement insisted on speaking English to his children, and many were despicable enough to be ashamed of their language and their nation. To my mind, humanity can sink no lower than this.*

¹⁹ *There is, perhaps, no Welshman, however patriotic and loving of his native language he may be, who would dare say today that an intelligent knowledge of the English language is not one of the essentials of worldly success".*

plith”²⁰ (B. G. Evans, 1889, p. 75). This turns Morris-Jones’s (1926) description on its head, seeing the moral debasement of the Welsh as a result of the abandonment of Welsh, rather than positioning that abandonment as the result of a lack of moral courage.

1.vi.c Saesneg and subordination: the denigration of Welsh English

In the days when Welsh people were largely monoglot Cymrophones, there was much debate about the best way for them to learn English. The abandonment of Welsh by Cymrophone parents in an attempt to turn their children into competent Anglophones was not, on the whole, depicted as a successful process, producing what was condemned as “[t]afodiaith fratiog, nad yw na Chymraeg na Saesneg”²¹ (G. H. Jones, 1926, p. 15), and as “[c]ymysgfa hagr o rywbeth nad yw yn Gymraeg nac yn Saesneg”²² (D. E. Evans, 1897). Writing of growing up in a small Welsh village, C. Williams (2002) describes the English spoken there as “a second language when we had no other language”. Her community’s language is a liminal construct, too Welsh to be English, but with an empty space where, in former generations, the Cymrophony which shaped their English would have stood alongside it as a mother-tongue.

The English spoken by Cymrophones has traditionally not been well regarded by English people either, leading to fears that being raised as a Cymrophone was “anfanteisiol er dysgu Saesonaeg; oblegid y mae yn beryglus i ddyn drwy hyny siarad Saesonaeg gydag acen Gymreig, a chael ei ddirmygu o’r herwydd gan y dosbarth mwyaf diwylliedig”²³ (Môn, 1889). As late as 1938, the Welsh were being reminded by a fellow Welshman “that the Welsh accent in English is a foreign accent” which is “like a French or a German accent—it is foreign and uncongenial to the spirit of the English language” (S. Jones, 1938, p. 234). The Welsh were urged to learn to speak

²⁰ *for the contemptible and shameful servility of the common Welshman in the presence of the Englishman; for the bowing and capping and ‘sir’-ing, to every foreign fop who comes into our midst*

²¹ *a ragged dialect, which is neither Welsh nor English*

²² *an ugly mix of something that is neither Welsh nor English*

²³ *disadvantageous for learning English; for there is a danger that one will, as a result, speak English with a Welsh accent, and be mocked for it by the most cultured class*

like the English, and particularly like “cultured people in and around London” (S. Jones, 1938, p. 233).

By avoiding speaking English with a pronunciation enregistered as Welsh, Welsh people hoped to access the prestige and culture of Englishness. For some authors, avoiding the enregisterment of a Welsh variety of English was a way to keep the two languages distinct:

Pe ceisid dysgu plant Cymru i ynganu'r Saesneg fel y gwna'r Saeson eu hunain, byddai'n llawer haws iddynt gadw'r ddwy iaith ar wahan. Dywedir bod Cymry a sieryd Saesneg gydag 'acen Gymreig' yn cael eu gwrthod weithiau o achos hyn pan geisiant swyddi yn Lloegr.²⁴
(E. D. Jones, 1933, pp. 50–51)

Keeping English as an *English* language, Jones suggests, may be a way to maintain Welsh as a Welsh one.

Cymrophone hostility towards the English language has translated, for some, into hostility towards Welsh Anglophony itself. Non-Cymrophone Welsh Anglophones “are the new strangers, doubly excluded, in Wales and in Britain” (Penhallurick, 1993, p. 31). In the latter decades of the twentieth century, nonetheless, some Cymrophone cultural commentators began moving away from viewing Welsh English as “as a vulgar, Anglicised betrayal of ‘Welshness’” instead asserting “that English is a Welsh language” (R. Williams, 1985). English is increasingly recognized as belonging to Wales, and there is a sense of pride that “nid Saesneg Lloegr yw ein Saesneg ni”²⁵ (Hopwood, 2003, p. 59). The growth in Cymrophones’ confidence in claiming an identity as Anglophone has, nonetheless, gone hand in hand with a growth in insecurity about their Cymrophony.

1.vi.d Innocence and insularity: the moral superiority of Cymrophony

As we will see in Chapter 3.i.a, there has long been a sense that authentic Cymrophony belongs to the rural parts of Wales, whose people are romanticized as

²⁴ *If Welsh children were taught to pronounce English as the English themselves do, it would be much easier for them to keep the two languages separate. It is said that the Welsh who speak English with a ‘Welsh accent’ are sometimes refused as a result when they try for jobs in London.*

²⁵ *our English is not the English of England*

simple, honest, and faithful. “Paham y rhaid i’r rhai sydd wedi dysgu–drwy lafur mawr yn ddiau–air neu ddau o Saesneg mawreddog, eu gwasgu ar ystyriaeth Cymry bach cyffredin?”²⁶ (‘Gweinidog Heb Ofal’, 1903), asks an anonymous correspondent to the radical Nonconformist periodical *Y Diwygiwr* (Johnston, n.d.). The “Cymry bach cyffredin” are depicted as an ordinary, unassuming populace, lacking the kind of social status which can only be achieved by the English, but which the Dic Siôn Dafydd so hopes to emulate with his smattering of pretentious English terms. The representation of English words here not only links their use by Cymrophones directly with the moral laxity described in the previous section, but also by implication divorces the Dic Siôn Dafydd who uses these words from the authentic Cymry.

The enregisterment of Welsh as a language of quiet rural life, strict morality, and old traditions reaches an extreme of a 1977 opinion piece in scientific periodical *Y Gwyddonydd* by Owain Owain, whose opinions about Welsh language and culture had a profound effect on modern Welsh nationalism, as Chapter 3.i will show. Owain claimed that the use of Welsh in science would encourage a more ethical and humanitarian practices, asking: “allwch chi–alla i ddim–ddychmygu gweithgor o wyddonwyr, Cymraeg eu hiaith, yn trafod y dulliau mwyaf effeithiol o wella napalm a nwy C.S.?”²⁷ (Owain, 1977, p. 128). This is, evidently, a purely emotional appeal with no factual or evidential basis, and an extreme example of linguistic determinism. Welsh is imagined as functioning like a morally-beneficial version of George Orwell’s “Newspeak”, preventing the expression and even the thinking of proscribed ideas. Owain’s 1976 dystopian science fiction novel, *Y Dydd Olaf*, which was indebted to Orwell for many of its themes, implied that to maintain the Welsh language was to maintain an ethical humanity in the face of an implacably mechanized and technologized modernity (Morus, 2016; Turner, 2020).

Owain was a nuclear scientist who worked on the production of weapons-grade plutonium for nuclear bombs at the Windscale site in England in the 1950s (*Safle Owain Owain*, n.d.), and who became an outspoken critic of nuclear weapons. His 1977 comments must be read in this context. We might hypothesize that they are an attempt by Owain to deal with his own complicity in the British nuclear programme

²⁶ *Why must those who have learnt–undoubtedly with great pains–a word or two of grandiose English, foist it upon the attention of ordinary little Welshmen?*

²⁷ *can you–for I cannot–imagine a committee of Welsh-speaking scientists discussing the most effective ways of improving napalm and C.S. gas?*

by enregistering English as the language of his scientific persona, leaving a core, Cymrophone persona, which is innocent of complicity because the language itself renders one incapable of expressing such scientific evils.

Welsh is enregistered as an inherently less worldly language than English: the language of rural life, of the innocent “Cymr[o] bach cyffredin”. As such, it is a language in which harmful ideas are hard, or impossible, to express. In its innocence, nonetheless, Welsh is also revolutionary, subverting hegemonic British culture. In this context, Owain’s repudiation of English takes on additional cultural weight, symbolizing a rejection of the deadly industrialization associated with dominant, Anglophone culture.

1.vi.e Partition and proscription: the (im)possibility of bilingualism

Cymrophone discourse of linguistic prescriptivism is rather different from that of British Anglophones. For many Cymrophones, there is a sense that authentic Welshness is the key factor in determining what is “Cymraeg gloyw”²⁸. Welshness is placed on a cline of authenticity, with English influence a prime determinant of one’s position on that cline. To allow English a foothold in one’s linguistic habits is therefore to risk diluting one’s entire identity as a Welsh person.

The twin ideals of English-free Welsh and Welsh-free English lead to rejection of the possibility that the two languages might exist in a creative symbiosis, producing a bilingualism which transcends the resources offered by either language individually. Indeed, total rejection of this possibility has characterized the more extreme opinions on both sides of the debate. Many Victorian educationalists insisted that Cymrophone schoolchildren should encounter and use nothing but English in the classroom, and be forbidden from using Welsh. Opponents of this policy believed that it led to “ofn a diffyg hunan-ymddiried”²⁹, as the child was left without the language to express what was clear in his mind (B. G. Evans, 1889, p. 74). Cymrophone children in English-only schools were essentially silenced, their intellectual ability unexpressed and inexpressible in a language which was not their own.

²⁸ *polished Welsh*

²⁹ *fear and lack of self-confidence*

The English-only policies of earlier schools in Wales are mirrored by Welsh-only policies in modern Welsh-medium schooling (Lee, 2016). During a row about Welsh-only policies in Ceredigion, it was reported that one child had been “told by a teacher to ‘moved [sic] to England’ if she wanted to speak English in the playground” (Misstear, 2012). Attitudes which relate the use of English in Welsh contexts to sociocultural Englishness are by no means new. The county of Pembrokeshire has for centuries been known as ‘Little England Beyond Wales’, because of its linguistic and cultural Anglicization:

a stranger travelling from England, and having ridden four-score myles and more in Wales, having heard noe English, nor English names of people, or of places, and coming hither to Penbroke-shire, where he shall heare nothing but English [...], would think that Wales were environned with England, and would imagine he had travelled through Wales, and came into England againe. (G. Owen & Lewis, 1799)

Monolingual policies in Welsh-medium schools have faced opposition from some parents as “archaic” (Traynor, 2012). Others express the sense that their children are “being ‘admonished’ for being bilingual” (Misstear, 2012). Schoolchildren in a bilingual (Welsh/English) secondary school in Wales interviewed by Selleck (2013) in 1996 reported that speaking English in the classroom would lead to a reprimand: “they say there’s an English school down the road you can go there” (Selleck, 2013, p. 31). Such attitudes are part of a conception of ethnolinguistic identity which imagines the non-Cymrophone Welsh as “ethnically no more than a version of provincial English” (Aitchison & Carter, 1994, p. 3).

1.vii Why does the status of ELUs in Welsh matter?

Whereas Victorian schools expected Welsh children to cease being Cymrophones on school premises, many modern Welsh-medium and bilingual schools are accused of expecting them to cease being Anglophones. The effects of monolingualist policies will be explored throughout this study, and proposals for schooling that celebrates multiple interwoven linguistic identities will be made in the concluding chapter. In order for such potential to be realized, however, it is necessary to understand the ideologies supporting Welsh-only policy, including the (often unspoken) association of Anglophony with Englishness.

The prior literature on code-switching and borrowing surveyed above provides a number of useful starting-points for this study, particularly by establishing measures relating to integration, frequency of use, and community language norms which will inform those described in Chapter 5 and applied in Chapter 6. Unlike Deuchar and colleagues, however, I will take a different standpoint when considering the use of ELUs by Cymrophones. Starting with the attitudes towards language-mixing in Wales surveyed above, and following Clair and Guadalupe (1980) and Lauttamus (1999) and others, I understand code-switching and borrowing as sociolinguistic phenomena, arising from culturally determined choices and behaviours of linguistic communities in contact, rather than as phenomena to be determined primarily by the structural and distributional patterns of EL lexemes.

The focus of this study is on establishing the semiotic functions of English-origin ELUs in Welsh, in particular their relationship to enregisterment (Agha, 2005) and to the construct I have called “*languageness*”. Instead of seeking the characteristics which distinguish a loanword from a code-switch, this study instead asks how an ELU becomes established in the IL lexicon and gains acceptability for users of the IL, losing its “foreignness” and becoming an unmarked, Welsh word. The categories of code-switch and borrowing, interpreted as manifestations of the sociolinguistic property of *languageness*, are not left open as objects to be defined. Instead, they are established here as terms for English-origin ELUs in Welsh which are either interpreted as retaining an English identity, or as having gained a Welsh one.

The remainder of this study therefore begins with the premise that loanwords and code-switches are differentiated primarily by consensus of linguistic communities, and that their differentiation is contextually contingent. In particular, this conception of ELUs accepts the possibility that the same lexeme may have a different status for different people, and/or in different contexts. As Chapter 2.iii.b will show, one and the same wordform may be deprecated by one linguistic authority as an “English” word, and accepted by another as a “Welsh” one. This cannot be explained by structural factors or analyses of integration, but only by paying attention to the highly contested nature both of Welshness and Englishness as linguistic constructs, and of language mixing as a practice. This inevitably involves attention to the social, cultural, and psychological aspects of Welshness and Englishness, since these are the ultimate sources of the conflicts which make language-mixing in Welsh-English bilingualism so contentious.

As a result of this framework, macrolinguistic factors, such as cultural and

pedagogical attitudes towards ELUs, frequency of use within the language community, and the status conferred by dictionaries, are prioritized in understanding how ELUs function. There is perhaps no area of language use which is fully separated from the sociocultural construction of the language community, however.

Phonology, for example, realized on the idiolectal level as accent, can be a potent indicator of social group membership. Potential factors in ELU categorization will not be discounted from this study solely because they sit nearer the microlinguistic end of the scale, therefore. They will instead be integrated into a holistic semiotic model which traces the pervasive influence of sociolinguistic context at all levels of linguistic analysis.

Conceptualizing borrowing and code-switching in terms of the languageness of the ELU in context explicitly rejects the “nonce-loan” as a category of ELU (Poplack et al., 1988), since this category consists of ELUs used spontaneously without prior establishment in the IL community. These words are argued to be loanwords rather than code-switches because they are integrated to some extent into the IL morphological and syntactic system (Sankoff et al., 1990). Conceiving of loanwords as words whose Welshness is accepted, however, makes it impossible to imagine nonce-loans as anything other than code-switches. They are words which remain English despite the presence of the necessary apparatus to make them congruent with IL syntax. Such words are often stigmatized as “foreign” and as a sign of lacking proficiency in the IL, while established loanwords do not face such stigma (Waltermire & Valtierrez, 2019), a fact which supports the overall conception that such ELUs in Cymrophone discourse retain their Englishness.

Using a macrolinguistic lens, a convenient basis for differentiating code-switching and loanwords is to understand them as behaviours of bilingual and monolingual language users, respectively. In this view, loanwords must be a part of the lexicon of the IL, because they can be used by monolingual IL users who are, by definition, not users of the EL (Pfaff, 1979). Such words cannot be the result of “switching”, because that would require the user to have knowledge of both of the codes in question in order to switch between them. The use of loanwords therefore requires only monolingual competence, while code-switching requires at least some degree of bilingualism (Pfaff, 1979).

This is, on its surface, an impractical choice of criterion in the Welsh context, since there are essentially no monolingual Cymrophones whose behaviour may be studied. However, the attitudes expressed by Cymrophones towards what Poplack calls “nonce-loans” mirrors that of the monolinguals cited by Waltermire and

Valtierrez (2019). As we will see, using “English words” in Welsh utterances is deprecated, considered unfit for formal or academic discourse, and may cause primary-language Cymrophones to doubt their ability to speak Welsh at all in official contexts. The idealized, high-status Welsh is a monolingual Welsh. The processes which drive assimilation of English-origin ELUs into Welsh are processes of Cymricization, enscribing and enregistering Welshness. As a result of such processes, the ELU becomes a Welsh word, appropriate for use by an imagined Cymrophone monolingual.

Monolingual adult Cymrophones may now be extremely rare, if any truly exist at all. But the community-level tendency to classify spontaneously-produced ELUs as “English” or as “not-Welsh”, mirroring the use of terms connoting foreignness to deprecate spontaneous ELUs in other contexts (Waltermire & Valtierrez, 2019), reifies a monolingual variety of Welsh as an imaginary standard. Subsequent chapters will show that monolingualism is at the heart of language ideologies which pervade Welsh-medium education as well as Welsh popular discourse and public policy. As Chapter 6 will show, however, monolingualist discourses are at odds with the natural linguistic behaviours of Cymrophones and, more than this, attempt to suppress resources which Cymrophones deploy strategically to enact creativity and humour, and which function as valuable differentiators of different registers in contemporary Welsh.

Chapter 2

“What’s ‘englyn’ when it’s english?”: A semiotic account of languageness

‘Da’ is ‘fine’; ‘pysgod’ is ‘fish’; – ‘isio bwyd’
Is ‘Boy! Am I peckish!’
‘Dwyn’ is ‘steal’, ‘desgil’ is ‘dish’;
What’s ‘englyn’ when it’s english?

(Morys, 2001)

2.i Languageness in the context of enregisterment

Like translanguaging theorists, Agha (2009, p. 253) critiques the “code-centric” model of bilingualism, on the grounds that structural models of code-switching “reveal very little about the social significance” of language-mixing behaviour (including, but not limited to, code-switching and borrowing) Agha (2009, p. 253). Unlike translanguaging, however, Agha’s theoretical model does not require the repudiation of languages as distinct objects. Instead, he encourages a view of bilingualism as “a discursive practice where the *transposition* of speech-tokens across contexts of use [...] influences their ‘type’-level construal” (Agha, 2009, 254, italics in original). In this conception, individual instances, or tokens, of linguistic units in use are evaluated in terms of abstract exemplars, or types, of those units in a language system. Bilingual behaviour often results in hybridized tokens, which combine features drawn from more than one language, and these tokens are evaluated for their faithfulness to the abstract types of more than one system. A phonologically-adapted ELU is therefore interpreted as a token of the phonological types of *both* ELU and

ILU, and its fidelity to both systems is evaluated. Agha suggests that this simultaneous activation of the types of two linguistic systems in the evaluation of a single token is itself a meaning-making resource; the relative fidelity of the phonological form to both ELU and ILU can carry gradations of social meaning (Agha, 2009, p. 255). Where translanguaging blurs bilingual language into a single competence, Agha's model insists upon the simultaneous activation of two discrete linguistic systems which work in concert.

A broad perspective to complement the token-level view of language given by Agha (2009) is provided by Wasserscheidt (2014, p. 309), who concludes that what I have called *languageness* is “a pragmatic construct and a network of co-used elements” rather than a property of individual linguistic items in and of themselves, a description which mirrors Auer's (2022) “feature constellation”. Agha's token-level instances reproduce types, which represent the potential of the language system(s) in play. Because language is made up of nested structures formed from categories of varying levels of complexity, a single token may be an instance of multiple types, representing dimensions including the phonological, morphological, and syntactic, as well as discourse-level structural principles such as coherence. As a result, “acts of token reproduction are fraught with multiple possibilities of nonfidelity to type” (Agha, 2009, p. 256). An ELU token set amongst ILUs has multiple possible ways to display fidelity to either or both languages, and each of these provides a potential site for attaching cultural meaning.

Evaluating a particular token for its fidelity to an abstract type is possible because the type is recognizable as a member of a linguistic system known to the user. If we take this proposal a little further, the ideal example of what is called a “borrowing” by linguists can be characterized as a word which is so faithful to the IL that it ceases to be evaluated according to the EL type-system at all. This is not an impossible target; modern languages often contain words borrowed so long ago that they are no longer obviously non-congruent with other tokens of the IL. The English words “pajamas” (Hindi “पाजामा” - /pɑːdʒɑːmɑː/) and “bungalow” (Hindi “बंगला” - /bəŋɡləː/) are very unlikely to be experienced by modern monolingual Anglophones as uninterpretable; even monolinguals who are aware of these words' etymologies will not be able to evaluate them for fidelity to the types exemplified by the Hindi etymons. In contrast, ELUs which are unassimilated may be “merely diacritics of otherness” when they are encountered by someone who is not familiar with the EL (Agha, 2009, p. 255). They can only be evaluated against the type information available in the IL, which essentially results in their labelling as non-language for the purposes of meaningful

interpretation.

2.i.a Enregisterment: ascribing social meaning to linguistic features

Agha's (2009) model of bilingualism is situated within his earlier theoretical concept, enregisterment (Agha, 2005): the process "whereby distinct forms of speech come to be socially recognized (or enregistered) as indexical of speaker attributes by a population of language users" (p. 38). In turn, enregisterment is related to the concept of register, originally theorized within the tradition of Systemic Functional Linguistics (SFL) (e.g. Halliday, 1964) to explain the potential of language to make meaning in context, and to use context to make meaning (Lukin et al., 2011). Registers are broadly defined in SFL as "language used for different purposes" (Halliday, 1964, p. 40), and as "functional contexts" (Halliday, 1990, p. 140), which embody and realize aspects of the communicative situation, conventionalizing them as bundles of linguistic features. In some cases, linguistic features become markers of register because they are particularly appropriate to the context; in others, features become associated with the context purely through convention (Biber, 1994). The context may also influence decisions about how to interpret a particular linguistic feature, ascribing different semantic meaning to it in different situations (Lukin et al., 2011).

Register is "a useful fiction" (Halliday, 1992, p. 358); a theoretical construct which forms "the basis on which we theorized the relation of language to social context" (Halliday, 2001, p. 273). Using register as a lens, every use of language can be seen as an enactment of context (Halliday, 1992). The context is understood by (Halliday, 1965, 1994) as a complex of associations between the subject-matter, the participants and their relationships, and the linguistic medium, as well as more abstract characteristics which describe the role played by language, for example for persuasion, instruction, or information.

The meaning potential of register is imagined as an outcome of users' experience of the language system. "What can be meant [...] is very much a product of history: of what could be meant, and what has been meant, before" (Halliday, 1992, p. 358). As language users gain detailed experience of language in use, they develop an intuitive sense of the probability that a given linguistic feature will or will not occur with a given meaning in a given context (Lukin et al., 2011). As new instances of language in context accrue, the patterns they embody influence the communal model

of language, making certain linguistic features more or less salient as markers of the register in question. Registers therefore require the existence of a language community capable of recognizing and deploying them (Agha, 2005).

Enregisterment (Agha, 2005) is the process “whereby distinct forms of speech come to be socially recognized (or enregistered) as indexical of speaker attributes by a population of language users” (p. 38). The “social range” of a register is the set of stereotypical values associated with it (Agha, 2015), which determines the range of personae which can be performed using that register (Agha, 2005). This is complemented by the register’s “social domain”, or the language community which is socialized in the use and/or interpretation of the register (Agha, 2005, p. 39). The representation of people with identifiable social characteristics in the British broadcast media, for example, is interpreted by Agha as a particularly important source of the enregisterment of accents of British English (Agha, 2005). The former leader of Welsh political party Plaid Cymru, Leanne Wood, reported continual discrimination against her on the basis of her accent, which is geographically associated with the industrialized Valleys of southern Wales, and was characterized by her abusers as “thick” and “working class” (McCarthy, n.d.; Mears, n.d.; Wood, n.d.). This is a typical example of enregisterment, as a distinctive linguistic feature (in this case, a pattern of pronunciation) becomes stereotypically associated with low intelligence and low social status.

Enregisterment constitutes registers in an ongoing process, adapting to social and demographic pressures. Registers represent a synchronic view on enregisterment; a “snapshot” (Agha, 2015) capturing groups of associated enregistered features at a single point in time. From one point of view, therefore, registers are artefacts made up of enregistered features. From another, they are systems of potential which members of a community can deploy to make meaning in context.

It is possible to think of languageness itself as a matter of enregisterment. Returning once more to the concept of a language as a “feature constellation”, selected according to the communicative situation (Auer, 2022, p. 135), enregisterment offers a theoretical model to explain how it is that communities relate linguistic objects and features together in this way. By experiencing certain linguistic items in use by particular people in particular contexts, language users come to associate those features with particular varieties. Imagining this process to proceed in the same way as enregisterment has the advantage of not requiring that bilinguals should develop cognitive processes which are separate from those ascribed to monolinguals. Instead, seeing languageness as a matter of the enregisterment of

groups of linguistic features which commonly co-occur in particular situations means that using a particular language is the same kind of process as using a particular register.

2.ii Language as a connotative semiotic

In the context of the current study, enregisterment gives us a framework for discussing the use of ELUs in Cymrophone discourse which explains the languageness of these units as an outcome of the same process which also determines whether they are stigmatized or treated as neutral items in the Welsh lexicon. Before moving on to consider how enregisterment is enacted in Cymrophone discourse, it will be useful to contextualize enregisterment as a process within a semiotic model of language. Agha's (2009) description of bilingualism in terms of the fidelity of individual linguistic instances to abstractions belonging to the language system, in particular, must be placed within semiotic model of language in order to be properly understood.

Register is a theoretical construct embedded in a conception of language as a social semiotic system, and associated particularly with systemic functional linguistics (SFL) (Shore, 2015). Enregisterment, too, is situated within a semiotic framework of language (Agha, 2005, 2015). Both SFL and semiotics are well established schools of thought in linguistics; as a result, a number of related, but not identical, models of the language system have been proposed by scholars working within these frameworks. The nature of the relationship between various parts of the system will be central to the analysis of data later in this study. Agha's (2009) description of bilingualism in terms of the fidelity of individual linguistic instances to abstractions belonging to the language system, in particular, requires a coherent model of language in order to be properly understood. The final section of this chapter will therefore offer an outline of language as a connotative semiotic system, explaining how enregisterment functions at all levels of language (Shore, 2015), connecting even apparently non-meaningful units such as the individual characters in a written word to the cultural identity embodied in a whole language.

The model which will be used here designates natural languages as connotative semiotic systems, based in large part on the theoretical models developed by Hjelmslev (1954, 1961) and in the tradition of SFL (e.g. Halliday, 1979; Halliday & Matthiessen, 2014; Hasan, 2013; Taverniers, 2008). Hjelmslev's work was initially

inspired by that of de Saussure (1911), for whom language is ‘un système de signes’¹ (de Saussure, 1911, p. 71). Signs are unified dualities, composed of a signifier and a signified², which (roughly speaking) represent a linguistic unit and the meaning it conveys. One of Saussure’s most important insights is that the relationship between signifier and signified is arbitrary. A particular linguistic form means something purely because a community of users has agreed to associate that form with that meaning.

Hjelmslev’s approach to language is more rigorously schematic than Saussure’s, theorizing the sign as the union of two purely formal categories (Hjelmslev, 1954), rather than a union of a specific form with a specific meaning. Hjelmslev’s abstruse presentation of his theoretical concepts³ means that his work has been interpreted in a variety of ways by later theorists (Taverniers, 2011), and untangling his original ideas from their later representation is not an easy task. Although the model of language adopted in SFL is generally credited as a Hjelmslevian model (e.g. Halliday, 1978; Hasan, 2013; Matthiessen et al., 2010), there are fundamental differences between the two which make them hard to reconcile (Bache, 2010; Taverniers, 2011). This discussion will identify what I believe to be the crux of the incompatibility between the two traditions, taking elements from both to propose a semiotic model of language which is consistent with the notions of register and enregisterment.

2.ii.a Layers and layouts: The stratified nature of the linguistic sign

In a standard semiotic conception of language, there are two types of stratification which separate different aspects of the linguistic sign. The first is the separation of expression, linguistic items that convey meaning, from content, meaning that is conveyed linguistically. These two aspects of the sign roughly correspond to Saussure’s signifier and signified. Hjelmslev (1961, p. 58) understood content and expression to be the two inputs to a “sign-function”; signs come into being when an instance of content and an instance of expression are provided to the sign-function.

The second stratification involves the separation of form and substance, terms

¹ *a system of signs*

² “signifié” and “signifiant” – de Saussure (1911)

³ Hjelmslev’s prose has been described as “more exhausting than exhaustive” (Larsen, 1993, p. 42).

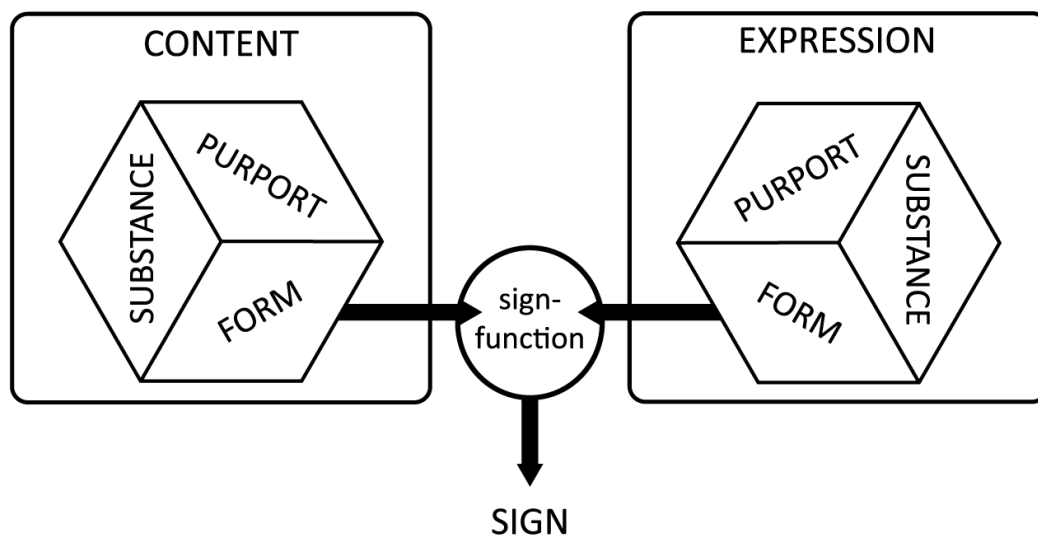


Figure 2.1: A visualization of Hjeltmslev's sign.

which require unpacking. For Hjeltmslev, there are three aspects to both content and expression: purport, form, and substance (fig. 2.1). Purport exists “provisionally as an amorphous mass” (Hjeltmslev, 1961, p. 50): “provisionally” in that it is not accessed except in the process of sign-creation; “amorphous” because it lacks the boundaries which will subsequently be imposed upon it by the language system. Purport is a space of pre-linguistic potential. On the plane of content, we might call purport *the space of everything that could be meant*. In the realm of expression, we might call purport *the space of everything that could be meaningful*. All human languages work with the same purport. Differences in their schemata, which demarcate the boundaries of linguistic objects, account for differences between languages.

Form is, in some ways, the absolute opposite of purport: it is a schema of categorical slots, defined purely by its position in relation to all the other slots in the schema. If purport is disordered “stuff”, form is “stuffless” order. Substance appears as the interaction of purport and form, which produces ordered “stuff”, the empty categories of form characterized by features extracted from purport. The formal schema represents its units' relations to each other, and the projection of this schema onto the undifferentiated mass of purport identifies the potential manifestations of the schema's forms. Substance is therefore a segmentation of purport according to the schema of a given linguistic system (Hjeltmslev, 1961).

We may begin to understand the relationship between schema, purport, and substance by thinking of a grille cipher (Gardner, 1972, p. 62): a sheet with windows



Figure 2.2: Grille cypher as a metaphor for Hjelmslev's "purport"

cut out at set positions which is laid over a text, revealing a hidden message. The text, as pictured in fig. 2.2, represents purport⁴ The purport is a zone of abstract potential, whether of meaning or of expression. It is unformed, but not unformable. A schema (fig. 2.3a), in contrast, is nothing but form. It is a set of empty slots, of a particular shape and in a specific formation. When the schema is laid over purport, it selects particular areas and groups them together as coherent units (fig. 2.3b). In this case, the grille reveals a sequence of words, making up the first line of the Welsh national anthem.

The areas of purport which are selected by the schema are substances. On the expression stratum of language, for example, the schema of a language selects an area of possible speech sounds which can all fulfil the same role in the language's sound system; this is a phoneme, and the set of individual sounds selected by the schematic category are the phones which users of the language recognize as allophones of the same phoneme.

Purport is pre-linguistic, limited only by the constraints of human physiology and cognition. Nonetheless, it exists only to become "substance for one form or another" (Hjelmslev, 1961, p. 52). It is pre-linguistic, yet also semiotic, as implied by the earlier formulations *the space of everything that could be meant* and *the space of everything that could be meaningful*. What could be meaningful and what could be

⁴ We must imagine, for the sake of the metaphor, that the individual characters are actually not discrete, but represent a continuous, unsegmented field.

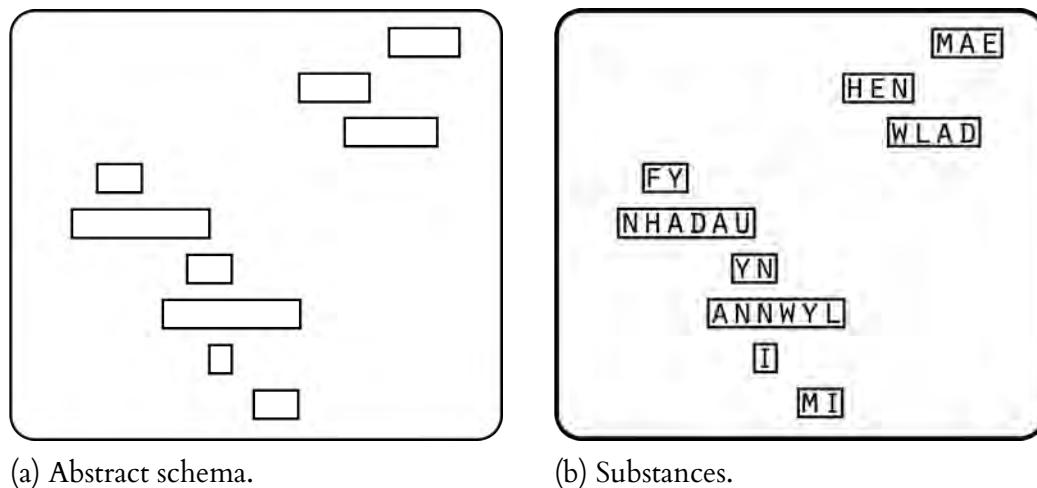


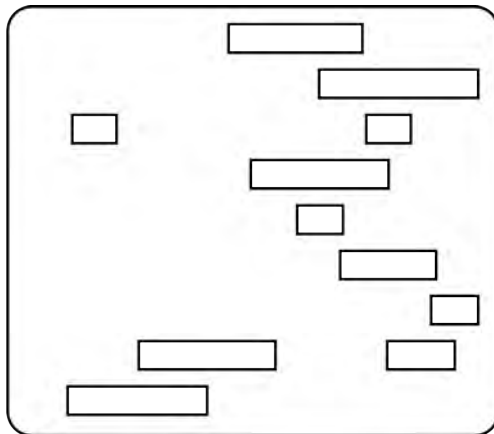
Figure 2.3: Grille cypher: a “schema”, and the resulting “substances”.

meant are not wholly neutral spaces, but ones which are already shaped by the “could be”, the limitations and potential of human capacity.

The schema of a language determines how purport is segmented to form substances, and different languages may have radically different schemata. The schemata depicted in figures 2.4a and 2.4c take the same purport depicted in fig. 2.2, and reveal lines from the French and South African national anthems, respectively (figs. 2.4b and 2.4d). In each case, the same purport, with a different schema, produces a different set of substances.

In some cases, the different schemata select overlapping areas of purport, just as different human languages may have words which share some meanings but not others. The Welsh word “pren”, for example, means the same as English “wood” when referring to a material used for carpentry; unlike “wood”, however, “pren” cannot be used to refer to a number of trees growing together in a dense group. The schematic category which selects purport to form a content-substance which can be matched to the expression “pren” is not of the same extent as the category whose content-substance is matched to “wood”, but the categories do overlap in the area of purport they select. The relationship of a schematic form to a zone of purport is arbitrary. This is because the precise organization of any given language schema is arbitrary. The boundaries around the purport represented by “pren” are not pre-determined outside the Welsh language community; if they were, it would be impossible for English to draw those boundaries differently for the word “wood”.

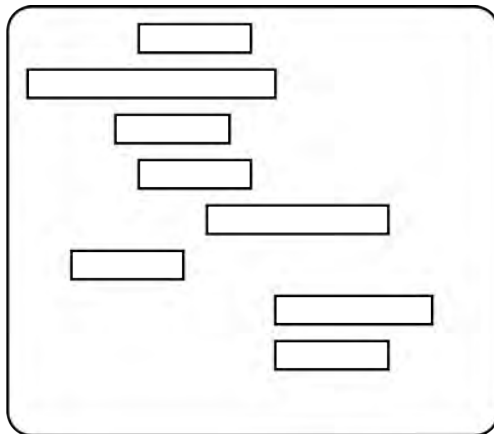
Hjelmslev (1947) credits Saussure as the inspiration for his focus on language as a relational system; the conception of form as a set of category slots which are related to



(a) : A second schema.



(b) : A second set of substances.



(c) : A third schema.



(d) : A third set of substances.

Figure 2.4: Two alternative schemata for the grille cypher, producing two alternative sets of substances.

each other in particular ways resonates with Saussure's notion of *value*⁵. Terms in the system have no positive meaning, in the sense of concrete linguistic features, but rather have “valeur purement négative et différentielle”⁶ (de Saussure, 1909, p. 7). The system need not know what the articulatory or acoustic properties characterizing a phoneme are; it only needs to know that a language recognizes so many phonemes as distinct units. This is why large-scale sound change in a language often manifests as a chain of shifts; as the boundaries of one phoneme move, it impinges on others, which also move in order to maintain their distinctiveness (Giancarlo, 2001; Smith, 1993). Of course, this is not inevitable; change can ultimately manifest as change at the level of formal categories themselves, affecting the shape of the system as a whole (for example when phonemes merge, as in the tense-lax neutralization which is characteristic of many varieties of North American English (Baxter, 2010; Dinkin, 2005)).

For Sonesson (2017), Hjelmslev's model, in its insistence on pure form and its concern with establishing exact symmetry between the two strata of content and expression, gives insufficient motivation for differentiating content and expression as the two inputs to the sign-function. This critique, while valid in the case of natural languages, misses an important point about Hjelmslev's purpose. He is not attempting to describe language in semiotic terms, but to delineate the general properties of semiotic systems, which natural languages happen to be. As a result, Hjelmslev's architecture is deliberately symmetrical: “expression and content are arbitrarily assigned designations for elements that are defined only oppositively and negatively” (Hjelmslev, 1961, p. 120). This model does not attempt to explain why content-forms are *categories that are meant* and expression-forms are *categories that mean*, because this distinction is arbitrary. Expression-forms become *categories that mean* only by their existence as the expression-forms of a particular semiotic system. In different circumstances, they could equally be *categories that are meant*.

2.ii.b Content and construal: how substantial is substance?

The key outcome of the absolute symmetry of Hjelmslev's content and expression is that both halves of the sign must be entirely conceptual. The existence of a phoneme on the plane of expression means the existence of phones which can realize that

⁵ “valeur” – de Saussure (1911)

⁶ *purely negative and differential value*

phoneme. A phone, in turn, can be manifested as a set of articulatory actions creating vibrations that move in a wave through the air. However, the same phone may be the object of discussion in a metalanguage (Hjelmslev, 1961, p. 119), in which case its ultimate exponent is the concept of the phone, not the articulatory movements, the vibration, and the sound wave. Concrete acts of expression are not entirely missing from Hjelmslev's system, but they are external to the creation of the sign. To understand this, we must turn to Hjelmslev's distinction between language as schema, as usage, and as act.

The schema has already been described, and provides the "shape" of the language without specifying anything about its features. Usage is "ce qu'il y a de stabilisé"⁷ of language in use, contrasted with the linguistic act, which contains "tout ce qui est arbitraire"⁸ (Hjelmslev, 1954, p. 188). Usage abstracts information from all the individual acts which together fill the same formal category. A phone, although defined according to its articulatory and acoustic properties, is therefore ultimately made of the same conceptual stuff as purport. Phones can be related to phonemes, which abstract from phones by discarding anything non-definitional. The exact quality of the phone used by a PSEE speaker to realize the phoneme /k/, for example as unaspirated [k] or aspirated [k^h], is not important to its function as a unit of expression; the distinctions between these realizations are arbitrary, in Hjelmslev's terms. However, it *is* important to know that /k/ is unvoiced, since this distinguishes it from /g/.

Concrete linguistic acts manifest etic instances of substance which, in the aggregate, are stabilized as emic substances. The exact nature of the form and substance strata of the expression plane of language will be crucial in establishing the relationship between orthography and enregisterment. It is here, however, that disagreement about the nature of the sign becomes most obvious, particularly in the reformulation of Hjelmslev's model in systemic functional linguistics (Bache, 2010; Taverniers, 2002, 2011).

The conception of content as prior to expression in SFL's model is driven by Halliday's interpretation of expression in material terms, as the "more 'concrete' stratum [...] realizing a second more 'abstract' one" (Halliday & Martin, 2003, p. 29). Hasan's (2013, p. 278) approach to the stratification of the sign, while still consistent

⁷ *the stabilized part*

⁸ *everything that is arbitrary*

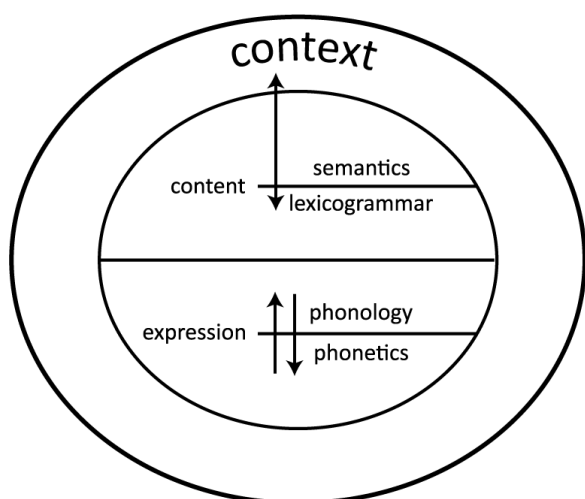


Figure 2.5: Relationship between Saussurean and systemic-functional strata, based on the account in Hasan (2013, p. 272).

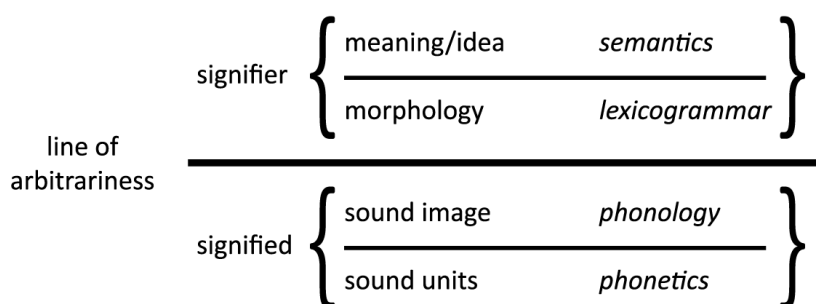


Figure 2.6: The structure of the sign, adapted from Hasan (2013, p. 272).

with the standard SFL model in most ways, refines it by recognizing the line separating content and expression as different in kind from the line separating form and substance within each of those strata (fig. 2.5). Nonetheless, Hasan, like Halliday, understands content as logically prior to expression. The context in which language is used activates semantics (content-substance), which activates lexicogrammar (content-form). Lexicogrammar, in turn, activates phonology (expression-form), which activates phonetics (expression-substance), to produce instances of language.

However, this relationship is not fully reversible (fig. 2.6). When a higher level activates a lower one, in most cases the lower level is said to “construe” the higher; it provides “clues” to the “identity as an instance of a particular category” of the higher-level unit which activated it (Hasan, 2013, p. 279). The word *clues* is telling: the fact that a lower stratum construes a higher stratum allows an addressee to reconstruct the relevant meaning from the evidence provided. The line demarcating

content from expression is a “line of arbitrariness” (Hasan, 2013): construal does not pertain between lexicogrammar and phonology. There are no “clues” in phonology about the nature of the category of lexicogrammatical meaning which activated the phonological form. This is the essence of semiosis: content and expression are arbitrarily related, and we only know that an expression is meant to signify a content because we have learned the rules of the language system, not because there are inherent features of the expression that would permit us to deduce the content from it. Like Sonesson (2017), therefore, Hasan differentiates content and expression by means of the subjectivity of the participants in a linguistic exchange: when we produce language, there is a chain of activation from context, through content, to expression; when we are the recipients of language, however, we must use our knowledge of the language system to make the leap across the line of arbitrariness, because there are no other cues which tell us why particular contents are linked to particular expressions.

Hasan’s identification of the line of arbitrariness is essential to the nature of the sign, and adds important nuance to the SFL model. However, like the standard model, Hasan’s is inconsistent with Hjelmslev’s conception because it imagines content-substance and expression-substance to be fundamentally different in kind. SFL tends to conflate Hjelmslev’s expression-substance with the material phenomena of language: the actual sound-waves, physical gestures, ink marks on paper, and so on (Taverniers, 2008, p. 277). In doing so, SFL moves the stratified model out of the purely conceptual realm and into the realm of physical instances of language in use. Bache’s (2010) view that SFL is focused on substance in the Hjelmslevian sense is therefore misleading. Hjelmslev’s *substance* is a conceptual abstraction, and not the concrete material phenomena of languages themselves.

2.ii.c Context and connotation: enregistering languages

So far, we have primarily considered language as a process of denotation (Hjelmslev, 1954), by which extra-linguistic content and extra-linguistic expression are formed into linguistic abstractions and combined to make a sign. However, denotation alone is not sufficient to explain the range and complexity of human meaning-making. This is achieved through the notion of a connotative semiotic, formally defined as a semiotic whose plane of expression is, itself, a denotational semiotic (Hjelmslev, 1954, 1961). The complete signs constituted by the the denotative system make up the expression-substances of the connotative system; these signs are connected to a plane

of content, becoming the signifiers of a new collection of signifieds (Barthes, 1965; Lagopoulos & Boklund-Lagopoulou, 2021).

Diagrams illustrating the connotative semiotic often leave the content plane almost entirely undetermined, with the expression plane imagined as a container for the denotative semiotic in some form. Nonetheless, the content plane of a connotative semiotic can be analyzed in the same terms as that of a denotative semiotic (Hjelmslev, 1961, p. 118). It consists of the same stratification into purport, form, and substance. The schematic categories of this content-plane, its connotators, are related to both content and expression planes of the denotative semiotic. The expression-forms of the connotative semiotic are categories of signs (or “sign classes” – (Hjelmslev, 1961, p. 118)), which share the same connotator.

The content-purport available to the connotative semiotic is imagined by Hjelmslev as very broad: alongside the more obvious sense of “connotation” (meanings which are implied or suggested), the fact that some linguistic object can be recognized as belonging to an idiolect, style, dialect, or language is all the result of connotation (Hjelmslev, 1961, p. 115). Our ability to enregister certain linguistic units or groups of units as having languageness is therefore a matter of connotation. Languages are not individual semiotic systems, but classes of denotative signs which share the same connotator: “it is the semiotic schema(ta) and usage(s) which we designate as the Danish language that are expression for the connotator ‘Danish’.” (Hjelmslev, 1961, p. 118). Hjelmslev’s “connotation” and Agha’s “enregisterment” are, therefore, theorizations of the same phenomenon.

2.ii.d Reciting and writing: phonology, orthography, and the double articulation

With this understanding of Hjelmslevian stratification, and having clarified how it differs from later reformulations in SFL, it is possible to move on to a model of the sign which explains how orthography relates to other aspects of the language system, and what possibilities for enregisterment it offers. For Hjelmslev, speech and writing are simply two conventions for the representation of linguistic objects. The formal categories are the only inputs to the sign-function, and these categories have no definitive features outside the schema to which they belong. Whether the substance that realizes a categorical unit is ultimately manifested in speech or in writing is of no consequence: “qu’il se manifeste par tel ou tel phonème ou par telle ou telle lettre d’un alphabet (soit l’alphabet latin ou l’alphabet morse), tout cela n’affecterait en rien

la définition de notre élément”⁹ (Hjelmslev, 1942, p. 34). This conception creates difficulties in fully explaining how orthographic practices become imbued with meaning, however. This section will therefore suggest a different conception of the relationship between phonology and orthography, and situate them as elements of a connotative semiotic system.

The expression plane of linguistic signs has, at its root, a comparatively limited number of atomic units which do not carry denotative meaning themselves but can be combined to produce meaningful units. In order to be both cognitively manageable yet (essentially) unlimited in its expressive potential, language must have a resource for constructing new signs which is easily acquired and used (Barthes, 1965; Hjelmslev, 1961). Natural languages therefore operate with what semioticians call a “double articulation” (Martinet, 1980, p. 13), using permutations of a small set of distinctive units (which Hjelmslev (1961, p. 46) calls “figuræ”) to create meaningful units of expression-substance.

The term *articulation*, as it is used in “double articulation”, refers to the segmentation of linguistic syntagms into their constituent units. Double articulation is understood as a top-down process, beginning at the higher level of expression and continuing down to the level of the figuræ (Barthes, 1965). The results of the first articulation are units “douées chacune d’une forme vocale et d’un sens”¹⁰ (Martinet, 1980, p. 13); in other words, signs. The first articulation identifies the significant units of expression, as a side-effect of identifying signs as a whole. The second articulation decomposes these significant units into figuræ. Like content-forms and expression-forms proper, figuræ are defined relationally, as empty categories. Through linguistic acts, the categories are associated with features that define them, and through the aggregation of acts as usage, the bounds of the segment of purport which is selected by the formal category are delineated.

Although Hjelmslev’s extremely schematic approach to language leads him to disregard its physical manifestations as incidental details, other semioticians do not agree. From the very beginning, Saussurean semiotics privileged the spoken over the written: “il est impossible de prendre pour base de la linguistique le mot écrit [...]. Le but de l’alphabet est de fixer par des signes conventionnels ce qui existe dans la

⁹ *whether it is manifested by such and such a phoneme or by such and such alphabetic character (be it the Latin alphabet or the Morse alphabet), none of that affects the definition of our element in the slightest*

¹⁰ *each endowed with a vocal form and a meaning*

parole”¹¹ (de Saussure, 1907, p. 5). For Martinet, too, “les signes du langage humain sont en priorité vocaux”¹² (Martinet, 1980, p. 8). Writing is understood as dependent on speech; yet the cultural weight of orthography is such that decisions about pronunciation begin to be influenced by orthographic rules, particularly in the pronunciation of ELUs. In these circumstances, “le signe est considéré comme quelque chose de donné en soi ayant une valeur absolue et l’orthographe comme un bien national, un patrimoine”¹³ (de Saussure, 1907, p. 10). The result is “la falsification de la prononciation par l’écriture”¹⁴ (de Saussure, 1907, p. 10).

In opposition to the view that “la lecture vient doubler la parole, jamais l’inverse”¹⁵ (Martinet, 1980, p. 8), Hjelmslev explicitly, even contemptuously, dismisses the idea of a direct relationship between orthography and phonology: “On pourra, si on le désire, se complaire à bâtir des allégories plus ou moins spirituelles, en prétendant par exemple que la substance graphique ‘exprime’ la substance phonique [...]. Mais nous avons justement voulu prémunir contre cet abus de métaphores”¹⁶ (Hjelmslev, 1954, p. 188). In treating all forms of linguistic expression as equally external to his abstract semiotic architecture, however, Hjelmslev’s view is remarkably ahistorical. He ignores the fact that individuals learn to speak before acquiring literacy (if they do the latter at all), and that human beings developed the capacity to use spoken language long before developing writing as a technology. Although he mentions the Latin alphabet, he does not mention that languages using that alphabet all use it in such a way that particular configurations of its characters represent particular phonemes or sequences of phonemes.

If, as Hjelmslev argues, the orthographic and phonological units of a language are merely different ways of representing the same schema, which have no direct relationship, then it is impossible to explain many of the instances of enregisterment

¹¹ *it is impossible to take the written word as the foundation of linguistics [...]. The purpose of the alphabet is to fix what exists in speech using conventional signs.*

¹² *the signs of human language are, first and foremost, vocal*

¹³ *the sign is considered as something which has an absolute value in itself, and orthography as a national asset, a heritage.*

¹⁴ *the falsification of pronunciation by writing*

¹⁵ *speech mirrors writing, never the other way around*

¹⁶ *We might, if we so wished, amuse ourselves by constructing more or less spiritual allegories, by claiming for example that orthographic substance “expresses” phonological substance [...]. But we wished to guard against this abuse of metaphors.*

we will encounter, both in the overview of prior work on orthography in chapter 4 and in the analysis of data in chapters 6. In particular, it is impossible to explain fully how orthographic adaptation of loanwords functions as a matter of relative type-faithfulness to two different systems (Agha, 2009) if we must evaluate written ELUs as examples only of orthographic types, and are unable to evaluate them simultaneously as examples of phonological types. Rather than accepting Hjelmslev's judgement, therefore, the model of language used for the remainder of this study instead imagines the second articulation of natural languages as a primarily phonological phenomenon. In alphabetic literacy, orthographic figuræ enrich the second articulation, becoming linked to the phonological figuræ according to the conventions of the language in a micro-semiotic process; this parallels the semiotic process, which treats the morpheme as the expression of a denotative content-form, and the macro-semiotic process which treats the full sign as the expression of a connotative content-form.

An important detail of this model is that the substances of both content and expression are understood as collections of etic units, rather than of emic ones. The emic units are properly thought of as belonging to the interface between form and substance; they are etic substances which, through their conventional assignment to the same formal category, are understood as emic substance-categories. Nonetheless, in denotation and connotation, the units of expression substance are syntagms of the emic units of the previous level of semiosis. The expression-substances in denotation are morphs, which are etic units of morphology, but are simultaneously syntagms of phonemes, which are emic units of the second articulation.

The relationship of substance to substance in the linguistic act embodies the potential for making meaning. The relationship of form to form in the language as schema implies the potential for a communal understanding of that meaning. But the core of semiosis comes in a relationship of content-form to expression-substance, where there is no corresponding relationship between expression-form and content substance. This is the true nature of the line of arbitrariness identified by Hasan (2013): that the schematic categories of content select not only the substance of what is meant, but also of what means. Because the expression-substances of the connotative and denotative semiotics come pre-formed, as syntagms of emic units of the previous semiotic process, their schemata need not perform the same segmentation of purport that is required of the content-schemata. In fact, the only real purpose for the expression-schemata beyond the level of the second articulation is to shadow the content schemata. The work of deciding how to place boundaries

around areas of purport has already been completed. The content schemata of the denotative semiotic and the connotative semiotic are the primary organizing principle of their respective levels of semiosis.

Because the lower-level process generates the substances for the higher-level one, there is no interaction with extra-linguistic expression-purport except in the second articulation. The second articulation does the work of segmenting and organizing purport, creating phones and phonemes. The phonemes are organized according to the language's phonological (and phonotactic) schema, which determines what the phonemes are and how they can legally be combined (as well as managing the realization of other linguistic features with phonological implications, such as prosody). The syntagms which are legal productions of the phonology are therefore the input to the denotative semiotic: not phones or phonemes, as SFL argues, but sequences of phonemes, or morphs. In the case of alphabetic literacy, moreover, the morph in its guise as a sequence of phonemes is no longer a syntagm of simplex units, but carries with it the extra freight of an orthographic counterpart.

As a whole, this model (fig. 2.7) situates orthography as a counterpart to phonology, as Hjelmslev argues it is, and yet also explains it as a secondary means of expression, as Saussure and Martinet describe it. Inasmuch as the planes of content and expression are symmetrical, phonology and orthography are each equally capable of acting as the other's representation. Inasmuch as the semiotic act itself imbues content and expression with different roles, however, relating phonology and orthography in a micro-semiotic process establishes the primacy of phonology as the ultimate source of meaning in the second articulation. Unlike the other semiotic processes, the relationship of orthography to phonology imagines both content and expression as having a direct relationship to purport; because graphemes must be formed directly from purport, as phonemes are, they cannot be controlled solely by the phonological schema. This is the reason why Saussure balks at "*la falsification de la prononciation par l'écriture*"¹⁷ (de Saussure, 1907, p. 10). He recognizes the mismatch between the schema of orthography and that of phonology as, in many ways, a violation of the elegance of language in its natural state of evolution.

A language's primary expressive modality, the one which evolved with the language itself, is linked to its content purport arbitrarily. Any additional modality imposed after the language itself has already evolved, however, must stand in an

¹⁷ *the falsification of pronunciation by writing*

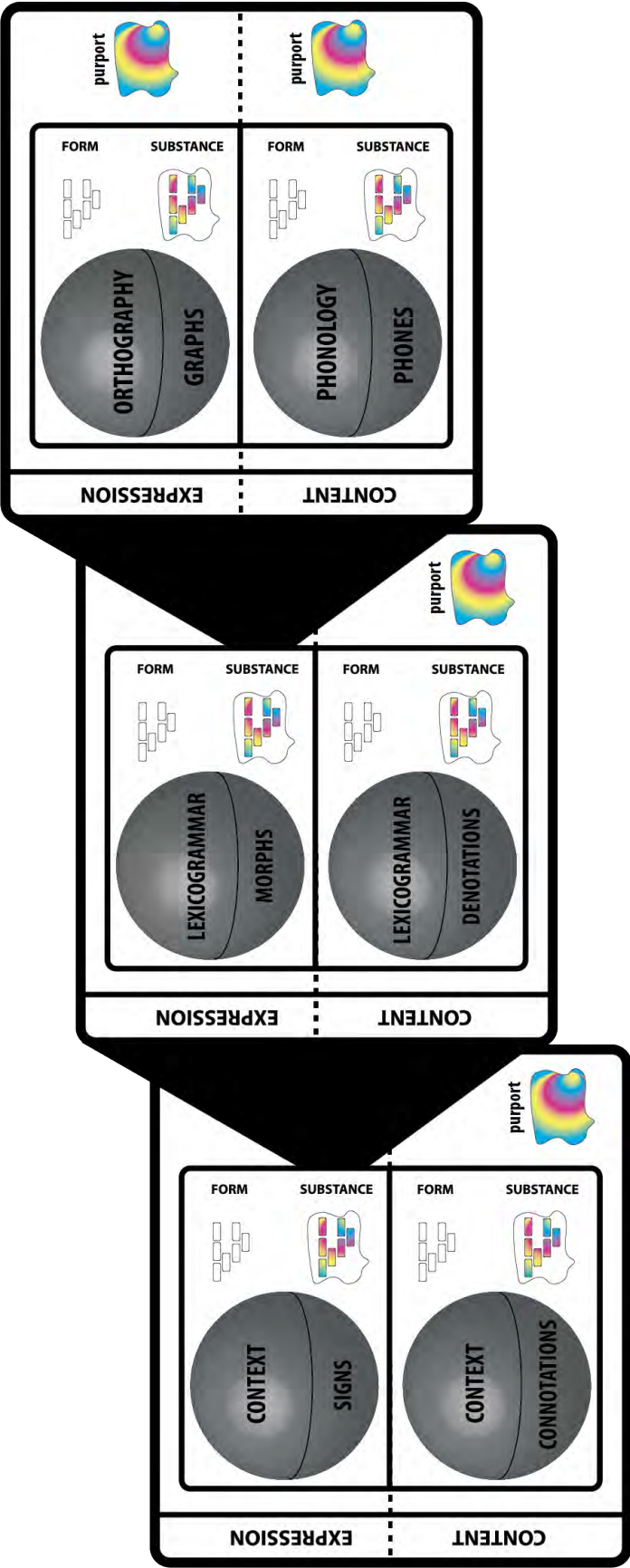


Figure 2.7: The expanded structure of language as a connotative semiotic, including alphabetic literacy at the level of the second articulation.

arbitrary relationship *both* to the content *and* to the primary modality. Although Martinet argues that the linguist “fait donc par principe abstraction des faits de graphie”¹⁸ (Martinet, 1980, p. 8), the model I have proposed will form the basis for an analysis of written forms which positions them as unique resources for creating meaning. Orthographic patterns are not only vehicles for conveying phonology, or for representing higher-level units such as lexemes, but, through enregisterment, for a broad range of contextual meanings. The conventions of phoneme-grapheme relationships, established in the micro-semiosis of the second articulation, offer additional opportunities for negotiating type-faithfulness in mixed-language contexts. Far from being an abuse of metaphor, the theorization of a semiotic relationship between phonemes and graphemes is, as the data analyzed later in this study will show, a necessary part of describing many creative and non-standard orthographic practices.

2.iii Language-mixing and enregisterment in Welsh

An advantage of using this semiotic model, including Agha’s enregisterment, is that it can account for both monolingual and bilingual language behaviours in much the same way. The enregisterment of linguistic features, which happens when connotators become attached to signs, is responsible for all kinds of co-associated feature constellations, whether we experience them as registers, dialects, or as languages. Indeed, (MacSwan, 1999) argues that pragmatic constraints on code-switching are analogous to constraints on language use in different registers in monolingualism. Combining features from two different languages is therefore the same kind of semiotic act as combining features of two dialects or registers of the same language. Enregisterment explains our sense that certain lexemes, or graphs, or phones have languageness. Context permeates the linguistic system at all levels, so that any linguistic object or property can become saturated with all kinds of contextual meaning. Linguistic features, experienced often in a particular context, become unavoidably entwined with that context. This has specific ramifications for Welsh: the contexts in which it may be fully experienced are limited for many Cymrophones, and the kinds of language experienced in those contexts are often

¹⁸ *therefore, as a matter of principle, disregards acts of writing*

artificially restricted by monolingualist ideologies.

2.iii.a Formality and function: enregistering Welsh for education

The crucial importance of Welsh-medium education in transmission of the language has steadily increased since the mid-twentieth century. Many children now acquire Welsh in immersion education, through social contact with teachers and friends as well as in the formal context of lessons. This places a significant linguistic burden on the Welsh-medium education system. Welsh children in English-medium schools are likely to have many opportunities to acquire colloquial English outside school, in their communities, from social media, and from popular culture, even if they do not come from Anglophone homes. Children in Welsh-medium schooling, in contrast, may have few comparable opportunities to acquire colloquial Welsh, especially if they live in largely Anglophone areas of the country (Price, 2020).

For children in Welsh-medium education from Anglophone homes, their primary experience of Welsh, in all its forms, comes through school. There are numerous risks associated with this situation. As E. M. Thomas et al. (2014b, p. 341) have argued, if Welsh-medium education is to contribute meaningfully to the revitalization of the language, it must aim to promote “the productive social and communal integration of young people as speakers of the minority language outside school”. Yet young Cymrophones often experience Welsh and English increasingly as functionally separate, the former predominantly enregistered as the language of educational and formal contexts, and the latter as the language of social and informal use (Price, 2020). For schools to develop Cymrophones’ competence in educational registers of Welsh is not sufficient. Where traditional schooling in a majority language can focus on developing children’s confidence in the standard language and ability to use its formal and academic registers, minority-language schooling must also ensure that children develop confidence in the kinds of colloquial, informal, familiar, and creative registers to which they are not naturally exposed in the home and the community.

Recent linguistic research studying younger Welsh-English bilinguals is generally supportive of language mixing. It is now often recommended that educators learn to capitalize on students’ multilingual repertoires through practices including code-switching and translanguaging (D. M. Lewis, 2011; E. M. Thomas & Webb-Davies, 2017). For some teachers, skilful use of code-switching supports their

classroom management and rapport-building skills, performing a range of pragmatic functions (Clapham, 2010). Nonetheless, the enregisterment of English-origin ELUs as “bad” Welsh is a common theme in pedagogical spaces. Examiners reports for Welsh-medium public assessments consistently criticize candidates for “Defnyddio geiriau Saesneg/geiriau Saesneg wedi’u Cymreigio”¹⁹ (CBAC, 2017), (see also e.g. CBAC, 2019, 2022). Other points noted in these reports suggest that candidates are failing to acquire Welsh registers appropriately, either lacking the register itself, or failing to deploy it in the right context. For the written portion of the 2017 Welsh Language GCSE examination, candidates’ work in the written portion was “rhy llafar”²⁰ (CBAC, 2017). Candidates offering an oral presentation in 2019, on the other hand, were using “Cywair rhy ffurfiol”²¹, as well as “Defnydd o eiriau Saesneg”²² (CBAC, 2019). The picture emerging from these reports is of young Cymrophones whose awareness of situationally appropriate language is limited, and who are either unaware of the stigma against language-mixing, or unable to find an appropriate ILU to use in context. In the framework discussed earlier in this chapter, this can be described as a failure of enregisterment of the features characterizing formal and informal registers. The following sections explore some characteristics of the Welsh context which help to explain why this might be the case.

2.iii.b Pedantry and plausibility: technical vocabulary and terms of art

The complexity of linguistic attitudes amongst Cymrophones is particularly notable in the case of technical vocabulary. On the one hand, the idealization of the language of rural Wales as the most authentic form of Welsh led some early commentators to criticize the Welsh taught to schoolchildren as “lliprynaidd lwyd yn ymyl cryfder ymadrodd y bugail a’r crefftwr gwlad”²³ (G. Jones, 1926, p. 84). Hawke (2018) argues that deliberate neologism in Welsh, if supported by extensive use of the new terms in official and popular media, can be (and has been) successful, pointing to the coining

¹⁹ *Using English words/English words adapted to Welsh.*

²⁰ *too oral*

²¹ *An overly formal register.*

²² *Use of English words.*

²³ *listlessly gray beside the discursive power of the shepherd or rural craftsman*

of what are now everyday, colloquial terms such as *teledu* (television) and *trydan* (electricity). He notes that such terms, while they are derided by some as puristic when first introduced, can increase speaker confidence by lessening reliance on English loanwords.

Nonetheless, disparagement of neologisms is also found in other minority-language contexts. Neologisms employed by new speakers of Breton have been described as “incomprehensible”, and described as “inauthentic” in comparison with the use of French-origin ELUs by traditional speakers of the language (Hewitt, 2017, p. 132). In Ireland, terminological innovations introduced by language planners have met with a mixed reception in the traditionally Irish-speaking Gaeltacht, as the centralization of linguistic authority in the State rather than the language community leads to the sense that neologisms are arbitrary or artificial, and that the use of adapted ELUs is often the clearest and most natural choice for technical terminology (Ní Ghearáin, 2011).

There is a corresponding thread in earlier Cymrophone discourse which identifies the language of dictionaries and language planners as inauthentic in comparison to the words used in rural Cymrophone communities, even when the latter words are ELUs. The use of English-origin ELUs may, indeed, be preferred by Cymrophones:

Y mae ‘gwefr-hysbysiad’ er esiampl yn air da, ac felly ‘prwygraffft,’ er nad cystal nac mor hunan-esboniol â’r cyntaf; ond er mor dda, ‘telegram’ a ‘program’ yw y geiriau a ddefnyddir yn eu lle gan naw Cymro o bob deg, ac yn iawn, fel y tybiwn. Ni wiw myn’ d i’r geiriadur am ‘eiriau gwneyd’; y mae gan y wlad ffordd ferach a sicrach o lawer na hyny i benderfynu’r mater. ²⁴(D. E. Evans, 1897)

The enregisterment of language choices as markers of identity is expressed by a mid-twentieth century Welsh author, Alun Llywelyn-Williams, in these terms:

Y peth syn yw yr amheuir pawb sy’n ceisio osgoi’r ffurfiau estron, a’i gyhuddo’n aml o siarad Cymraeg annaturiol. O’r braidd yn wir y cyfrifir

²⁴ ‘Gwefr-hysbysiad’, for example, is a good word, as is ‘prwygraffft,’ although not as good or as self-explanatory as the first; yet however good they are, ‘telegram’ and ‘program’ are the words used in their place by nine Welshmen out of every ten, and correctly, I would say. It avails nothing to go to the dictionary for ‘made-up words’; the people have a much quicker and more certain way than that to decide the matter.

ef yn Gymro o waed coch cyfan ; odid nad rhyw goegyn digrif ydyw
sydd wedi dysgu'r Gymraeg o'r geiriadur, a'i iaith o'r herwydd yn
anystwyth a thu hwnt i amgyffred y werin a fu gynt mor ddeallgar a
diwylliedig. (Llywelyn-Williams, 1940, p. 146)²⁵.

The author feels that his very claim to Welshness is being attacked by those who find his preference for neologistic ILUs inauthentic. Seen through the lens of enregisterment, this is another example of Agha's (2005) notion of the "social range" of enregistered features: the personae performable using Welsh characterized by such neologisms are over-educated oddballs, outsiders to the community of "true" Welshmen.

In his turn, Llywelyn-Williams enregisters his opponents' language choices as signs of negative personal characteristics. The rural folk's insistence on using the popularly preferred ELUs is a sign that they have lost their previous status as "intellectual and cultured". The most authentic Welsh is now, for this author, that which takes full advantage of modern academic developments. This attitude is supported by other early Cymrophone authors, who lamented that academic philosophers were "difwyno'n hiaith â bastardiaid fel 'seicoleg,' 'teip,' 'simbol,' 'sentimentaliti,' etc., pryd y gallent drwy ymgynghori â'n hysgolheigion Cymraeg gael geiriau cydawns ag anianawd ein hiaith"²⁶ (E. D. Jones, 1933, p. 49). Similar attitudes persist in contemporary minority-language contexts. Reversing the denigration of Breton neologisms, new speakers of Breton who acquired the language through formal education often comment disfavouredly on traditional speakers' propensity to use French loanwords and avoid neologisms (Hornsby, 2015).

Nonetheless, concerns that English is monopolizing the lexicon in certain fields is driven by the fact that, in many multilingual contexts, English has become enregistered as the default language for classroom technical discussions, even when other aspects of education are not English-medium (Farrugia, 2003). English terms are almost always embedded in the language community before new

²⁵ *The astonishing thing is that everyone who tries to avoid foreign forms is suspect, and often accused of speaking unnatural Welsh. He is barely accounted a full-blooded Welshman; no doubt he is some strange reprobate who has learned Welsh from the dictionary, his language stiff as a result and beyond the comprehension of the gwerin, who were once so intellectual and cultured.*

²⁶ *defiling our language with bastards such as 'seicoleg' [psychology], 'teip' [type], 'simbol' [symbol], 'sentimentaliti' [sentimentality], etc., when they could, by consulting with our Welsh scholars, have congenial words which are innate to our language*

minority-language terms can be coined. As a result, the ELU becomes enregistered as the appropriate term for technical use, and changing to a later-introduced ILU may feel artificial or awkward (Ní Ghearáin, 2011).

For scientific subjects, terminological precision and lack of ambiguity are particularly important aims, and it is here that effective terminology planning is most needed (Krägeloh & Neha, 2014). When working in an academic register, both students and educators benefit from a sense of confidence that they are using the correct technical term (Prys, 2003). Modern efforts towards Welsh terminological standardization for education aim to improve the consistency of existing resources, providing a single preferred label for a specific concept, in the context of a particular domain (Andrews & Prys, 2016). Despite efforts to provide appropriate standardized terminology, nonetheless, children learning subjects such as mathematics through a minority language reportedly find the use of neologistic ILUs for technical terminology particularly difficult (Krägeloh & Neha, 2014).

The tension between using newly-coined ILUs and (adapted) ELUs to refer to technical and scientific innovations has been present in Welsh discourse since at least the time of the Industrial Revolution, when the linguistic Anglicization of many areas of Wales coincided with unprecedented scientific and technological development. New ILUs coined by lexicographers and terminologists to keep up with scientific innovation have not always proven successful. The English lexemes almost inevitably form a standard against which new terminology is measured; they are perceived as unmarked, neutral linguistic units, and the minority-language neologism's acceptability may therefore be judged based on how well it reproduces the familiar English term, rather than on its own merits (Ní Ghearáin, 2011). Moschkovich (2007) suggests that bilinguals switch between languages in technical discussion for many reasons, including differences in register between semantically equivalent terms in the two languages; as a result, confining students to a single language in technical discussion robs them of an important tool of communication.

The need for technical and scientific terminology is repeatedly invoked as a significant reason for the use of English-origin loanwords in Welsh, or for using English to gloss potentially unfamiliar Welsh terms. For some, this is neither strange nor particularly problematic. For one early-nineteenth-century commentator, the fact “bod rhai yn devnyddiaw *geiriau saesneg er egluraw y gymraeg*”²⁷ seemed perfectly

²⁷ *that some use english words to explain the welsh ones*

natural: “nid yw hyny mewn un modd yn dangaws tlodi yr *iaith*; yr ačaws yw, mae y Saeson gymmain o flaen y Cymry, yn yr *elvau* ac y *gwyddiannau*, val pan ddefnyddir geiriau cymreig *newydd*, gorvod yw arnom roi geiriau seisnig i egluraw ein meddwl”²⁸ (‘Caervallwch’, 1834). Antia (2000) is reluctant to allow a place for sociocultural concerns in making decisions about the appropriate treatment of technical vocabulary. He is concerned that the practical communicative needs of language users are being overshadowed by decisions such as the blanket exclusion of borrowings from politically disfavoured languages, seeing such decisions as reproducing unnecessarily divisive cultural conflicts. Nevertheless, terminology planning guidelines for minority languages often have a stated preference to avoid borrowing from the majority language (Krägeloh & Neha, 2014; Virues-Ortega et al., 2015), in an attempt to avoid exacerbating the appearance that the smaller language is being overwhelmed by the larger (Prys, 2003).

The criteria that have been proposed as essential for an effective technical neologism are varied, and the avoidance of loanwords is often placed below other considerations. These include formal characteristics such as brevity, conciseness, and systematic orthography; grammatical characteristics such as the ease of using the term as the basis of further morphological derivations; semantic consistency and clarity; and cultural considerations, taking into account community reactions to the term (Akbari, 2020). There is no agreement on which features are most important for the success of an introduced term, or indeed whether the satisfaction of all of the criteria is essential (Akbari, 2020). On the whole, however, brevity, consistency, and the minimization of redundancy are repeated precepts in the literature on coining terminological equivalents for terms in another language (Akbari, 2020; King’ei, 1999; Virues-Ortega et al., 2015). Although the use of periphrasis can be a valuable tool for translating terminology from one language to another, it has numerous drawbacks, including the difficulty of using the new term as the basis of either inflectional or derivational morphology, making it less flexible and functional than a single lexeme (Virues-Ortega et al., 2015).

Welsh-medium pupils are already at a substantial disadvantage as regards acquiring many domains of technical vocabulary. Much of the basic vocabulary of computer science and information technology, as well as of the natural sciences, is

²⁸ *this in no ways suggests the poverty of the language; the cause being, the English are so far ahead of the Welsh, in the elements and the sciences, so that when new welsh words are used, it is needful for us to give english words to explain our meaning.*

seen whenever one uses a computer or a smartphone, or watches a science fiction film²⁹. This everyday exposure is reduced or absent in Welsh. In schools, there has historically been a lack of appropriate Welsh-language teaching materials, a situation that was noted as far back as the mid-nineteenth century, when the editors of a Welsh educational journal noted that “nid yw y iaith Gymraeg, fel y gwyr pawb o honom, yn feddiannol ar lyfrau Ysgol mewn gwahanol ganghenau gwybodaeth”³⁰ (Anon., 1851). Contemporary Welsh-medium pupils are still faced with a lack of appropriate textbooks (C. Baker, 1995; BBC News, 2016; C. H. Thomas, 1982; C. Williams, 1994), having either to use English books or rely on teachers’ translations, which are unlikely to benefit from the kind of terminological regularity found in published textbooks. This further limits pupils’ exposure to standardized Welsh-language terminology, and may actually increase their familiarity with English terms if they resort to English-language materials to support their studies. As a result, young Cymrophones have less opportunity to develop confidence in using Welsh technical terms, and may feel more comfortable that they know the English term for a scientific or technical concept than the Welsh one.

The need for appropriate vocabulary in technical and scientific registers is therefore particularly pressing for children in Welsh-medium education, where the use of English technical terminology risks harming their achievement as measured by standardized assessments. One recent examiners’ report, for example, noted that “Er mwyn codi marciau’r ymgeiswyr [...], mae angen osgoi defnyddio termau Saesneg megis ‘bold’ yn hytrach na ‘print trwm/tywyll/wedi’i dduo”³¹ (CBAC, 2019). This comment exemplifies numerous serious problems with the insistence on using ILUs wherever possible in academic registers of Welsh. Unless the terminology is properly standardized, it will be difficult for students to feel confident that they are using the correct term in context. The fact that the examiners offer three possible Welsh equivalents for “bold” suggests that there is disagreement about the proper term, when there is no such disagreement in English. If even the examiners cannot select

²⁹ S. W. Evans (2006), however, notes that the familiarity of terms may occasionally be a disadvantage. He proposes that a word which belongs exclusively to a technical register may act as a cue that a particular kind of scientific thinking is required, where a term which is also encountered in everyday language does not have the same effect.

³⁰ *the Welsh language, as each of us knows, is not possessed of schoolbooks in the various branches of knowledge.*

³¹ *In order to raise candidates’ marks [...], it is necessary to avoid using English terms such as “bold” in place of “heavy/dark/blackened print”.*

the appropriate term of art in this context, then it is difficult to imagine how pupils, or even their teachers, are expected to know the “correct” term to use.

As already noted, commonly accepted standards for technical neologisms strongly prefer concise lexemes, a single term for a single meaning, and unambiguous representation of the semantic meaning. The examiners’ report violates these precepts in various ways with its choice of terms. Three possible alternatives are given, immediately violating the criterion that a single meaning should have a single term. In one case, the alternative is structurally a prepositional phrase (although it functions rather like a non-finite clause, since it is the equivalent of a verb phrase in the perfect tense): “print wedi’i duo” is, very literally, “print after its blackening”, but has the same function as “blackened print”³². This alternative therefore goes against the ideal that technical terms be as concise as possible. Two of the proposed terms fail to match the semantics of “bold” unambiguously. Bold type is not necessarily blackened (“wedi’i duo”) or dark (“tywyll”); white print on a black background may still be bold. The word “trwm” (*heavy*) is more appropriate semantically, since it conveys the idea that the elements of each character are made up overall of heavier strokes than those of the standard Roman form of the same typeface. A word such as “amlwg” (“obvious”, “prominent”) might offer an equally good, if not a better, alternative, since it conveys the semantics of visual prominence which are also conveyed by “bold”.

Nonetheless, more significant than any inadequacies of the proposed Welsh alternatives is the fact that “bold” is listed in *GPC* as a Welsh equivalent for the English typographic term of art “bold” (fig. 2.8), with an attestation date of 2002. Since *GPC* is freely available online, and is an authoritative dictionary of Welsh, it is not unreasonable to imagine a pupil or teacher consulting it and interpreting the entry as confirmation that “bold” is acceptable for use in Welsh as a typographic term of art. Yet a student who used “bold” in a Welsh-medium examination might be docked a mark for using “termau Saesneg” (CBAC, 2019).

This reveals a fundamental, and serious, problem in Welsh-language pedagogy. The enregisterment of English-origin ELUs as “bad” or “substandard” Welsh, suitable only for (at best) colloquial and informal registers of Welsh, leads to the

³² As Morris Jones (2010, p. 367) notes, this type of construction can often be replaced with an adjective using the suffix “-edig”: “dŵr berwedig” and “dŵr wedi’i berwi” are both acceptable equivalents for “boiled water”. Not all verb forms can take this suffix, however, and *“duedig” seems unlikely to be acceptable.

bold²[bnth. S. *bold*]

a.

Arg. A chanddo ymddangosiad trwm ac amlwg (am wyneb teip):

bold (of typeface, in typogr.).

2002.

Figure 2.8: GPC Online entry for the word “bold”.

stigmatization of linguistic elements which are not only commonplace and conventionalized within the language community, but are also in some cases legitimated by other sources of linguistic authority. The next section will analyze the approach to ELUs in two texts which offer advice on using the Welsh language, revealing that this confused and confusing approach to ELUs is not limited to this one instance, but rather is a trope of the Cymrophone language-advice genre.

2.iii.b.1 Protocol and perplexity: the enregisterment of English-origin lexical units as “bad Welsh”

The increased presence of English-origin ELUs in a Welsh text is an acknowledged stylistic signal of informal register (Rottet & Morris, 2018, p. 421). As a result, authors discussing the successful use of formal Welsh regularly refer to English-origin ELUs. We will focus here on two language-instruction texts for Cymrophones: *Cymraeg Clir* (*Clear Welsh*) (C. Williams, 1999) and *Cymraeg yn y gweithle* (*Welsh in the workplace*) (R. H. Williams, 2018). This discussion will illustrate the importance placed on the avoidance of what are termed “English” words in formal Welsh, but also reveal the significant confusion in these texts about what an “English” word is.

Both authors situate English-origin ELUs as important variables in establishing register formality in Welsh. C. Williams (1999) claims that the purpose of his guide is to help users “i benderfynu ar gywair ffurfiol”³³ (C. Williams, 1999, p. 22). He goes on to offer three features of such a register which, he says, is one:

- y mae pawb yn gallu ei ddeall;
- sy’n cadw urddas yr iaith trwy gadw geiriau Cymraeg;

³³ to choose a formal register

- sy'n cadw at batrymau naturiol y frawddeg yn hytrach na rhai seisnigedig.³⁴ (C. Williams, 1999, p. 22)

It is impossible to be certain whether the phrase “urddas yr iaith” (*the dignity of the language*) refers to the nature of the specific instance of formal language which the reader wishes to create, or to the nature of the Welsh language as a whole. Williams may only be exhorting the reader to ensure that their formal Welsh is dignified; but he may equally be charging them with ensuring that the language itself is not degraded in their use of it. Whichever reading is to the fore for an individual reader, linguistic “urddas” is implied to result from one stricture: retaining Welsh words.

C. Williams’s (1999) stated focus is on encouraging clear and simple language where possible, and avoiding longer words and unfamiliar vocabulary. Discussing information leaflets and questionnaires or forms intended for the general public, he argues that “bod yn gryno ac eglur yw’r prif nod. Mae hynny’n golygu ein bod yn defnyddio iaith symlach. [...] Ond nid yw’n golygu defnyddio geiriau Saesneg.”³⁵ (C. Williams, 1999, p. 35). And yet, in a previous section discussing the replacement of longer words with shorter, more familiar ones, he recommends replacing the words *hysbysiad*, *gostyngiad*, and *galw* with the words *bil*, *disgownt*, and *ffonio*. Each of the replacements is an English-origin ELU. It is therefore not clear exactly what Williams is referring to when he expresses disapproval of “geiriau Saesneg”³⁶.

Like C. Williams (1999), R. H. Williams (2018, p. 277) outlines what she calls “y cywair ffurfiol”³⁷ of Welsh, again giving a prominent role to replacing “English” words with “Welsh” ones in order to establish formality. She offers a list of such words and appropriate replacements:

- watsho > gwylio
- telefishion > teledu
- diseido > penderfynu
- wel > [hepgor]

³⁴ -that everyone can understand; - that maintains the dignity of the language by retaining Welsh words; -that holds to the natural patterns of the sentence rather than Anglicized ones.

³⁵ the main aim is to be concise and clear. That means using simpler language. [...] But it does not mean using English words.

³⁶ English words

³⁷ the formal register

- joio > mwynhau

³⁸ R. H. Williams (2018, pp. 278–9)

It is worth comparing this list to the one which directly precedes it, offering examples of “geiriau llafar/anffurfiol” (*colloquial / informal words*) and appropriate formal replacements:

- cwpwl > ychydig / un neu ddau / pêr
- peint > diod
- amboiti > am / ynghylch
- mas > allan
- ffrind > cyfaill
- siŵr > sicr

³⁹ R. H. Williams (2018, p. 278)

These examples, like the vocabulary replacements recommended by C. Williams (1999) and quoted above, highlight Cymrophones’ difficulty in establishing what, exactly, constitutes a “Welsh” word.

³⁸ (Items in square brackets are explanations of the structure of the lexical item in question). [English *watch* + Welsh verbnoun suffix *-o*] > to watch; [English *television* written according to Welsh orthographic rules] > television; [English *decide* + Welsh verbnoun suffix *-o*] > to decide; [English *well* written according to Welsh orthographic rules] > (delete); [a reduced form of the English *enjoy* + Welsh verbnoun suffix *-o*] > to enjoy

³⁹ [a loanword derived from English *couple*] > a little / one or two / a pair; [a loanword derived from English *pint*] > a drink; [one of many colloquial forms of a Welsh preposition ultimately derived from ‘o peutu’ – *from all about*] > about / around; [a dialectal adverb meaning *out*] > out; [a loanword derived from English *friend*] > friend (or companion); a loanword derived from English *sure* > certain.

Table 2.1

	GPC	Cyn.	New.	Maw.	Aca.	Mod.	WLD	Gwy.	Gom.	Total listed
Williams (2018): ELUs classified as English words										
<i>wel</i>	✓	✓	✓	✓	✓	✓	✓		✓	8
<i>telefshion</i>	✓*									1
<i>watsho</i>	✓*				only 'gwatsiad'					1
<i>joio</i>	informal/ colloquial				only 'joio mas draw'					1
<i>diseido</i>										0
Williams (2018): ELUs classified as informal words										
<i>siŵr</i>	✓	✓	✓	✓	✓	✓	✓		✓	8
<i>cwpwl</i>	✓*	✓*	✓	✓*	✓		✓		✓	7
<i>ffrind</i>	✓		✓	✓	informal/ colloquial	✓	✓	†	✓	7
<i>peint</i>	✓		✓*	✓*	✓*	✓*	✓*		informal/ colloquial	2
Williams (1999): ELUs classified as simpler alternatives										
<i>bil</i>	✓		✓	✓	✓	✓	✓		✓	7
<i>ffonio</i>	✓		✓	✓	✓	✓	✓	✓	✓	8
<i>disgownt</i>	✓			✓	✓		✓		✓	5
Total listed	11	3	6	7	7	5	7	1	8	

✓* – word listed only as a variant form; † word used in defining text, s.v. “cyfaill”

All of the dispreferred words listed by R. H. Williams (2018) as “English” are, clearly, ELUs derived from English lexical items. However, four out of six of her colloquial and informal words are also English-origin ELUs. Table 2.1 shows the ELUs listed by R. H. Williams (2018) and C. Williams (1999), either as terms regarded as inappropriate by the former because they are “English” or “informal”, or as terms preferred by the latter as replacements for words which are harder to understand. The dictionaries consulted are placed in order of their publication date, except for *GPC Ar-lein* (2024), which is placed first, since it is an ongoing project. Dictionaries are identified by three-letter abbreviations in the header row, as follows: GPC: *GPC Ar-lein* (2024); Cyn.: *Geiriadur Cynaniaethol* (1903); New.: *Geiriadur Newydd* (1953); New.: *Geiriadur Mawr* (1983); Aca.: *Geiriadur yr Academi* (1995); Mod.: *Modern Welsh Dictionary* (2007); WLD: *Welsh learner’s dictionary* (2007); Gwy.: *Gwybod y Geiriau* (2015); Gom.: *Geiriadur Cymraeg Gomer* (2016).

The table notes the presence or absence of each word in a range of Welsh dictionaries. In each case, the table shows whether the word appears as a headword in the dictionary (except in the case of Griffiths and Jones (1995b), which is an English–Welsh dictionary; here, words are marked according to whether they are listed in definitions of an English headword). Empty cells indicate that the word is not present in the dictionary. Other cells either contain a tick mark, indicating that the word is listed in the relevant dictionary, or text indicating that the word is listed under a variant spelling or tagged as informal or colloquial.

“Ffrind” is present in the defining text of *Gwybod y Geiriau* (2015) as part of the definition of the word *cyfaill*, but not as a headword. Griffiths and Jones (1995a) does not feature “watshio”, but lists the form “gwatsiad”, which *GPC Ar-lein* (2024) considers as a variant of *watsio*. Griffiths and Jones (1995a) also features “joio” only as part of an idiom, “joio mas draw” (*to enjoy immensely*). The ELU “peint” is included in most dictionaries, but not in the specific sense that R. H. Williams (2018) means it; it is listed in the general sense of a measurement of volume, but not in the colloquial sense of a pint of beer drunk socially.

Of R. H. Williams’s (2018) list of ‘English’ words, most appear only in *GPC Ar-lein* (2024). The exceptions are *diseido*, which is entirely unlisted, and *wel*, which is listed in all the dictionaries consulted, except the small children’s dictionary *Gwybod y Geiriau* (2015). Indeed, *wel* is one of the three most frequently listed in this table, with the same score as *siŵr*, one of R. H. Williams’s (2018) “informal” words, and *ffonio*, approved by C. Williams’s (1999) as a simpler alternative form.

Although *peint* is technically only listed twice, this is because only dictionaries

listing the specific sense intended by R. H. Williams (2018) were counted. The word is also listed in its general sense by a further five dictionaries. Surprisingly, *ffrind* (the usual colloquial word for a friend) does not appear as a headword in the children's dictionary, although the more formal *cyfaill* does, and *ffrind* is offered as part of the definition of the latter word (*Gwybod y Geiriau*, 2015). Two of R. H. Williams's (2018) "informal" words are marked as informal or colloquial by at least one dictionary: *ffrind* (by Griffiths and Jones (1995a)) and *peint* (by D. G. Lewis and Lewis (2016)). One of her "English" words, *joio*, is also marked as informal or colloquial (by *GPC Ar-lein* (2024)).

The final group of words, C. Williams's (1999) preferred simpler alternatives, again shows variation. Although none of the three words is listed by all dictionaries, *ffonio* (*to phone*) might be counted as universally listed, since only the dictionary published before the telephone came into common use omits it⁴⁰. If this is taken into account, *ffonio* is the only word in this list which is present in all dictionaries consulted, since it is the only word present as a headword in *Gwybod y Geiriau* (2015). There is far less agreement about C. Williams's (1999) *disgownt*, which is present only in five of the consulted dictionaries.

Of course, listedness of words does not rely purely on whether the lexicographer counts them as a genuine part of the lexicon. A dictionary is constrained by practicalities such as the time available to produce it, the desired number of headwords, the physical size of finished product desired by the publisher, and so on. It is also constrained by the projected audience and their likely need for vocabulary at various levels of difficulty and in various semantic areas. The fact that some of the words in Table 2.1 are missing from a given dictionary cannot be taken as proof that the authors or compilers of those dictionaries did not accept those words as part of the Welsh lexicon.

Nonetheless, some suggestive information arises from the comparison of the various words. The ambiguity of the word *ffrind* is particularly interesting. Both R. H. Williams (2018) and Griffiths and Jones (1995a) consider this an informal or colloquial word, but it is otherwise listed in all but the two earliest adult dictionaries without any comment on its register. Nonetheless, it is omitted from *Gwybod y Geiriau* (2015), despite its obvious relevance to an audience of children. The fact that

⁴⁰ Indeed, apart from the children's dictionary, the earliest dictionary is also the least likely to list the ELUs considered here, suggesting that English-origin ELUs may have become more acceptable as part of the Welsh lexicon over time.

Gwybod y Geiriau (2015) uses “ffrind” as part of its definition of the word *cyfaill* suggests that the author expects children to be familiar with the word. Its omission from the list of headwords is therefore odd. This prompts speculation about whether its status as an English-origin ELU, and the evident attitude from some sources that *ffrind* is not an appropriately formal word, may have influenced its omission from *Gwybod y Geiriau* (2015).

Considering Table 2.1 as a whole, it is clear that the categories of words considered here differ in their profile. R. H. Williams’s (2018) “English” words are, with one notable exception, rarely listed in any dictionary other than the comprehensive historical dictionary *GPC*. The exception is the discourse marker *wel*, which is included in all the adult dictionaries, without any comment on its register appropriacy (such as labelling as informal or colloquial). Its profile is identical to that of *siŵr*, which appears in R. H. Williams’s (2018) list of informal words. This second list has a profile which is fairly similar to that of C. Williams’s (1999) words which he offers as simpler alternatives to be preferred when trying to write clear, comprehensible Welsh. There is apparently a difference of ideology here. C. Williams (1999), whose focus is on clarity and on reducing complexity and jargon in formal written Welsh, sees ELUs with wide community acceptance as a useful resource, which can be mined for terms which are easier for the average reader to understand. For R. H. Williams (2018), in contrast, such words are unacceptably informal.

It seems possible, therefore, that the establishment of an item in the lexicon of the language community, as judged by its presence in dictionaries, correlates with R. H. Williams’s (2018) understanding of English-origin ELUs as “English” or as “Welsh”. Since R. H. Williams’s (2018) does not explain why some English-origin ELUs are dispreferred because they are “English”, while others are dispreferred because they are informal, it is not possible to know for certain whether she considers the latter set of words truly “Welsh” or not. However, it is interesting that only one of the four loanwords in her list of informal words is marked as such (by Griffiths and Jones (1995b)). Williams’ judgement that a word is informal may therefore correlate to its status as an English-origin ELU, rather than to a general acceptance of its informality in the language community as a whole.

Even this explanation of Williams’ principles of categorization is complicated by her examples, however. She categorizes *wel* as an “English” word to be removed from formal Welsh, despite its being one of the most universally listed words amongst the examples considered here, and one of only three listed by the two earliest dictionaries

consulted. R. H. Williams (2018) lists no Welsh alternative for *wel*, instead instructing that it should be removed altogether, suggesting that the word's supposed Englishness is not the main factor making it inappropriate for the register of formal business Welsh. After all, if Englishness were the only problem with the word, one would expect a preferable "Welsh" replacement to be offered. The real problem is that *wel*, like its etymological parent, English *well*, is a colloquial discourse marker. It may be entirely appropriate in speech, but it seems unlikely that a writing guide would recommend the use of *well* in formal business English. It would be reasonable to categorize *wel* in R. H. Williams's (2018) list of informal words, alongside the similarly well-established ELU *siŵr*. The fact that she instead lists it as inappropriate because it is "English" suggests a confusion about what exactly an "English" word is, in her estimation.

Since only a few examples are offered by each author, only very tentative conclusions can be drawn about their attitudes towards ELUs. Nonetheless, even these few examples show a lack of clarity in defining words as "English" or "Welsh". For both R. H. Williams (2018) and C. Williams (1999), the construct of formality is inseparable from the presence or absence of "English words" in one's Welsh. Yet neither author clarifies what exactly is meant by an "English word". English-origin lexical items are treated variously as "English" or as informal "Welsh" by R. H. Williams (2018), and are variously prohibited or recommended as clearer and simpler alternatives by C. Williams (1999). The authors are implicitly in agreement that English lexical items act as a register marker, yet their texts raise complex questions about which lexical items, in which contexts, are actually considered to be "English".

Chapter 3

“Do I not love my country because I haven’t properly learnt mutations?”: The uneasy relationship of Welshness and Cymrophony

Can you only be a patriot if you speak Welsh fluently then? Do I not love my country because I haven’t properly learnt mutations?

Anonymous Twitter user

Chapter 2 has proposed a semiotic model of language as a resource for meaning-making, and described enregisterment as the process which turns linguistic systems into culturally coherent codes. In preparation for a detailed discussion of the enregisterment of Welsh orthography, this chapter explores Welshness, and the role played by language in its construction. This discussion is embedded in the sociopolitical history of Welsh and English as languages in contact.

Studies suggest that Cymrophones are thought of, by themselves and by non-Cymrophone Welsh people, as particularly patriotic and nationalistic (Bourhis et al., 1973), and that many Cymrophones derive a sense of identity at least partly by differentiating their Welshness from Englishness (Johnson & Giles, 1982). The act of speaking Welsh becomes imbued with political values as a consequence, and may be perceived by non-Cymrophones as a deliberate expression of separatist Welsh nationalism (Klein & Wardle, 2008). Understanding Wales as a former colony, whose culture and language have been threatened and repressed, but never erased, by English and British state power, helps to explain the ideological weight of language in modern Wales (Aaron & Williams, 2005; Bohata, 2004; Mac Giolla Chríost, 2005). Yet, as we saw in the introduction, Welsh people have always been both peripheral and central to imperialism; both colonizers and colonized (ap Gareth, 2009; Iyengar,

2015; M. Johnes, 2019).

For any construction of Welshness, the binary opposition of Welsh and English is crucial (M. Johnes, 2019). Other facets of identity, such as race, class, religion, and education, are mapped onto this binary in order to construct various visions of “authentic” Welshness. This chapter examines modern Welsh in context, with a focus on the development of language ideologies and their contribution to the construction of Welshness. In examining popular discourses of Welsh identity, I will show that ideological constructions of Welshness and Englishness have long been inherent to conceptions of the Welsh and English languages in Wales.

3.i “Nehemiah of the Vale of Clwyd”: the Fro Gymraeg and the languageness of the land

In 1889, Emrys ap Iwan¹ wrote an article entitled “Plicio Gwallt yr Hanner Cymry” (*Plucking the Hair of the Semi-Welsh*) (ap Iwan, 1889), decrying what he understood as a deterioration in the linguistic standards of contemporary Cymrophone writers. The title is an allusion to an episode in the Book of Nehemiah, in which the Jews of Jerusalem are castigated for marrying non-Jewish women and bearing children who cannot speak Hebrew. Nehemiah beats the men, pulling out their hair, and makes them promise that their children shall not marry non-Jews.

The article, which appeared in periodical *Y Geninen*, was signed “Nehemiah o Ddyffryn Clwyd” (*Nehemiah of the Vale of Clwyd*). The reference to Nehemiah places ap Iwan’s article in the biblical prophetic tradition, albeit in a somewhat tongue-in-cheek manner. It also founds his critique of modern Cymrophony upon a biblical precedent which imbues matters of linguistic propriety with broader moral and ethical meaning. Although ap Iwan’s text itself focuses almost entirely on his complaints about the standard of modern written Welsh, therefore, the figure of Nehemiah stands in the background, embodying a relationship between language preservation, religious salvation, and ethnic identity.

ap Iwan’s preferred type of Welsh is firmly located within the Welsh landscape. For him, the purest form of the language is “yr hen iaith Gymraëg, yr hon sydd etto

¹ The penname of Robert Ambrose Jones, a Methodist minister, author, and literary critic (Johnston, n.d.).

yn fyw y'nghilfachau y mynyddoedd”² (ap Iwan, 1889, p. 116). The mountains of Wales would soon come to represent a line of linguistic demarcation, as the north-western uplands beyond the Cambrian Mountains became the site of resistance to the increasing Anglicization of Wales as industrialization took hold (P. Gruffudd, 1995). In 1964, the Welsh nationalist activist and language campaigner Owain Owain, whose opinions about Welsh were discussed in Chapter 1.vi.d, introduced into Welsh political discourse the concept of *Y Fro Gymraeg* (Owain, 1964a; C. H. Williams, 1976), referring to the western and north-western areas of Wales, whose inhabitants have traditionally maintained higher levels of Cymrophony and a strong identification with a Welsh identity (R. Jones & Fowler, 2007). It delineates a core area of the country, held up as the locus of true Welshness (and consequently as the preferred focus of political activism) by prominent and respected Cymrophone thinkers of the earlier twentieth century such as Saunders Lewis (P. Gruffudd, 1995; R. Jones & Fowler, 2007).

For Owain, Wales without the Fro Gymraeg is not Wales at all: “Enillwn y Fro Gymraeg, ac fe enillir Cymru, ac oni enillir y Fro Gymraeg, nid Cymru a enillir”³ (Owain, 1964b). The Fro Gymraeg has since been a central ideological construct of Welsh nationalist thought. Balsom’s (1985) “Three-Wales Model” brought the concept of the Fro Gymraeg to greater prominence in Welsh political thought in the latter part of the twentieth century (Osmond, 2002). Building on previous conceptions of a politically and culturally divided Wales (D. G. Williams, 2014), the model contrasted the Fro Gymraeg with “Welsh Wales”, largely located in the industrialized valleys of southern Wales, where low levels of Cymrophony were nonetheless paired with a strong sense of Welsh identity; and “British Wales”, located along the Western border with England and in Pembrokeshire, whose inhabitants largely identify as British, rather than as Welsh (Balsom, 1985).

In early- to mid-twentieth-century Welsh nationalist thought, the Fro Gymraeg was not defined solely by its linguistic makeup, but also by its rural and agrarian character (P. Gruffudd, 1995; R. Jones & Fowler, 2007). As a result, the rural areas of Wales were idealized as the site of authentic Welshness (C. Williams, 2007, 2015), while the industrialized south of the country was perceived as Anglicized, a threat to

² *the old Welsh language, which lives yet in the recesses of the mountains*

³ If the Fro Gymraeg is won, Wales is won, and unless the Fro Gymraeg is won, what is won is not Wales.

Cymrophony (P. Gruffudd, 1995). Indeed, M. Johnes (2015, p. 683) argues that, in early-twentieth-century Wales, the sense of national identity amongst the non-Cymrophone industrial working classes themselves was “complicated by their belief that Welshness was rooted in a rural culture that belonged to the past”.

Agriculture and Cymrophony, land and language, were bound together in nationalist political thought, particularly that of the first half of the twentieth century. The Fro Gymraeg symbolized a highly idealized Welshness. On the other hand, an (equally idealized) American-style “melting pot” in industrialized southern Wales was envisioned by some internationalist Welsh Anglophone historians, who dreamed that the assimilationist spirit would eradicate class and cultural difference (D. G. Williams, 2014). For these thinkers, the agrarian Cymrophony of the Fro Gymraeg was the regressive ideal of a separatist cultural elite, and contrasted with the modernity of “Welsh Wales” (D. G. Williams, 2014).

3.i.a Blue Books and barbarism: constructing the gwerin as a Cymrophone archetype

Whether understood as an aspirational ideal or as a relic of the past, the association of Cymrophony with rural Wales results in establishing the rural Welsh “gwerinwr” (*common man, peasant*) as an archetypal persona (Agha, 2005) imbued with Welsh languageness, imagined as a guardian of authentic Welshness and authentic Cymrophony (Neal & Agyeman, 2006; C. Williams, 2007). In late-Victorian and early-twentieth-century Wales, the gwerin were often positioned as the true Welsh nation, whose Cymrophony attested to their authenticity (C. Williams, 2007).

As an attributive adjective, “gwerin” is often translated as *folk*, as in the compounds ‘cân werin’ (*folk song*) and ‘dawns werin’ (*folk dance*) (GPC *Ar-lein*, 2024, s.v. *gwerin*). As a noun, it means a class or group of people, often of humble social station; it is therefore, on the face of it, close in meaning to the English *peasantry*. Nonetheless, the gwerin need not be seen solely as a vulgar segment of the populace, but may equally be understood as the *citizenry*, those with whom political power would lie in a republic (“gweriniaeth”) (F. P. Jones, 1965).

The notion of the gwerin has complex roots which draw on Welsh political, intellectual, and religious ideologies. Stereotypes of religious nonconformism and political anti-conservatism were cemented in the Victorian era, when the Liberal Party gained a strong hold on Welsh politics, and issues such as temperance, educational reform, and the disestablishment of the Anglican Church came to

dominate the Welsh political agenda (R. M. Jones, 1992; D. D. Morgan, 1999).

Vicious English depictions of the Welsh as immoral, uncivilized, and ignorant had come to a head with the infamous official report whose publication came to be known as “Brad y Llyfrau Gleision” (*The Treachery of the Blue Books*, an allusion to the colour of the report’s binding). The Blue Books linked the supposed bestial ignorance of the Welsh to two primary factors: the use of the Welsh language, and religious nonconformism (S. Baker et al., 2014; H. G. Roberts, 2009). For some, the Blue Books were a spur to embrace the English language, rejecting the ignorant Cymrophone persona in favour of alignment with dominant English moral values (G. T. Roberts, 1996) and commitment to commercial and industrial modernity (Pryce, 2011). Others pressed the revival of the pseudo-historical scholarly and bardic tradition, largely invented by Iolo Morganwg (Constantine, 2007), as the key to asserting the Welsh claim to civilisation and literacy (Pryce, 2011). However, the pagan resonances of Welsh (and Cymrophile English) antiquarians’ bardic and folkloric traditions made them an uncomfortable fit with the nonconformist worldview (G. T. Roberts, 1996).

Amongst rural Cymrophones, the dominant identity constructed in response to stereotypes of barbarism was one that thoroughly rejected English cultural and linguistic hegemony (G. T. Roberts, 1996): Toryism and Anglicanism were repudiated in favour of Liberalism and nonconformism. The construction of the gwerin was rigorously socially conservative, however, conforming to an English model of respectability which was ironically very like the one that the Blue Books’ authors had themselves sought to impose (H. G. Roberts, 2009; Sheppard & Byrne, 2017). If anything, the Welsh outdid the Victorian English in the performance of piety (D. D. Morgan, 1999).

Education and literacy were highly valued by the gwerin (S. Baker et al., 2014) as keys to piety, industriousness, and self-improvement, characteristic pillars of British Victorian morality (O’Malley et al., 1995). The gwerinwr, “imbued with the best qualities of Welsh Liberal-Nonconformity; sober, literate and honest” was “a culmination of the history of Wales” (Lynch, 2007). Nonetheless, as R. M. Jones (1992, p. 339) argues, the Welsh identity embodied in the gwerin was never a universal one. Rather, it was “based on an assumed and proclaimed homogeneity” which limited true Welshness to those with particular religious, linguistic, and political allegiances.

The growth of the Labour party as the dominant force in Welsh politics after the First World War gave rise to an increasingly secularized, socialist, working-class

identity, epitomized by the coal miners of southern Wales, as a competing model of Welshness (R. M. Jones, 1992; D. D. Morgan, 1999). The Welsh language was uncoupled from this new idea of Welshness, not only because of the comparatively greater proportion of Anglophones in the industrial south, but also because “English was deemed to be a much more appropriate means of expression” for working-class solidarity which transcended national borders (D. D. Morgan, 1999, p. 333). The gwerin of rural Wales therefore remained as the primary Cymrophone Welsh identity, increasingly seen as puritanical, old-fashioned, and reactionary (D. D. Morgan, 1999).

3.i.b Clans and councillors: the ethnic/civic divide in modern Welsh political thought

The connection of Cymrophony and a traditional, conventional, and respectable rural society has remained central to the political ideals of some Welsh nationalists (R. Jones & Fowler, 2007; D. G. Williams, 2014). Nonetheless, a faultline has become increasingly prominent in contemporary nationalist thought. The nationalists for whom the Fro Gymraeg and the preservation of Cymrophony remain a focus of their activism are sometimes placed as “ethnic” nationalists, as distinct from the “civic” nationalists whose concern is with issues facing Wales as a geopolitical unit, and who make up the majority of contemporary mainstream Welsh nationalist politicians (Brooks, 2010; R. Jones & Fowler, 2007).

Although the usefulness of this binary taxonomy of nationalism, introduced by Kohn (1944), has been challenged (e.g. Bugge, 2022; Kaufmann, 2019; Tamir, 2019), the terms remain ubiquitous in modern political discussion (Kuzio, 2002). Civic nationalism is constructed as democratic, liberal, and enlightened (Tamir, 2019), encouraging individuals to think of themselves as citizens and to develop a political consciousness (Kohn, 1944). Ethnic nationalism is instead grounded in the past, encouraging ties founded (or supposedly founded) in genetic relationships and in shared cultural traditions (Kohn, 1944). Kohn’s distinction therefore provides a convenient means for the nationalist parties of both Wales and Scotland to differentiate their “civic” ideology from illiberal, racist, “ethnic” thought (McAngus, 2015; McBride, 2022).

As Brooks (2010, 2021) notes, the Welsh language is often situated as an ethnic marker within this model of nationalism. This, he argues, is the inevitable result of the minoritized status of Welsh culture and language (Brooks, 2010). The English

language gains a status as universal and multicultural because it is enforced as the default mode of civic belonging in the British state, while Welsh, restricted to a segment of the population, is marked as monoculturally ethnic and, therefore, as exclusive and exclusionary (Brooks, 2021).

Proponents of English as a lingua franca position the language as a passport to educational, professional, and economic success, as well as highlighting the convenience and efficiency for institutions of having a single official language. Lying behind much of this support for English are instrumentalist assumptions about language, which emphasize ease of communication, efficiency, mobility, and economic advantage as the primary goals of language policy (May, 2012). In contrast, minoritized languages are regularly constructed as carriers of identity, with little instrumental value. This dichotomy parallels the civic/ethnic distinction. Dominant, majority languages are understood as bolstering the efficiency of civic institutions, and aiding individuals to become successful, integrated members of society. Minoritized languages, in contrast, are seen merely as markers of ethnic identity, which fuel social divisions and impose needless economic burdens on the state (May, 2012).

Contemporary immigration policy in the United Kingdom requires immigrants to demonstrate knowledge of English⁴. Proponents of these language requirements represent English as a tool of individual integration and national cohesion (Blackledge, 2009a), yet this civic depiction of English coexists with an ethnic one which racializes the other languages of immigrants and associates the English language with British cultural values (Blackledge, 2006). To a large extent, ethnic and civic framings of language are not mutually exclusive. The civic arguments for dominant, majority languages are not the only ones available, but the very dominance of these languages means that they are normalized as “the unmarked (linguistic) category” (May, 2012, p. 219), allowing their users to overlook the identity-related functions they perform. Civic and ethnic approaches to language are

⁴ The current requirements pay lip service to other languages, offering applicants the opportunity to demonstrate knowledge of either Welsh or Scottish Gaelic, instead of English (KoLL, 2024, p. 5). In practice, however, English is the expected language for establishing an immigrant’s suitability: no formal criteria are provided for demonstrating proficiency in Welsh or Scottish Gaelic, whereas there are detailed regulations specifying how English-language skills can be demonstrated. Seeking to remedy this lack of clarity, a 2019 Freedom of Information request asked the Home Office to describe the requirements for proving proficiency in Welsh or Scottish Gaelic for the purposes of an application for British citizenship by naturalization (Ray, 2019). The Home Office refused to provide an answer, deeming this not to be an appropriate FoI request.

two sides of the same coin (Fozdar & Low, 2015), with civic arguments for language testing for immigrants, for example, described as “the language of racism [...] dressed in the clothes of liberal, educated, articulate, common-sense discourse” (Blackledge, 2009b)

The civic-nationalist criticisms aimed at the languages of immigrant communities are broadly similar to those aimed at the United Kingdom’s Celtic languages (Blackledge, 2002). The promotion of monolingualism by the British government privileges English at the expense of all other languages, not only those with their origins outside the United Kingdom. In response, modern Welsh political discourse regularly uses civic framing to promote the Welsh language, not only as a language of national cohesion (see section 3.ii below), but also in economic terms, stressing that bilingualism can be “yn fantais wirioneddol yn economaidd”⁵ (*Record of Proceedings*, 2020), and “bod hyn yn sgil, bod hyn yn agor drysau, efallai, yn ddiwylliannol, yn ychwanegol i’r byd gwaith, wrth gwrs, ar lefel broffesiynol”⁶ (*Record of Proceedings*, 2022a). Although the neoliberal and consumerist values embedded in this representation of minoritized languages concerns some critics (Barakos, 2016; Mac Giolla Chríost, 2005), others see potential for the institutionalization of Welsh to protect the language as an integral part of the civic state (May, 2000).

Wales at the turn of the second millennium has been described as “a place with citizens, not a cause with adherents” (R. M. Jones, 1992). The use of a civic framing in political discourse implicitly rejects the association of Welsh with ethnic nationalism, instead insisting that the status of the language rests on its instrumental role in civic society. The increasing civic equality of status between English and Welsh brings its own problems, however. Putting minoritized languages on an even political and legal footing with a dominant language is often perceived by opponents as discrimination against those who cannot use the minoritized language, and particularly as an unfair disadvantage in employment (May, 2012), as we shall see.

Nationalist monolingualism (Blackledge, 2002), the standard language ideology of modern European (Van Avermaet, 2009) and British (Blackledge, 2002, 2009c)

⁵ *a genuine advantage economically*

⁶ *that this is a skill, that this does open doors culturally, in addition to the workplace, of course, on a professional level*

All translations of quotations from the Senedd’s Record of Proceedings are the official ones provided in the Record itself.

political policy, imagines that civic unity can only be cemented by the use of a single, common language. Multilingualism, whether officially sanctioned or not, threatens to fragment the nation. From the Treachery of the Blue Books to modern claims of Cymrophones' unfair advantages, the use of Welsh has been depicted by monoglot Anglophones as subversive, immoral, and inimical to British identity.

We have already seen how Cymrophones and non-Cymrophones respond to the language-mixing behaviours of Welsh-English bilinguals. This illustrates the ways in which broader ideas and ideals of Welshness are implicated in language attitudes. The remainder of the chapter will explore some of the ways in which Welshness manifests itself. The question of who may claim authentic Welshness, and Welsh-languageness, is inseparable not only from constructs of race, but also from others such as national origin, place of residence, family background, and linguistic competence.

The lines separating Welshness from Englishness, and Cymrophony from Anglophony, are not necessarily coextensive with one another, or with the borders separating Wales from England. Wales is an internally-divided geopolitical landscape, whose ethnic, linguistic, and cultural borders are perpetually contested from within and from without. As such, Wales is not a unified ideological construct; rather, it is a palimpsest of overlapping landscapes, each inscribed with a particular ideal of Welshness. Whether we imagine this Wales as fragmented or as a plurality, awareness of the nation's multiplicity provides essential context for the discussion that follows.

3.ii Who speaks Welsh? Sociolinguistic and raciolinguistic ideologies of Welshness

For some, the Welsh language has been and remains the only truly *Welsh* language (Penhallurick, 1993). English, for these people, indexes England specifically, and not the entirety of the United Kingdom. Children who use English in a Welsh-medium school, like adult Cymrophones whose Welsh is mixed with English words and phrases, may therefore be left feeling that they are not genuinely Welsh at all when their normal bilingual behaviours are stigmatized (Selleck, 2013). Cymrophones whose natural variety features language mixing are left in a position where their right to claim a Welsh national and linguistic identity is questioned, as a result of the enregisterment of an idealized, monolingual Welsh as the true language of Welshness.

3.ii.a Welshness and whiteness: raciolinguistic constructions of Cymrophony

The association of Cymrophony with the *gwerin* persists in modern Wales. In contemporary discourses, rural Wales evokes Welshness, whiteness, and cultural Christianity; symbols of multiculturalism, such as Islamic religious buildings, are conspicuously absent from rural Welsh landscapes (Dafydd Jones, 2010). As a result, Cymrophony is also implicated in this image of conservatism and monoculturalism. Even if the modern *gwerin* are not necessarily imagined in quite the same terms as were the agrarian workers of Victorian times, rural Wales has not yet fully broken free of the picture of reactionary conservatism constructed in that era. Scourfield and Davies (2005), reporting qualitative research about views of Welshness and ethnicity amongst children aged 8–11 in Wales, interviewed a young Muslim girl attending an English-medium school in Cardiff. The child stated that she would not want to attend a Welsh-medium school because the children there “might be racists or something” (Scourfield & Davies, 2005, p. 91).

If it is uncoupled from Cymrophony, Welshness can safely be positioned as a “civic” identity, grounded in a sociopolitical conception of citizenship rather than in membership of an ethnolinguistic tribe. Cymrophony, on the other hand, remains an “ethnic” identity, both as a marker of the “indigenous” culture of Wales (Kohn’s (1944) sense of “ethnic”), and in the common modern sense, which equates “ethnic” with “racial”. Hafferty (2020) reports that being of a white British background, identifying as Christian, and living in a rural area were significant predictors not only of self-reported ability to use Welsh, but also of self-reported use of Welsh in everyday life. The identification of Welshness with the overwhelmingly white Cymrophone *gwerin* of rural Wales may not be intended as racially exclusive (C. Williams, 2007), but it provides a foundation for unthinking or deliberate racism to flourish, for doubting the Welshness of Cymrophones of colour, and for people of colour to feel (or actually to be) excluded by policies which encourage Welsh-language use (Welsh Arts Anti Racism Union, 2021; Welsh Language Commissioner, 2020; C. Williams, 2007).

Despite white Cymrophones’ assertions that Welshness is open to all who call the country home, the public perception is often that Welshness is a narrower, ethnolinguistic category (Brooks, 2010, 2021; Jackson & Jones, 2014). Another of Scourfield and A. Davies’s (2005, p. 90) respondents, a child attending a Welsh-medium school in the valleys of southern Wales, asserted that he would not

want to be Black because “I’d probably be an English person then”. A girl from the same school, asked whether children in English-medium schools were different in any way from her own schoolmates, answered “Mae nhw’n gallu bod yn plant du”⁷ (Scourfield & Davies, 2005, p. 92).

Raciolinguistic ideologies (Alim et al., 2016; Rosa, 2018), centering on the assumption that Cymrophones will be white, are the inevitable result of the whiteness of modern Cymrophone society. Such ideologies mean that Cymrophones of colour regularly experience scepticism about their linguistic identity (BBC One, 2022; Critically Speaking, 2020; D. Davies, 2020; S. Jones, 2020). Cymrophone culture may, theoretically, be open to anyone who engages with the language. In practice, however, the whiteness of this culture is self-sustaining, even when it is not necessarily deliberately or overtly racist. The obvious demographic imbalances in Cymrophone culture, regardless of their cause, limit the social range (Agha, 2015) of the Welsh language, promoting only a fairly restricted set of personae which can easily be performed using Welsh (Agha, 2005).

Recent racist incidents in prominent Cymrophone cultural locations, including the National Eisteddfod (E. L. Jones, 2018), Cymrophone television channel S4C (Anffyddiaeth, 2012; Golwg360, 2012), and a gig by legendary Cymrophone singer Meic Stevens (BBC Cymru Fyw, 2019; Golwg360, 2019) have displayed an unthinking white ethnocentricity in Cymrophone culture. There is an unspoken assumption of a monocultural audience whose identity is traditional white Welshness, for whom stereotyped mimicry of ethnic minorities or casual remarks about Africans being “anwariaid”⁸ are amusing because they do not touch upon their own identity. This is an imagined audience whose sympathies are likely to be constructed in an insular context of cultural and ethnic homogeneity, in which the concerns of people of colour are subordinate, because they are imagined to exist outside the community.

The establishment of a specifically Cymrophone identity in Wales is therefore far more complex than merely identifying linguistic skills. Rather, it is tied up with broader questions of race, ethnicity, and national identity. Understanding what Welshness means on an individual, social, and political level is an essential component in understanding the construction of linguistic Welshness. The nature of the relationship between Welsh and English people, and the nature of the animosity

⁷ *They could be black children*

⁸ *savages*

between Cymrophone and Anglophone communities, will prove to be crucial for explaining Cymrophone attitudes towards English-origin ELUs.

In the late-twentieth and early-twenty-first centuries, Cymrophones on average had lower unemployment levels, and held a high proportion of jobs in media, government, and the public sector in Wales (Day, 2002; Drinkwater & O’Leary, 1997). This, in turn, has led to the perception of an urban Cymrophone elite, derisively nicknamed the *Taffia* (C. Lewis, 1995), who wield disproportionate cultural power over the lives of those who do not speak the language (M. Johnes, 2012; A. Thomas, 1997). Chairman of the Swansea Bay Racial Equality Council Naz Malik, condemning policies to require some level of Welsh-language skills in certain jobs, complained that “the English are being discriminated against in employment either because they are English or because they cannot speak Welsh”, and that “this is racism” (Hastings, 2000). The success of efforts to promote the use of Welsh in professional contexts places the language in a double bind described by May (2012, pp. 220–221) as common to minoritized languages, with opponents mocking them “for their lack of utility, and then opposing their utility when it proves to be politically inconvenient”.

As a result of discourses equating Welsh-language revitalization with anti-English racism, Cymrophones may fear that their Cymrophony, in and of itself, will be interpreted as anti-English or as racist. Madoc-Jones and Parry (2012) describe some of their respondents’ use of “non-extremist positioning”. One man, for example, who describes using English very rarely as most of his social circle is Cymrophone, nonetheless explains that he is happy to speak English with non-Cymrophones: “I don’t you know exclude them or anything racist” (Madoc-Jones & Parry, 2012, p. 173).

On the other hand, the repeated claim that anti-Cymrophone bigotry is “the last acceptable form of racism” (Youde, 2018) can be illustrated with examples from Twitter. “Pam bod hiliaeth at y Gymraeg dal yn dderbyniol?!”, asks one user. Another recounts a complaint against residents in a nursing home speaking Welsh, asserting that “Tasa hyn yn unrhyw iaith arall, ma’n siwr fase fo’n cael ei drin fel hiliaeth”¹⁰. This discourse is “dominated by White people’s ideas of what it means to be Welsh” (Youde, 2018). It centres a white Welshness which seems unaware that

⁹ “Why is racism towards the Welsh still acceptable?”

¹⁰ “If this were any other language, surely it’d be treated as racism.”

true racism, far from being treated as “unacceptable” by society at large, is still a daily threat to non-white people, systemically and on the level of individual interactions. Both Anglophones and Cymrophones in Wales construct their linguistic identities in ethnic terms when doing so permits them to use the language of racism to give weight to their grievances. Not only does the use of an ethnic framing show insensitivity towards the experiences of people of colour, however. It also encourages the oppositional construction of Welsh and English as mutually exclusive and immutable categories of identity, denying any potential for the two to meet and meld in novel ways.

3.ii.b Chauvinism and sharing: the competition for authentic Welshness

Aitchison and Carter (1994, p. 3) claim to find a thread of ethnolinguistic chauvinism in Cymrophone discourse which asserts that “to be Welsh, in any meaningful way, a person must speak, or at least understand, Welsh”. The corpus of Twitter messages collected for this study certainly shows examples of such attitudes amongst contemporary Cymrophones¹¹:

- (1) a. Be am i ni ddechre eu dysgu yn yr ysgolion nad ydy neb 100% yn Gymro os nad ydy nhw’n siarad Cymraeg.

Why don’t we start to teach in the schools that nobody is 100% a Welshman if they don’t speak Welsh.

- b. Esiamplau o #oxymorons “Compassionate banke”, “cyfalafiaeth ddemocrataidd”, “Cymru ddi-Gymraeg”.

Examples of #oxymorons “Compassionate banke[r]”, “democratic capitalism”, “non-Cymrophone Welsh people”.

Some of Madoc-Jones and O. Parry’s (2012) Cymrophone respondents expressed similar views, with one man asserting “If you can’t speak Welsh you’re not Welsh at all to me” (Madoc-Jones & Parry, 2012, p. 170). Another respondent questioned

¹¹ In ex. 1b, the author has used a spelling which, in standard Welsh, refers to the name of the country, “Cymru” (*Wales*), not the demonym “Cymry” (*Welsh people*). It is therefore possible that they wished to convey the meaning “non-Cymrophone Wales”; however, given the established nature and frequent use of the phrase “Cymry di-Gymraeg” (*non-Cymrophone Welsh people*), it seems much more likely that this was the intended meaning.

whether Cymrophones who choose to speak English are truly Welsh (Madoc-Jones & Parry, 2012, p. 171).

Nonetheless, the exclusionary definition of Welshness as reliant on Cymrophony is by no means the only one expressed in modern Cymrophone discourse. The Cymrophone press has, at times, given a platform to those who condemn such attitudes as “chwilio am rwystr o hyd i’w osod rhyngom ni a’r Cymry sy’n methu neu’n dewis peidio siarad yr iaith”¹² (L. Ellis, 2024). For some, promoting and strengthening Welsh requires Cymrophones to embrace their shared existence with non-Cymrophones, stepping “allan o’r cilfachau ac mewn i ganol [eu] cymunedau”¹³ (Golwg360, 2018), in the words of community organizer Einir Siôn.

In the field of public policy, Members of the Senedd have stressed the importance of an inclusive understanding of Cymrophony and of Welshness, which encourages people to learn Welsh, to use Welsh, and to support Welsh. Samuel Kurtz, MS, noted in 2021 that “am rhy hir, mae yna wedi bod meddylfryd o ’ni a nhw’ o ran yr iaith Gymraeg”¹⁴, separating Cymrophones who consider themselves non-fluent, including learners, from those who can use the language confidently (*Record of Proceedings*, 2021a). In the same vein, Jeremy Miles, MS, has argued that one need not speak fluent Welsh “er mwyn teimlo’r cynhesrwydd yna tuag at y Gymraeg a’r synnwyr bod y Gymraeg yn perthyn i ni i gyd”¹⁵ (*Record of Proceedings*, 2024). This idea of the language as a patrimony which all Welsh people hold in common informs official approaches to defining Cymrophony. Answering a question about how the Senedd defines a “Welsh speaker”, Jeremy Miles, at the time the Welsh Government’s Minister for Education and Welsh Language, stated that “Welsh belongs to us all, and part of that is recognising that there is a range of confidence and a range of competence, and that, whatever level of proficiency you have, you’re entitled to call yourself a Welsh speaker” (*Record of Proceedings*, 2023a).

The position often repeated by Senedd members is that, in the words of Rhun ap Iorwerth, MS, “mae eisiau cefnogi a dathlu pob ffordd mae pobl yn defnyddio’r

¹² constantly looking for a barrier to place between us and the Welsh people who can’t, or who choose not to, speak the language

¹³ out from the hideouts and into the centre of [their] communities

¹⁴ for too long there’s been a mindset of ‘us and them’ in terms of the Welsh language

¹⁵ in order to feel that warmth towards the Welsh language and have that idea that the Welsh language belongs to us all

iaith”¹⁶ (*Record of Proceedings*, 2022b). Rejecting the “us and them” mindset which separates people by their fluency in Welsh, Samuel Kurtz argued that the aim should be to help people “deimlo’n gyffyrddus i siarad Cymraeg, beth bynnag eu safon nhw, heb ofni am gamgymeriad bach fan hyn a fan co”¹⁷ (*Record of Proceedings*, 2021a). During a 2021 one-day inquiry on the Welsh language in the Senedd, Caryl Haf, representing the Wales Federation of Young Farmers Clubs, gave the following contribution:

Dwi’n gweithio mewn siop amaethyddol. Mae gen i nifer o bobl sy’n dod mewn yn dysgu’r iaith, ond mae yna bobl yn cadw i’w cywiro nhw yn ormodol. Mae nhw yna i drïo, a dwi’n credu ei bod hi’n bwysig ein bod ni yn pwysleisio bod y Gymraeg ar gael i bawb.¹⁸ (*Record of Proceedings*, 2021b)

As we shall see, linguistic prescriptivism is an issue often raised in popular discourse as a barrier to the growth of Cymrophony. Mike Hedges, MS, told a plenary session of the Senedd that “I speak Welsh every day, but rarely in debates, and I only speak to people I am confident will not try and trap me with my lack of fluency” (*Record of Proceedings*, 2023b). A strong official commitment in Welsh politics to making Welsh accessible and to encouraging its use seems, therefore, to face practical obstacles from within the Cymrophone community itself, with people lacking confidence to use the language because of a fear of being ridiculed or censured by fellow Cymrophones.

¹⁶ *we need to support and celebrate the ways in which people use the language*

¹⁷ *to feel comfortable in speaking Welsh, whatever their level, not worrying about the odd mistake here and there*

¹⁸ I work in an agricultural supply store, and I have a number of customers who are Welsh learners, but people are constantly correcting their language. But they’re doing their best, and it’s important that we support those people and emphasise that the Welsh language belongs to us all.

3.iii “I dunno if she was Welsh in the sense of a Welshwoman”: the porous semantics of Welshness and Cymrophony

Caryl Haf’s assertion in the Senedd that “Does dim rhaid bod yn rhugl yn y Gymraeg i fod yn Gymro”¹⁹ (*Record of Proceedings*, 2021b) faces a number of challenges, both direct and indirect, in popular Cymrophone discourse. The relationship between Welsh identity and the Welsh language is complex and varied, built up from both explicit and implicit discourses about language, linguistic standards, and authenticity. As the next section illustrates, the Welsh language itself differs from English in its terminology for linguistic identity and ethnic, civic, or geographical identity. For Cymrophones, talking about Welshness and Cymrophony as discrete phenomena is often a matter of some complexity.

3.iii.a Cymraeg and Cymreig: Welsh semantics and the semantics of “Welsh”

The semantic range of the English adjective “Welsh” is covered by two distinct words in standard modern Welsh: “Cymraeg” refers specifically to the Welsh language, while its etymological doublet “Cymreig” refers to Welsh national, geographic, or cultural identity²⁰. This allows Cymrophony and Welshness to be neatly separated, as in the following examples (collected from the Twitter corpus described in Chapter):

- (2) a. Pwysigrwydd creadigrwydd Cymreig (boed yn Gymraeg neu’n Saesneg).
The importance of Welsh creativity (whether Cymrophone or Anglophone).
- b. Am le mor Gymreig, pam does dim enw Cymraeg ar Cross Hands?
For such a Welsh place, why doesn’t Cross Hands have a Welsh-language name?

Nonetheless, the strict semantic separation of the two terms for *Welsh* is not always maintained in colloquial use. As the following examples (again taken from Twitter)

¹⁹ *You don’t have to be fluent in Welsh to be Welsh*

²⁰ Similar distinctions are made for some other language/nationality pairs: “Saesneg” (*English-language*) and “Seisnig” (*English by geography, nationality, etc.*); “Ffrangeg” (*French-language*) and “Ffrengig” (*French by geography, nationality, etc.*); etc.

show, it is not uncommon to find the term “Cymraeg” where the context requires “Cymreig” in standard Welsh:

- (3) a. ma bwytaï Cymraeg yn cwyno am y diffyg o bysgod Cymraeg sydd ar gael.
Welsh restaurants are complaining about the lack of availability of Welsh fish.
- b. Bi-product o paratoi'r pwdin dolig. 3 a hanner can o stowt Cymraeg.
By-product of preparing the xmas pudding. 3 and a half cans of Welsh stout.
- c. onibai am Abertawe, Caerdydd, Casnewydd a Wrexham, ma pobl Cymraeg yn cefnogi timoedd o Loegr.
apart from Swansea, Cardiff, Newport and Wrexham, Welsh people support teams from England.
- d. Gret i wylïo a gwaeddu dros Paralympwyr Cymraeg Beverly Jones a Tracey Hinton.
Great to watch and shout for Welsh Paralympians Beverly Jones and Tracey Hinton.
- e. Fi just ishe gwbo pwy nath adel i Bill Nighy gor neud acen cymraeg yn hari poter a'r dethli halows pt 1.
I just wanna know who let Bill Nighy over-do a welsh accent in harry potter and the deathly hallows pt 1.

In each of these examples, the intended meaning is “Welsh by nationality, citizenship, geographical origin, or location”, rather than “Cymrophone”.

The use of “Cymreig” where standard usage prescribes “Cymraeg” is rather rarer, but does occur on occasion:

- (4) a. mae cleifion ein Wardiau Adsefydlu [...] wedi dysgu geiriau Cymreig newydd
the patients on our Rehabilitation Wards [...] have learned new Welsh words
- b. bydd yr holl recordiadau ar gael gydag isdeitlau Cymreig / Saesneg
all the recordings will be available with Welsh/English subtitles

Despite the apparently clear terminological distinction in standard Welsh, examples of colloquial usage suggest that the differentiation of “Cymraeg” and “Cymreig” is not always firm. This may reflect transfer from English, which has only the single word “Welsh” to cover the semantics of “Cymraeg” and “Cymreig”.

Whatever the reason, for at least some Cymrophones in some contexts, “Cymraeg” (or, occasionally, “Cymreig”) has the same semantics as “Welsh”, rather than the narrower meaning prescribed in standard Welsh.

The confusion is reflected in the fact that the same corpus contains numerous tweets explaining the distinction between the terms:

- (5) a. Mae ‘Cymraeg’ yn cyfeirio at yr Iaith, a dim ond hynny. Mae nhw’n gallu bod yr un mor Gymreig â ti neu fi heb siarad Cymraeg, ond fydden nhw ddim yn ‘Gymraeg’.
- ‘Cymraeg’ refers to the Language, and that alone. They can be just as Welsh as you and me without speaking Welsh, but they would not be Cymrophone.*
- b. Mae angen i ti ddysgu fod ’na wahaniaeth rhwng Cymraeg a Chymreig.
- You need to learn that there’s a difference between Cymrophone and Welsh.*

In addition, there are a number of tweets pointing out that someone appears to have used the incorrect word:

- (6) a. Curry Cymraeg? Cyri Cymreig? [...] Mewn penbleth yma. Ydy’r ‘curry’ ’na yn siarad Cymraeg?
- Cymrophone Curry? Welsh Curry? Confused here. Does this ‘curry’ speak Welsh?*
- b. Buchod Cymreig neu buchod Cymraeg? (Neu buchod o Gymru...)
- Welsh cattle or Cymrophone cattle? (Or cattle from Wales...)*

The spoken data in the CorCenCC corpus include the following exchange between two speakers²¹ (Knight et al., 2020):

- (7) <Siaradwr 1> O’dd hi’n Gymraes ’lly?
- <Speaker 1> Was she a Welshwoman, then?*
- <Siaradwr 2> Na wel nid... doedd hi ddim ’n siarad Cymraeg...
- <Speaker 2> No well not... she didn’t speak Welsh...*
- <S1> ...ia.
- <S1> ...yeah.

²¹ The original CorCenCC text has been re-formatted for ease of reading, since details included by CorCenCC (such as background noises or pauses) are not relevant to this discussion.

<S2> 'Dwn i'm os o'dd hi'n Gymraeg o ran yn Gymraes...

<S2> *I dunno if she was Welsh in the sense of a Welshwoman...*

<S1> Ond doedd hi ddim 'n...

<S1> *But she didn't...*

<S2> ... doedd hi ddim 'n siarad Cymraeg 'lly.

<S2> *...so she didn't speak Welsh.*

This exchange illustrates the potential ambiguity of adjectives meaning “Welsh” and “Cymrophone” in colloquial use. Having stated that the woman was not a Welsh speaker, Speaker 2 then says “Dwn i'm os o'dd hi'n Gymraeg o ran yn Gymraes” (*I dunno if she was Welsh in the sense of a Welshwoman*). The term “Cymraeg” here should, according to standard usage, mean “Cymrophone”, yet Speaker 2 is using it to mean “Welsh by nationality”.

Something more than the ambiguity of Welsh words for “Welsh” is revealed here, however. The confusion which permeates this extract also shows the speaker's uncertainty about the interpretation of the word “Cymraes”. This uncertainty is difficult to understand in English translation. The Welsh words “Cymro” and “Cymraes” refer, respectively, to male and female Welsh people. As such, they are often translated as “Welshman” and “Welshwoman”. However, the Welsh terms are semantically broader than their English counterparts, as they may also mean “Cymrophone”.

Speaker 2 struggles to answer clearly, because it is not clear whether Speaker 1 is asking “was she a Welshwoman?” or “was she a Cymrophone?”. Speaker 2's repetition of the term “Cymraes” when they say “Dwn i'm os o'dd hi'n Gymraeg o ran yn Gymraes” suggests that they would naturally consider the word to be a term for a woman who has Welsh national or cultural identity, rather than for a Cymrophone woman specifically. Yet they are clearly aware of the ambiguity, and their initial response, “Na wel nid... doedd hi ddim 'n siarad Cymraeg...” contradicts the later use of “Cymraes”, implying that the speaker's immediate interpretation is that this is a question about language, not about Welshness in general.

The terminology of language and ethno-national identity is intimately intertwined in the Welsh language. The terms “Cymro”, “Cymraes”, and the gender-non-specific plural form “Cymry” are regularly used to mean either “Welsh” or “Cymrophone”. In some cases, context will determine the intended meaning, as in the following examples from Twitter, which refer to people who are known not to be Cymrophones:

- (8) a. Bale, Rush, Joey Jones, Giggs. Oes 'na Gymro arall wedi chwarae yn y Ffeinal?
Bale, Rush, Joey Jones, Giggs. Has any other Welshman played in the Final?
- b. Braint oedd cael cynnal cwrs Cymraeg i ffoduriaid a cheiswyr lloches yng Nghaerdydd yr wythnos yma. Dyma ddau Gymro newydd yn cael eu gwers gyntaf.
It was an honour to provide a Welsh course for refugees and asylum seekers in Cardiff this week. Here are two new Welshmen having their first lesson.

In many other cases, however, it is unclear whether the meaning is “Welsh person”, “Cymrophone”, or a combination of the two:

- (9) a. Y Gymraes @alysconran yn cyrraedd rhestr hir Gwobr Ryngwladol Dylan Thomas.
The Welshwoman/Cymrophone @alysconran reaches the International Dylan Thomas Prize longlist.
- b. Dyma be 'ma Cymro go iawn yn edrych fel! Person sy' di dewis y wlad 'ma a dysgu'r iaith.
This is what a real Welshman/Cymrophone looks like! A person who's chosen this country and learned the language.

The terms “Cymro” and “Cymraes” slip semantically between the linguistic, the national, and the geographic²². As a result, Speaker 2 in Ex. 7 asserts that the woman being discussed is certainly not a Cymraes, yet may nonetheless be a Cymraes: she does not speak Welsh, but she may have Welsh national or cultural identity.

²² This is not only true of words referring to the Welsh people and their language. The word “Sais”, for example, was historically used fairly broadly and could refer to an Englishman, or to a Welshman who either spoke English, had English ancestry, had lived in England, or had some political connection with England (Fowkes, 1981; Suppe, 2022). As late as 1941, Cymrophone poet Euros Bowen advised young Welshmen wishing to study philosophy and theology “nid yw bod yn Sais da yn anhepgor i gynhyrchu athronydd neu ddiwinydd o Gymro” (*being a good Anglophone is not essential to producing a Welsh philosopher or theologian*) (Bowen, 1941). In contemporary Welsh, however, it is rare to find “Sais” (or its feminine counterpart, “Saesnes”, or plural, “Saeson”) used in a neutral way to refer to a Welsh person’s linguistic ability or family background. Instead, these words tend to be insulting or disparaging when applied to Welsh people (Selleck, 2013), as when the poet R. S. Thomas, a Welsh nationalist with an extreme aversion to the Anglicization of Wales, referred to non-Cymrophone Welsh people as “Sais-Gymry” (quoted in V. Hughes, 2024).

3.iii.b Lack and legitimacy: the semantics of (non-)Cymrophony

The usual term for Cymrophones in modern Welsh is “Cymry Cymraeg”, while non-Cymrophones are called “Cymry di-Gymraeg”. In standard use, “Cymry Cymraeg” has the unambiguous meaning “Cymrophone Welsh people” (more literally, “Welsh-language Welsh people” or “the Welsh people of the Welsh language”), and not the tautologous “Welsh Welsh people”. The “Cymry di-Gymraeg” are therefore *non-Cymrophone Welsh people*. The pair of terms notionally establishes both Cymrophones and non-Cymrophones equally as *Cymry*; as Welsh people.

However, the pairing of these terms contains subtle biases. The alliterative symmetry and rhetorical balance of the choriambic *Cymry Cymraeg* lends an aesthetic force to the suggestion that “Cymraeg” is the natural consummation of the identity “Cymry”. The slippage in meaning of *Cymraeg* between “Cymrophone” and simply “Welsh” in colloquial language only adds to the sense that Cymrophony and Welshness are naturally synonymous.

Furthermore, to translate the term “Cymry di-Gymraeg” as “non-Cymrophone Welsh people” does not fully capture the semantics of the Welsh negative morpheme “di-”. English “non-” is, in general, a “strictly negative” morpheme (Lieber, 2004, p. 114). When added to a noun or adjective, “non-” usually expresses simply the opposite or negation of the unprefix word (Bauer et al., 2013; Lieber, 2004). Welsh “di-” often has a more specific semantic content, such as privation or pejoration, as in “di-gartref” (*homeless*), “digalon” (*disheartened*), “digywilydd” (*shameless*), and “di-lun” (*misshapen, deformed, or shapeless*). Lacking a fundamental characteristic of the poetically balanced “Cymry Cymraeg”, the “Cymry di-Gymraeg” might equally well be called “Welshless” as “non-Cymrophone”.

On the one hand, the designation “di-Gymraeg” has a political resonance, positioning these as people who have been deprived of a language. On the other, however, it defines non-Cymrophones in terms of what they lack. This definition also devalues the importance of Welsh beyond the Cymrophone community. Literate adults living in contemporary Wales will see the language all around them in the landscape. People who received state education in modern Wales are likely to have had at least some instruction in Welsh. The Welsh national anthem and Welsh-language songs play an integral ceremonial role in local and national sporting events which, Carolyn Thomas MS told a Senedd Plenary session, “has fostered a

sense of pride, I believe, and a desire to be part of the Welsh-speaking and Welsh-singing community as well” (*Record of Proceedings*, 2024).

The designation “di-Gymraeg” ignores this intimacy with the Welsh language which a modern Welsh existence brings. Cymrophone bard and academic Mererid Hopwood calls the term “yn broblemus ar sawl lefel”²³ (Hopwood, 2019). She notes that the English spoken in Wales shows plentiful evidence of Cymrophone influence, and that calling non-Cymrophones “di-Gymraeg” denies the role of the Welsh language in forming a distinctive Welsh ethnolinguistic identity. By negating the presence of Cymraeg in the formation of the Cymry di-Gymraeg, the term negates a powerful force which in fact underlies the identities of all Welsh people (Hopwood, 2019).

The semantics of Welsh-language demonyms and ethnonyms like “Sais” and “Cymro” present a further ideological difficulty in disentangling language from other aspects of identity, such as culture, nationality, and geography. “Cymro go iawn ydy fy ffrind. Dysgwr ydw i.”²⁴, states one Twitter user. The intention is probably not to imply anything about the speaker’s sense of Welshness, and yet it is impossible to exclude a reading which says that the friend is a “true Welshman” in addition to being a “true Cymrophone”.

Many primary schools in Wales have adopted the practice of awarding a weekly certificate to the child who is judged to have been most successful in using or promoting the Welsh language. The winner of the certificate is termed “Cymro yr wythnos” or “Cymraes yr wythnos”. In at least one school, this title is translated into English as “Welsh Speaker of the Week” (Pontybrenin, n.d.). Another school explains on its website that “These certificates have been awarded for keeping the Welsh language alive” (Penalltau, n.d.). A tweet from the official account of a third school, @YsgolPanteg, congratulates the child who has been chosen as Cymraes yr Wythnos “am ei chywirdeb wrth siarad ag esiampl maent yn dangos! Arbennig!”²⁵. Children who attend Welsh-medium schools are, therefore, often educated in environments in which to be a Cymro or a Cymraes is to support the use of Welsh, and to follow standards of correctness in using the language.

The restricted use of “Cymro” and “Cymraes” to denote Cymrophony is by no

²³ *problematic on several levels*

²⁴ *My friend is a genuine Cymro. I’m a learner.*

²⁵ *for her correctness in speaking and the example they set [sic]! Excellent!.*

means incorrect or unprecedented, yet it is by now apparently much less common than the use of these terms to refer to national or civic identity. In both this study's Twitter corpus and in the Senedd's official Record of Proceedings, the use of these terms with the meaning "Cymrophone" is quite rare, suggesting that this sense is no longer prominent in colloquial Welsh. Its use for weekly language awards in schools is therefore marked, and encourages pupils to enregister a strong connection between language and Welsh identity.

It is not only in Welsh that language and ethno-national identity suffer from a certain terminological ambiguity. Although Anglophones may use the terms "English-medium" and "Welsh-medium" to describe the linguistic status of schools in Wales, for example, it is not uncommon to hear them called "English" and "Welsh" schools instead. Selleck (2013) quotes a schoolchild at a predominantly English-medium school describing how she and her best friend "got separated cos their mum wanted them Welsh and my mum wanted me English" (Selleck, 2013, p. 36). As Selleck notes, the choice of medium of education "appears to take on a nationally identifying element – go to the 'Welsh school' and become Welsh, and go to the 'English school' and become English." (Selleck, 2013, p. 36). G. Lewis (2006), too, describes her experience at a Welsh-medium junior school, which "was next door to the 'normal' school, and there was frequent warfare between gangs of 'Englies' and 'Welshies' (even though we were all from Cardiff)". Figure 3.1 shows a plaque over the entrance to Bethel English Baptist Chapel in the town of Maesteg, in the valleys of southern Wales (Coflein, 2010). The designation "English" refers to the language in which services were held, rather than to any cultural or national allegiance to England.

For both Anglophones and Cymrophones in Wales, therefore, boundaries between terms relating to language, nationality, ethnicity, and geography are not always clearly defined. Despite the existence of standardized or officially-preferred terminologies which allow language to be separated from other aspects of identity, in practice (and particularly in colloquial use) these meanings are interwoven and enmeshed in polysemous, ambiguous terms such as "Welsh" and "English", "Cymry" and "Saeson", "Cymraeg" and "Saesneg".



Figure 3.1: Plaque above the doorway of Bethel Baptist Chapel in Maesteg.

3.iv “You’ll never be a member of our exclusive club, you second-class, linguistic failure”: purism, prestige, and power in the Welsh sociolinguistic context

Having sketched out the broad outlines of Welshness as a cultural identity, and explored the interwoven terminology of Welshness and Cymrophony, I will now examine prescriptivist attitudes in detail. Here we will find parallels between Welshness at the socio-political level and Welshness at the level of language use. Connotation, in the systemic-semiotic model I have proposed, is to a large extent the linguistic expression of various kinds of context. If the semiotic categorization of ELUs as “Welsh” words or “English” words is a matter of connotation, it follows that the context in which the Welsh language operates is an integral part of what is expressed by any individual word.

3.iv.a Register and remonstrance: linguistic prescriptivism and language confidence

It is common for people who were educated in Welsh to recall being scolded or punished for code-switching between Welsh and English at school. As a Twitter user puts it “Oh my God it’s only just dawned on me o ni arfer cal ein cosbi am seared [sic] Saesneg”²⁶. Prescriptivist attitudes towards language mixing are not left behind once a Cymrophone reaches adulthood, moreover, but persist in both traditional and social media. The ‘plismyn iaith’ or ‘heddlu iaith’ (*language police*), as they have come to be known in popular discourse, are Cymrophones who publicly criticize linguistic behaviours of which they disapprove.

King (2003) describes the high-prestige variety of written Welsh as an artificial standard, which is nobody’s native variety of the language, and which is kept deliberately distinct from colloquial Welsh. Critique of “the problematic nature of the dominant mode of formal literary Welsh” has been an important thematic strand in modern Cymrophone writing, with authors highlighting and subverting the often

²⁶ Oh my God it’s only just dawned on me *that we used to be punished for speaking English*.

stilted uniformity of the standardized written language (L. G. Lewis, 2019, p. 673). Literary experimentalism notwithstanding, it is in general rather harder for a Welsh-English bilingual to acquire standardized Welsh than to acquire standardized English. The latter can be heard and read across a variety of domains, whereas the former is largely restricted to formal, written registers, acquired through instruction or extensive study. If enregisterment is the extrapolation of contextual meaning from instances of language in use, as Agha (2009) argues, acquiring register distinctions in English is likely to be an easier task for Cymrophone children than acquiring them in Welsh. The expanded use of English in a wide variety of contexts means that English features which mark language as formal, or scientific, or colloquial, have a better chance of becoming firmly enregistered for Cymrophones.

The strongly diglossic nature of Welsh, with a significant gulf between colloquial and literary forms of the language (Borsley et al., 2007), contributes to a sense of linguistic inferiority for Cymrophones who have not fully acquired the literary register. The pseudonymous Twitter account @BeiroCoch (“Red Biro”), which was active on Twitter in the mid-2010s, correcting the language of official organizations, journalists, and other public figures (“@BeiroCoch”, 2015), castigated the linguistic standards of Cymrophones, lamenting “eu bod nhw’n mynnu siarad ac ysgrifennu llediaith sy’n mynd yn debycach bob dydd i’r Saesneg (pan nad ydyn nhw’n defnyddio’r Saesneg, hynny yw – a Saesneg da iawn gan amlaf)”²⁷ (“@BeiroCoch”, 2015). As a consequence of discrepancies in the acquisition of high-prestige written varieties of the two languages, those whose Welsh fails to meet @BeiroCoch’s high standards are nonetheless described as writing “good” English.

Disparagement of colloquial Welsh by a perceived cultural elite, including an insistence on the avoidance of English loanwords, contributes to some Cymrophones’ sense of estrangement from their language, with many choosing to abandon it in favour of English rather than attempt to use an artificial register (C. F. Huws, 2011; M. Johnes, 2012). This compounds the dominance of English as the language of popular culture, the internet, and broadcast media, as well as its dominance as the community language in many parts of Wales. The result is that opportunities to enregister a full range of Welsh varieties are extremely restricted for many Cymrophones. Rather than gaining confidence in using their natural, code-switched

²⁷ *they insist on speaking and writing slang that becomes more like English every day (when they’re not using English, that is – and very good English, for the most part)*

variety of Welsh, and potentially being exposed to a broader range of Welsh as a result of participating in the Cymrophone community, many Cymrophones essentially behave as though they are monolingual Anglophones, believing that their Welsh is not Welsh enough to earn them a place within that community (S. Huws, 2021).

It is unclear whether the problem of purist language correction is particularly prevalent amongst Cymrophones, or whether Cymrophones have a disproportionate fear of being corrected. A Twitter user suggests that the problem is “Pryder am blismyn iaith, yn hytrach na phlismyn iaith felly? Ond pe na bai plismyn, ni fyddai pryder...”²⁸. Although @BeiroCoch, mentioned above, insists that he only criticizes the language of public figures (“@BeiroCoch”, 2015), for example, his criticism is still visible to other Cymrophones and contributes to a potential chilling effect on willingness to use Welsh.

The argument that prescriptivist censure means that Cymrophones will feel more comfortable using English in public leads another Twitter user to assert:

Yr heddlu iaith sydd yn lladd yr iaith. Dwi’n gweld hyder pobol yn dyddiol yn mynd lawr wrth derbyn sylwadau ddi-angen am cam dreiglo neu rhoi gair saesneg mewn brawddeg. So blydi what!²⁹

This condemnation of linguistic purism in informal contexts has been echoed in official spaces, as when Eluned Morgan, MS, told the Senedd in 2020 that “mae’n bwysig ein bod ni’n trio ymdawelu’r plismyn iaith”³⁰ (*Record of Proceedings*, 2020). Enforcing the standards of a puristic variety of Welsh is seen by many Cymrophones as actively harmful to the goal of encouraging Welsh as a natural, everyday language. The centrality of this goal in Welsh politics makes the language police a topic of serious political discussion, rather than an easily dismissed part of the social-media landscape.

The popular maxim “Gwell Cymraeg slac na Saesneg slic”³¹, attributed to poet Ifor ap Glyn (I. W. Rees, 2015), urges Cymrophones to prefer Welsh rather than

²⁸ *Anxiety about the language police, rather than the language police then? But if there were no policemen, there would be no anxiety...*

²⁹ *The language police are killing the language. I see people’s confidence decreasing daily from receiving unnecessary comments about mistakes with treigladd or using an English word in a sentence. So bloody what!*

³⁰ *it’s important that we should silence the language police*

³¹ *Better slack Welsh than slick English*

falling back on English, even if they have access to a more standard register of English than of Welsh. This phrase neatly encapsulates a recognition that, unlike monolinguals, bilinguals may choose to drop one language and rely on the other, particularly when the latter is a widely-used language and the former is not. This fact gives language policing greater power to do harm amongst Cymrophones than it does in Anglophone discourse, as noted by a Twitter user: “Does dim modd i’r rhan fwyaf o Saeson ddihengyd i iaith arall, felly dyw’r Saesneg ddim yn colli defnyddwyr oherwydd y plismyn [iaith]”³².

The issue of prescriptivism has, in recent years, been taken up in *Golwg*, a popular Cymrophone current-affairs and popular-culture magazine and website, often in response to social media comments criticizing language standards (Ceiro, 2021; Dwyfor, 2021; S. Huws, 2021; Mwyn, 2023; Onllwyn, 2022; *Golwg360*, 2021). When journalist Rob Harries tweeted about the football commentary on S4C, complaining that “Literally half the words [commentator and former footballer] John Hartson uses on S4C are English”, and asking “Can’t they get a Welsh speaker?”, an editorial in *Golwg* reported the strong and overwhelmingly critical response on social media (*Golwg360*, 2021). Responses noted that Hartson’s footballing talent and experience were of more importance to his ability as a pundit than his language skills, and pointed out that his background had not given him the best opportunities to develop and use his Welsh: Hartson was raised in a non-Cymrophone family, and had trained and worked in England since his teens (*Golwg360*, 2021).

Columnist Garmon Ceiro expands on these points, noting that “ma’ beirniadu iaith rhywun yn cymryd yn ganiataol ein bod ni i gyd wedi cael yr un cyfleoedd yn yr iaith—ac yn hollol amlwg nid yw hynny’n wir yng Nghymru”³³. Moreover, Hartson’s commentary style reflects the language Ceiro himself uses when discussing football with friends “yn ein Wenglish naturiol”³⁴ (Ceiro, 2021). Here, Ceiro acknowledges the natural tendency of bilinguals to code-switch, particularly when subject-specific terminology and terms of art are called for. *Golwg* columnist Gwilym Dwyfor expresses pleasant surprise that the public response to Harries’ tweet was so strongly critical, suggesting that “mae’n teimlo fel pe bai’r farn gyhoeddus yn troi’n

³² *There’s no way for most of the English to escape to another language, so English doesn’t lose users because of the [language] police*

³³ *judging someone’s language takes for granted that we’ve all had the same linguistic opportunities—and it’s totally obvious that that’s not true in Wales*

³⁴ *in our natural Wenglish*

ddiweddar a bod y plismyn iaith yn canfod eu hunain mewn lleiafrif”³⁵ (Dwyfor, 2021).

Nonetheless, public rows about language standards persist. In 2022, *Capten* by Meinir Pierce Jones won the National Eisteddfod’s prestigious Daniel Owen Memorial Prize, despite one of the three judges objecting to the use of English in some parts of the novel (BBC Cymru Fyw, 2022). *Golwg* columnist Huw Onllwyn challenged this criticism of *Capten*’s mixing of languages with an article which uses code-switching extensively and deliberately to make its point (Onllwyn, 2022). Like those who responded to criticism of John Hartson, Onllwyn asks what message puristic attitudes send to Cymrophones “nad sy’n byw mewn *linguistic-nirvana* Cymraeg”³⁶ and who have restricted opportunities to practice and hone their language skills:

Onid y neges fydd y rheiny’n ei chlywed wrth wrando ar y puryddion ieithyddol yw ‘*Your Welsh isn’t good enough—so don’t bother*’. A bod yna risg iddynt, wrth geisio ei siarad, i dderbyn beirniadaeth hallt gan y rheiny sy’n feistri ieithyddol³⁷.

‘*You’ll never be a member of our exclusive club, you second-class, linguistic failure*’.

Onllwyn’s skilful and humorous use of code-switching in the article is itself a challenge to those who object to the use of English in *Capten*. Manon Steffan Ros, one of the judges to award the prize to *Capten*, praised the author for “defnyddio’r Saesneg fel symbol o ddiethrwch a’r chwithdod rhwng fydoedd y ddwy iaith”³⁸ (BBC Cymru Fyw, 2022). Mwyn (2023) also suggests that code-switching and other features denigrated by purists can be an integral part of modern Welsh artistic creativity. He praises the musician David R. Edwards, frontman of pioneering band Datblygu (Gallagher, 2021), as “[m]eistr ar eiriau a bratiaith fwriadol greadigol”³⁹

³⁵ *it feels as though public opinion is changing of late and that the language police are finding themselves in the minority*

³⁶ *who don’t live in a Cymrophone linguistic-nirvana*

³⁷ *isn’t the message those people will hear when they listen to the linguistic purists ‘Your Welsh isn’t good enough—so don’t bother’. And that they risk, by trying to speak Welsh, receiving stringent criticism from those who are linguistic maestros. You’ll never be a member of our exclusive club, you second-class, linguistic failure.*

³⁸ *using English as a symbol of alienation and the dislocation between the worlds of the two languages*

³⁹ *a master of words and of deliberately creative bratiaith*

(Mwyn, 2023).

As Garmon Ceiro notes, “cyd-destyn sy’n bwysig”⁴⁰: John Hartson’s code-switching style might not be appropriate if he were the lead presenter of Welsh-language current-affairs programme, but “pan ddaw hi at y pyndits, mae gen i fwy o ddiddordeb yn eu profiad a’u dealltwriaeth o’r gêm na’u gallu ieithyddol”⁴¹ (Ceiro, 2021). Although he condemns the exclusionary pedantry of language purists, Onllwyn also acknowledges that the use of standard spoken Welsh is appropriate in many contexts, such as by newsreaders or teachers in Welsh-medium schools (Onllwyn, 2022). What his essay calls for, in essence, is better recognition of register variation in Welsh, including acknowledgement that code-switching is an inherent part of some registers of Welsh.

3.iv.b Hierarchy and heritage: Cymrophone notions of cultural prestige

Like the artificiality of standardized written Welsh, recognized by King (2003), the Welsh promoted by Cymrophone purists and prescriptivists is artificial in a number of ways, particularly in its condemnation of natural behaviours such as code-switching and lexical borrowing, and in its insistence on the total separation of languages. Despite the reality that Cymrophones have, for generations, lived in intimate contact with the English language, authors of literary works are expected to write as though Welsh communities are purely monolingual. Even skilful juxtaposition of the two language for rhetorical effect, as in Meinir Pierce Jones’ *Capten*, is unacceptable to some.

The use of literary Welsh on official documents after the passage of the Welsh Language Act of 1967 caused significant problems for Cymrophones for whom the language was enregistered as ecclesiastical; they were used to hearing it in church, but had little experience or confidence in using it productively as an everyday means of communication (B. P. Jones, 1988). For respondents in Madoc-Jones and O. Parry’s (2012) study, too, concern about their ability to participate in an interview in Welsh generally centred around the expected formality of the encounter. Ability to use the high-prestige variety of Welsh is tied to a very specific kind of cultural education,

⁴⁰ *it’s context that matters*

⁴¹ *when it comes to the pundits, I’m more interested in their experience and their understanding of the game than their linguistic ability*

including the formal traditions of Welsh literature, rather than to educatedness in the broader sense. Whereas English conceptions of literacy, education, and high culture are fairly strongly correlated with the construct of social class, the Welsh situation is rather more complex.

The roots of the Welsh bardic tradition are in the courts of the medieval Welsh rulers and the households of the nobility, who patronized professional poets. These poets worked in the strict metrical traditions of Welsh verse, *canu caeth* (*strict-meter composition*), which required extensive knowledge of the intricate and complex rules of *cynghanedd* (Constantine, 2017). The Anglicization of the Welsh upper classes meant the collapse of this tradition of patronage and professional bardcraft (Constantine, 2017), yet the tradition of *canu caeth* persisted in less formal locations, and amongst the working people. The “*beirdd gwlad*” (*country poets*) of the farming communities of north-eastern Wales “would gather regularly in taverns from the early eighteenth century onward to practise their craft” E. W. James (2019, p. 308), a tradition which inspired the revival of the *eisteddfod* in Wales (Constantine, 2017).

Although the Victorian *eisteddfod* relied on the patronage of gentry and English aristocracy, it nonetheless offered opportunities for working-class Cymrophone writers (E. Edwards, 2019; Lynch, 2007). In the agricultural and industrial North American settler communities of the 1800s, labourers in steel mills might win the bardic chairs of the local *eisteddfodau* (Tyler, 2018). Furthermore, it was the itinerant stonemason Iolo Morganwg, who was the central driver in shaping the modern *eisteddfod* from its roots in a fairly obscure medieval tradition (E. Edwards, 2019). Although, as Constantine (2017) notes, the practice of *canu caeth* may have attracted comparatively fewer working-class poets than the less complex *canu rhydd* (*free-meter composition*), the formal traditions were nonetheless quite available to poets from a range of backgrounds; in some cases, they even appear to have been passed down orally by non-literate poets Constantine (2017).

English working-class poets of the eighteenth and nineteenth century were in many ways quite similar to their Welsh counterparts, developing a self-taught knowledge of poetic tradition and working within supportive social networks of fellow-writers (Goodridge & Keegan, 2004). Yet mainstream English literary thought preferred to position such writers as anomalies. Rather than writers with skill developed through a combination of study, practice, and engagement with community traditions, they were mythologized as solitary figures filled with an unrefined poetic genius given by God (Fulford, 2024; Goodridge & Keegan, 2004; Van-Hagen, 2017). The idea of a poet as a craftsman, working with skills passed

down through the community, sat uncomfortably with the Romantic ideal of the solitary poetic genius.

In the later eighteenth century, as publishing, bookselling, and authorship were increasingly professionalized, it became profitable to market English poetry as the epitome of refined literary taste (Benedict, 2004). The working-class poets of England were repackaged as artless geniuses by professional publishers for the consumption of middle-class readers (Fulford, 2024). Unlike popular genres, particularly the novel, good poetry was imagined as complex and abstruse, and “increasingly became matter for the leisured elite” (Benedict, 2004, p. 17). In contrast, Welsh high poetic culture in the eighteenth and nineteenth centuries was “surprisingly down-to-earth” treating poetry as “not so much a mystery as a craft” Constantine (2017, p. 108).

From the Victorian period onwards, the canon of Anglophone poetry has been shaped and maintained by publishing houses, university scholars, and literary critics (Crawford, 2001; Harding, 2008), and “enshrined in the elite institutions of England and the United States” (Harding, 2008). Despite recent arguments for admitting popular genres such as rap, popular songwriting (notably that of Nobel poetry laureate Bob Dylan), and slam or performance poetry into the poetic canon (Camangian, 2008; Hill et al., 2008; Kelly, 2013; Ricks, 2011), canonical poetry is, on the whole, not part of Anglophone popular culture (Stapley, 2008).

In the Welsh tradition, by contrast, the canonical poetic forms have a prominent place in popular culture. At the turn of the second millennium, J. Rowlands (1999, p. 89) wrote that “[a]t a time when poetry in other countries has been relegated to the academic margins, here in Wales versifying is still the delight of the many”. *Talwrn y Beirdd* (*Y Talwrn*, n.d.), a contest of poetic skill in which teams of bards compete to produce verses on topical themes using a variety of traditional metrical forms, has been broadcast regularly on BBCRadioCymru since 1979 (BBC Cymru Fyw, 2020). These contests have a level of popular appeal (Gravell, 2005; Lynch, 2007; Stapley, 2008) which is far beyond that enjoyed by traditional Anglophone poetry in Britain (Lynch, 2007; J. Rowlands, 1999). Like the poetical gatherings of the eighteenth-century *beirdd gwlad*, these modern contests of poetic skill also regularly take place in pubs and other relaxed venues (H. D. Owen, 2015; S. P. Rowlands, 2001).

Just as British Anglophone and Cymrophone constructs of canonical literature place different weight on cultural notions such as class, heritage, and populism, so too do constructs of linguistic prestige. Unlike British Anglophone purism, which privileges the language of those with higher social class and education, Cymrophone

purism is more interested in the preservation of traditional, often dialectal, forms of Welsh (A. James, 2018; A. R. Thomas, 1987). Welsh dialect lexis is idealized by many both as more purely Welsh than borrowed English-origin ELUs and, equally, as more authentic than the neologistic standard Welsh lexis taught in Welsh-medium schooling (A. R. Thomas, 1987). Access to a prestige variety of Welsh is, therefore, a matter of social identity rather than one of social class, in that it is associated with factors such as growing up in a Cymrophone household, living in a largely Cymrophone community, engaging with a Cymrophone social circle, and consuming Cymrophone media, literature, and music.

Purism in Cymrophone discourse foregrounds heritage and tradition, weaving the notion of what is good Welsh together with issues of family, community, and cultural preservation. The Cymrophony of the *Fro Gymraeg* and its idealized association with Welsh agrarian traditions are part of a complex construction of authentic Welshness, while the industrialization and Anglicization of the southern Welsh valleys both appear as rejections of true Welsh heritage.

The intermeshing of Welshness and Cymrophony, and the privileging of linguistic tradition, explains why, for some Cymrophones, using Welsh which is “correct” according to purist standards may be a mark of authentic Welshness. Yet, as we have seen, there is a growing backlash to such prescriptivist attitudes, which not only criticizes linguistic pedantry, but also rejects the implication that Welshness requires a particular type of Cymrophony. As one Twitter user asks: “Can you only be a patriot of [sic] you speak Welsh fluently then? Do I not love my country because I haven’t properly learnt mutations?”

Members of the Senedd affirm a commitment to making Welsh accessible to everyone in Wales. In doing so, they place the language at the centre of Welsh identity, as a touchstone of Welsh uniqueness and cultural heritage. The civic character of modern Welsh nationalism rejects the exclusionary, ethnolinguistic rhetoric which makes “Welsh” synonymous with “Cymrophone”; instead, it frames the language in legal and civic terms as a matter of cultural heritage, of community cohesion, and of economic and professional advantage.

3.v Implications: “Bratiaith Guilt” and language ideologies in the development of Welsh

“Fi’n ddiolchgar iawn am lot o bethe gesi i yn ystod fy mhientyndod, Bratiaith Guilt isn’t one of em”⁴², asserts a Twitter user. Although the Welsh gwerin have long been prominent archetypes of Cymrophony in the popular imagination, this romanticized image was never universal. For Emrys ap Iwan, writing in 1889, a decline in Cymrophone linguistic standards began “pan ymwithodd gofaint a chryddion, saeri maen, saeri coed, a saeri clobos Ymneillduol i gadeiriau hen Gymreigwyr trylen Eglwys Loegr ac Eglwys Rhufain”⁴³ (ap Iwan, 1889, p. 116). Himself a Methodist minister, ap Iwan nonetheless saw the swift spread of religious Nonconformism in Wales as having necessitated the public advancement of working men who lacked erudition and formal learning. Rather than being the saviours of the language, the gwerin were contributing to its deterioration.

ap Iwan’s claim that “y buasai yn well gennyf, o’r ddau , weled fy nghydwladwyr yn arfer Saesneg gwyach na Chymraeg gwael”⁴⁴ (ap Iwan, 1889, p. 116) finds an echo in Owain Owain’s assertion that Wales without the Fro Gymraeg is not worth saving. For both men, the object to be preserved must be the idealized, authentic one; saving something diluted and debased is worse than useless. A different echo, in the popular modern phrase, “Gwell Cymraeg slac na Saesneg slic”⁴⁵ challenges ap Iwan’s words. Nonetheless, the previous chapter showed that an oddly confused and confusing monolingualist prescriptivism remains in Cymrophone pedagogy, and the current chapter has shown that fear of criticism from purists remains high amongst contemporary Cymrophones.

The reification of a literary, English-free Welsh as a cultural ideal has the effect of hierarchizing the dialects of the language on a cline of authenticity. The less they

⁴² I’m really grateful for a lot of things I got in my childhood, Bratiaith Guilt isn’t one of em.

⁴³ *when Nonconformist blacksmiths and cobblers, stonecrafters, woodcrafters, and clodcrafters*, thrust themselves into the chairs of the old, learned Welshman of the Anglican and Roman Churches*

* A “saer” is an artisan or skilled worker, so “saeri maen” and “saeri coed” would usually be translated as *stonemasons* and *carpenters*. However, “saeri clobos” is not a common idiom. “Clobos” are sods or clods, so the ultimate meaning is probably “farmworkers”. However, maintaining the rhetorical flourish of the repeated noun-phrase structure is important in order to convey ap Iwan’s satirical style.

⁴⁴ *of the two, I would prefer to see my compatriots employ polished English than poor Welsh*

⁴⁵ *Better slack Welsh than slick English*

resemble this ideal, and the more they are influenced by English, the less they are enregistered as authentically Welsh. At the far end of this cline are Welsh non-Cymrophones, whom an extremist minority of Cymrophones refuse to accept are Welsh at all (E. Williams, 2009). This model of a purified Welsh language as a prerequisite of Welsh identity, and the close connection of the prestige language with sites of cultural authenticity such as the Eisteddfod and the poetic tradition, goes some way to explaining why Cymrophones who speak English-influenced dialects feel a sense of inferiority about their Welsh (B. P. Jones, 1988; B. Thomas, 2000). “Bratiaith Guilt” is the inevitable result of the kinds of attitudes explored in this chapter, as bilingual Cymrophones are held to an artificial monolingual standard, and the Anglophone component of their identity is treated as a source of shame.

Denigrating a person’s use of language is not merely a matter of mechanical error-correction; it risks threatening their sense of individual and group identity (J. Edwards, 2010). It may also have the paradoxical effect of making the stigmatized behaviour more appealing. Cymrophone children who are punished for using English or English-origin ELUs may well find speaking English more attractive as a result: using Welsh “becomes a symbol of ideological acquiescence and, conversely, the use of English becomes a way of contesting the authority of the school” (Selleck, 2013, p. 33).

Research on the language of Cymrophones from non-Cymrophone families, who acquire the language through immersion schooling, suggests that they acquire a variety that is not as formal as literary Welsh, but also lacks many of the colloquial features of traditional Welsh dialects (Durham & Morris, 2016) which, as we saw earlier in this chapter, are often idealized as the most authentic varieties of the language. Those who acquire Welsh through immersion education are characterized by insecurity about their linguistic ability, and by the tendency to use English as the near-exclusive language of peer interactions outside the classroom (Morris, 2014). The phenomenon of those who learned a minoritized language through immersion schooling or adult education, rather than family transmission, doubting their claim to “speaker” status has also been documented for other languages, such as Basque (Ortega et al., 2015) and Canadian French (Juan-Garau & Lyster, 2019).

Equally, Cymrophones whose formal education has been wholly or partially English-medium are often reticent about the standard of their Welsh, particularly if they have low literacy in the language (J. Davies, 2014). Here, too, there are parallels with other minoritized-language communities, in which speakers whose route to the language was familial transmission alone may see themselves as speaking an

illegitimate sort of patois in comparison with the expanded linguistic repertoire of the formally-taught (Costa, 2015; Frekko, 2009). “Bratiaith Guilt” is therefore not only a phenomenon of those from non-Cymrophone families; it may also characterize people who do not have the opportunity of education in their mother tongue. In both cases, Cymrophones’ lack of confidence reflects the narrowness of their linguistic opportunities, and the correspondingly narrow repertoires they enregister as a result.

R. W. Jones (2004) has argued that the scarcity of Welsh-medium provision in university teaching is an influential contributor to the use of *bratiaith*. Despite the purist tone of his remarks, Jones raises the important point that Cymrophones who pursue English-medium higher education are likely to develop a formal, academic register of English which they will be unable to replicate in Welsh, leading to an understandable loss of confidence in their comparatively restricted Cymrophone repertoire. (In this, they are the mirror-image of Cymrophones who report feeling a sense of linguistic inferiority when transitioning to English-medium university study, since their prior (Welsh-medium) education led them to develop an academic register of Welsh which they could not easily replicate in English (R. Jones & Desforges, 2003).) This example raises the difficulty of challenging purist discourses in a minority-language context: those who argue against mixing English with Welsh are arguing against a dominant language. They are not wrong in highlighting the relative lack of resources and opportunities for Cymrophones who want to live their lives wholly in Welsh, or the difficulty of providing young Cymrophones with linguistic experience to match the experience they gain simply from existing in an English-dominant society.

Nonetheless, suggesting that there is some kind of authentic Welshness unachievable by Cymrophones who code-switch or use English-origin loanwords is no less harmful for being motivated in part by anxieties about the survival of the language. Unsurprisingly, attacks on the natural, colloquial language of Welsh-English bilinguals have been linked to reluctance to speak Welsh and to the abandonment of the language both by adult learners and by those fluent since childhood. What has been termed the “minority-group reaction” (Lambert et al., 1960), in which speakers of stigmatized varieties display negative evaluations of their own language, is therefore doubly triggered for Cymrophones whose main dialect of Welsh is strongly influenced by English. On the one hand, their language (as a stigmatized variety of Welsh) is enregistered within Cymrophone discourse as uneducated, lazy, semi-lingual, and not truly Welsh. On the other hand, their

language (as a minoritized language) is enregistered within Anglophone discourse as fossilized, old-fashioned, and exclusionary. As the next chapter will show, these enregistered characteristics are prominently displayed in discourses about Welsh orthography.

Chapter 4

“Rough as Bwlchardudwy”: Orthography and the foreignness of Cymraeg

“Where are you going, my pretty maid?”
“To Llyncochdywarchen, sir,” she said.
“Shall I go with you, my pretty maid?”
“If you climb Tecwynuchelochrau, sir,” she said.
“Is it as rough as its name, my pretty maid?”
“Rough as Bwlchardudwy, sir,” she said.
“I like not your Welsh, my pretty maid.”
“Glyndwr Fawrglod did, sir,” she said.
“I’ll never learn Welsh, my pretty maid.”
“Nobody asked you, sir,” she said.

Anonymous, undated postcard, postmarked 1923.

This chapter’s epigraph is a verse found on an old postcard¹ (fig. 4.1), depicting a young woman in stereotypical Welsh costume being sexually harassed by a man wearing tweed walking clothes. The verse is a parody of a traditional rhyme (Halliwell, 1886, pp. 107–108), in which a man interrogates a pretty woman about her background and prospects, concluding that she does not meet his matrimonial standards². In response to his declaration that he will not marry her, she responds

¹ All postcards depicted in this section are taken from my own collection. They all appear to have been produced in the late-nineteenth or early-twentieth centuries, but it is often impossible to determine the exact date of publication, except as a *terminus ante quem* established by a postmark if the card was ever sent through the post.

² In the version given in Halliwell (1886), she is rejected for refusing to promise faithfulness; in another common version she is rejected for being poor.

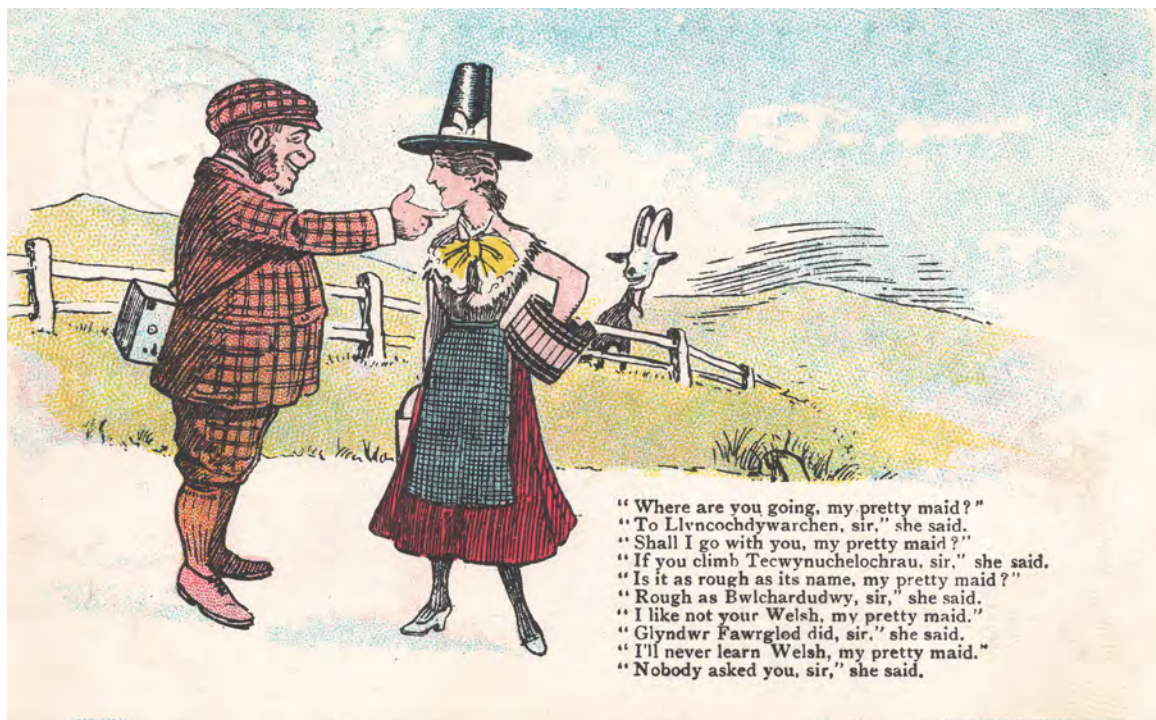


Figure 4.1: Postcard, postmarked 1905.

“Nobody asked you, sir!”. This rebuff is repeated on the postcard, but in response to the man’s assertion that he will not learn Welsh.

This postcard is somewhat unusual in providing a potential source of humour to both Cymrophone and non-Cymrophone readers (Henken, 1996). For a Cymrophone, the butt of the joke is the tourist, with his cultural and linguistic deficiencies (Henken, 1996). The reference to “Glyndwr Fawrglod”, [*Owain*] *Glyndŵr of Great Fame* evokes a stereotype of Welsh national pride which would have been recognized by non-Cymrophones (Henken, 1996), yet the Welsh epithet “Mawrglod”, omission of Glyndŵr’s given name, and use of the Welsh spelling rather than the Anglicized “Glendower” all contribute to making the young woman’s meaning obscure to non-Cymrophones, implying an intended Cymrophone audience. Furthermore, the verse’s final line at least notionally offers the young Welshwoman agency to reject the tourist’s advances and his denigration of her language.

On the other hand, the postcard permits a reading in which the Welsh language is the joke’s target, and the monoglot Anglophone is the intended audience. As we shall see, the use of long Welsh words and placenames with distinctively Cymrophone orthographic sequences, such as the use of <w> and <y> to represent vowels in <Glyndwr>, is a repeated trope in tourist postcards of this era. Indeed, it is

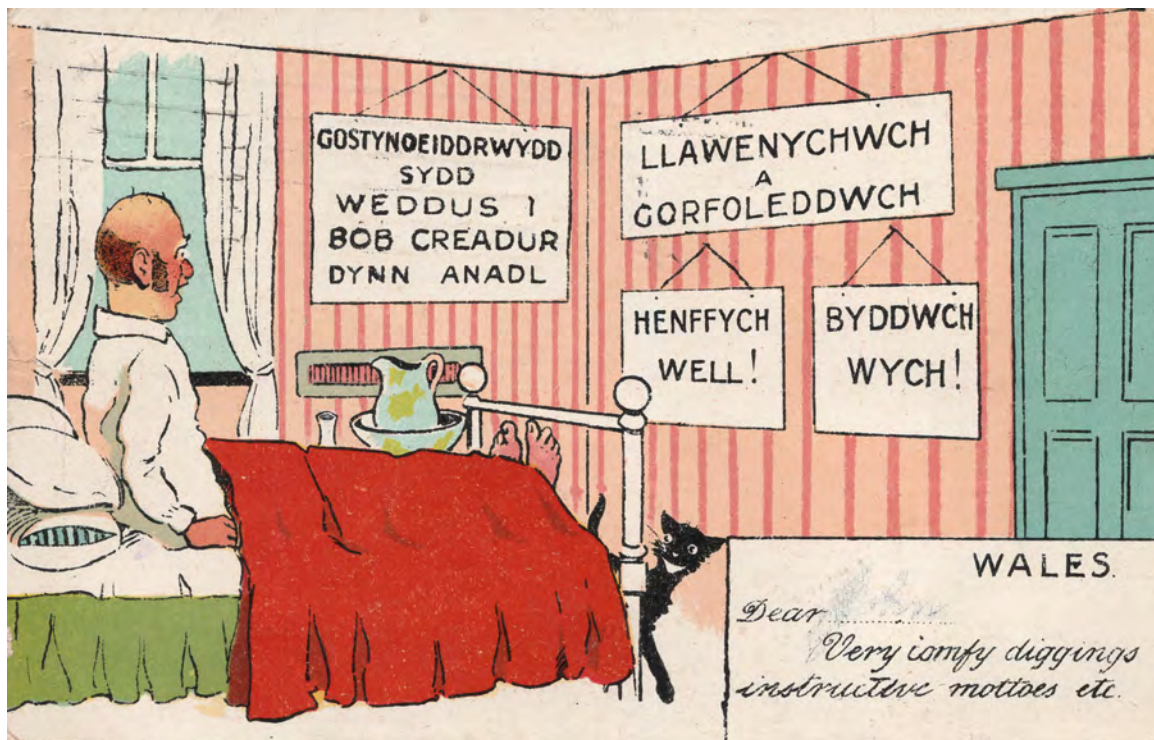


Figure 4.2: Postcard, postmarked ?1926.

more common to find postcards which take an exclusively Anglophone, mocking perspective on the Welsh language, without poking fun the presumptuous tourists who assume they are objects of envy for the simple Welsh peasants. As Pritchard and Morgan (2005) argue, depictions of Wales in postcards from the Edwardian period imply an English, colonialist gaze through which the Welsh and their language are stereotyped and othered. The card in fig. 4.1 replicates this gaze.

Figures 4.2 and 4.3 show two more early Welsh novelty postcards with similarly ambiguous audiences. The signs depicted on the walls of the guest house in the first card all contain real Welsh phrases (apart from the error of transcribing <gostyngeiddrwydd> as “GOSTYNOEIDDRWYDD” in the leftmost sign, an error which implies that the artist was not a Cymrophone, although his or her source for the sayings must have been). The meaning of the signs is roughly as follows, from left to right and top to bottom: “modesty is becoming to all creatures who draw breath”; “rejoice and be glad”; “greetings!”; “be excellent!”. They are therefore appropriate mottoes for a morally-inclined guesthouse owner to display to guests. In fig. 4.3, the tweed-clad tourist asks a Welsh farmer about the state of his crops. The response, while perfectly comprehensible, uses deliberately long words. An English translation which matches this feature of the Welsh text might read “Comparative with the

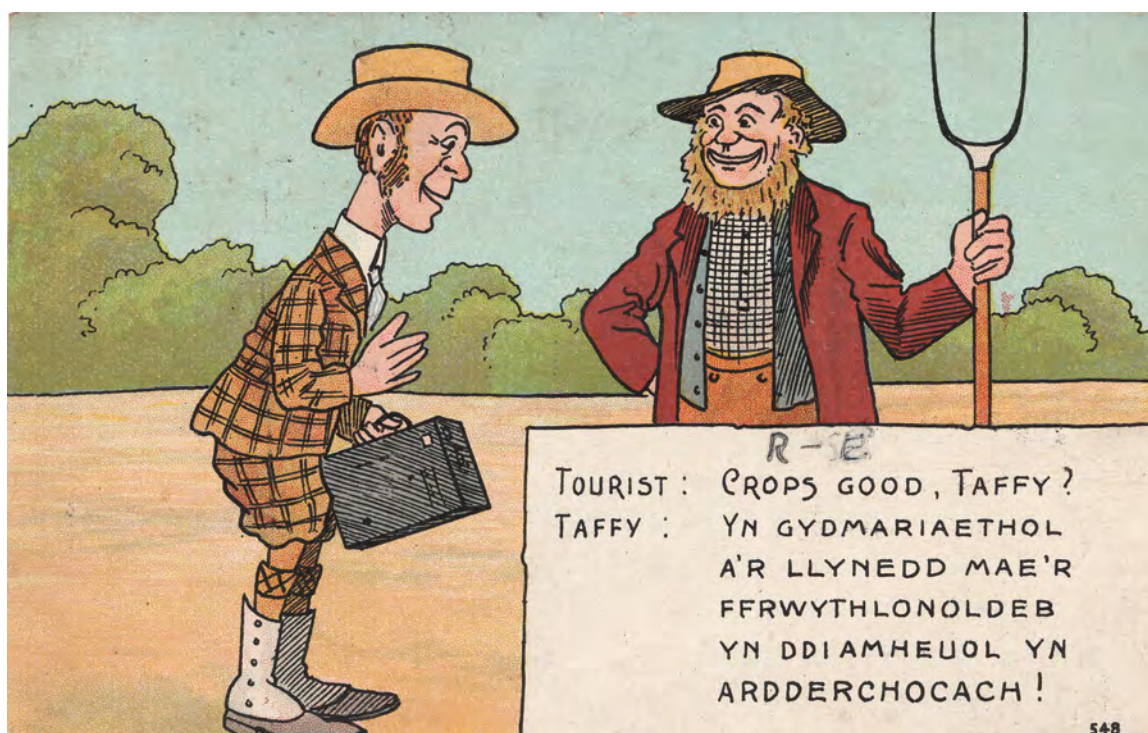


Figure 4.3: Postcard, postmarked 1908.

previous year, the fruitfulness is incontrovertibly splendifer!”

In both cases, a knowledge of Welsh is needed to understand the texts on the cards in full. For a non-Cymrophone, the appropriate content and formal register of the Welsh in both cases is invisible. Only a Cymrophone will appreciate the humour in differentiating the English tourist’s informal language (he uses a minor sentence, eliding the verb, as well as the familiar and derogatory nickname “Taffy”) from the farmer’s flights of literary language. Nonetheless, the texts on both cards are also notable for their inclusion of a few distinct orthographic cues, meant for the Anglophone gaze. These cues are a feature of this genre of postcard, including words made up of a large number of graphemes, words whose vowels are represented by <w> and <y>, and words with the Welsh graphemic units <ch>, <dd>, <ff> and <ll>. The use of these digraphs word-initially, a position they rarely or never occupy in English, highlights them.

The postcard in fig. 4.4 contrasts with those already considered, in engaging the Anglophone audience alone. It features supposedly Welsh place names, two of which are just nonsense words. The other, <cyfansoddwyd>, is a verb form (“was composed”), not a toponym. The nonsense words are a jumble of morphemes: <breuddwydiofwyn> is a compound of two words meaning “sweet dreaming”, and



Figure 4.4: Postcard, no postmark.

<cyflwyneraeg> seems to be made up of the stem of the verb “cyflwyno” (*to present*), and some form of the suffix {-eg}, found in the names of languages. This latter term might tentatively be translated as “presentationese” in Welsh, although it seems unlikely that any actual meaning should be attached to the makeup of these words. Rather, their function is simply to display the foreign nature of Welsh orthography to the Anglophone gaze. Once again, <w>, <y>, and <dd> are prominent here, as is the stereotypical discourse marker “look you”, a longstanding feature of the mockery of Welsh varieties of English (Pritchard & Morgan, 2005). Unlike the others, this card provides no source of interest or humour for a Cymrophone reader, merely an insulting, sexualized image of a Welsh woman, and a text that manages to denigrate both Welsh and Welsh English.

Many more images of vintage postcards are given in Appendix B, in which similar motifs of the Welsh language are repeated. Most of these, like the one shown in fig. 4.4, are primarily intended for Anglophone readers. The dominant theme is the length of Welsh placenames, and in particular of the notorious northern Welsh village Llanfairpwllgwyngyllgogerychwyrndrobwlllantysiliogogoch. Other Welsh placenames in these images, real and fictitious, share some or all of the features already mentioned: orthographic length, and prominence of <w>, <y>, <dd>, <ch>, <ll>, and <ff>. Some cards also highlight the orthographic representation of postalveolar affricates, with names such as “Shani Shones” rather than standard English “Johnny Jones”. There are other examples of mockery of Welsh personal names, and in particular of Welsh patronymics (and their supposed length).

The representation of Welsh people in these cards repeats the Anglophone, tourist gaze. The tourists are implicitly English, usually men, bowler-hatted or clad in country tweeds or plus-fours. Welsh people are either tradespeople, service workers, or the stereotypical rural gwerin, with a nod to the romantic fictions of Welsh tradition constructed by Victorian Cymrophiles (Constantine, 2007; E. Edwards, 2019). The Welsh appear as porters, postmen, and golf caddies; as countryfolk in old-fashioned, worn and patched clothes, sometimes filled with bardic inspiration; or as pretty Welsh maidens wearing (antiquarian-invented) Welsh costume (Freeman, 2007).

These postcards enregister Welsh as archaic, comical, inconvenient, and transgressive; the language is likened to swearing, causes accidents and injuries, and is generally portrayed as an impossibly long-winded medium of communication. The orthographic forms of Welsh play a prominent role in characterizing the language in these ways. Chapter 2 resituated graphemes as potential sites of meaning in a

connotative semiotic model of language, explaining why it is that the presence of <w> and <y> in the words in these postcards can carry not only a cultural meaning such as “Welshness”, but also potentially “humour”, “foreignness”, or “transgression”.

Cymrophone lexicographer Thomas Edwards (Johnston, [n.d.](#)) noted that, “To an Englishman, the Welsh language on paper has a most uncouth appearance, from the number of consonants or double letters, such as *dd, ll, ch, ngh, ng, rh, ff*, &c.”

(T. Edwards, [1847](#), p. 8). Some of Edwards’ contemporaries and near-contemporaries, as we saw in the previous chapter, espoused the view that a Welsh pronunciation is “foreign” to English, reproducing and adopting an Anglophone perspective that enregisters Welsh phonology as other. In contrast, Edwards reproduces the Anglophone view of Welsh orthography as “uncouth” only to reject it, arguing that “in whatever situation the letters are placed in the composition of words, they have but one invariable sound, which simplifies the language much”. As Chapter 2 argued, a language’s phonology and orthography are not simply two parallel ways of expressing the same content, and neither is orthography a mere artificial doubling of the phonology. Rather, orthography and phonology, while intimately related, are also semiotically independent.

In the face of Saussurean dismissal of writing, the model proposed in Chapter 2 rehabilitates orthography as an object of linguistic study, and shows how it might be enregistered with distinct connotational meaning. This understanding of orthographic units as semiotic objects in their own right provides the necessary context for understanding how orthography can, in itself, add to the meaning of a particular linguistic act. After briefly describing the orthographic conventions of Welsh, this chapter will review the literature on the sociocultural role of orthography and on the orthographic adaptation of ELUs, establishing a foundation for understanding the postcards discussed above, as well as for analyzing the orthographically adapted ELUs collected for this study.

4.i Welsh orthography and its relationship to Welsh and English phonology

The scholar and poet John Morris-Jones was the driving force behind the standardization of modern Welsh orthography. He oversaw The Society for the Utilization of the Welsh Language’s *Welsh Orthography* (1893), and *Orgraff yr Iaith Gymraeg* (1928), the official report on spelling reform of the University of Wales

Board of Celtic Studies' language and literature committee (Johnston, n.d.). The publication of these reports effectively ended the “orthography battle” of the nineteenth century, which had seen a number of authors attempt their own reforms of Welsh spelling (Hawke, 2018). *Orgraff yr Iaith Gymraeg* received minor updates in 1928 and 1987; the third edition (Orgraff, 1987) is the current standard for Welsh orthography (Hawke, 2018).

The modern orthography of Welsh is founded on the premise that the written form should be, as far as possible, a direct representation of the sounds of the spoken form (Orgraff, 1987, p. iv). The major departures from this rule are the use of <y> to represent two different phonemes ([ə] and [i]); and the use of two different digraphs, <ph> and <ff> for the phoneme /f/ (the former only word-initially, when aspirate treigladd has altered the initial phoneme /p/ to [f]) (Orgraff, 1987).

Unlike English orthography, therefore, Welsh orthography tends to be a fairly unambiguous guide to pronunciation once its rules have been acquired. The corollary of this fact, of course, is that Welsh orthography is much less flexible than that of English, and less tolerant of deviance. In section 4.iii.a below, I will discuss this consequence of Welsh orthographic transparency in more detail, since it has significant implications for answering this study's main research question.

In discussing Welsh vowels here, I will occasionally refer to “vowel sets”. This term is fully defined in Appendix A.ii. Essentially, it uses the close link between orthography and phonology to provide a convenient way of referring to the pairs of what are usually called long and short Welsh vowels, by labelling each pair with a grapheme:

- <a> /e/~/a:/
- <e> /ɛ/~/e:/
- <i> /ɪ/~/i:/
- <o> /ɔ/~/o:/
- <u> /ʊ/~/u:/
- <w> /ʊ/~/u:/
- <y> /ə/

The only set without two phonemes is <y>. This corresponds to the Welsh **ŷ** vowel, which has no “long” form. It should be emphasized that the grapheme <y> is the most

ambiguous in Welsh, since it can also, in some contexts, represent the sounds of the <u> set, or of the <i> vowel set, for varieties which do not distinguish <i> and <u> (this set will be called the <u>~<i> set, unless referring to the sounds of the <u> set alone). The grapheme <ŷ> in Welsh orthography always represents the long vowel of the <u>~<i> set.

4.i.a Syllables and stress: Diacritics in Welsh orthography

Diacritic, or accent, marks are a distinctive feature of Welsh orthography, and can influence the phonological interpretation of alphabetic characters representing vowels. In analysing the data collected for this study, I will also argue that diacritic marks are a highly enregistered feature of Welsh orthography, so that the use of a circumflex accent can, in some contexts, be sufficient to connote Welsh languageness. Before we proceed to a discussion of alphabetic characters in Welsh, therefore, this section will give an overview of conventions for the use of diacritics in Welsh.

Welsh monosyllables are usually considered to carry a primary stress, except in the case of proclitics such as *yr* (/əɾ/, “the”) and *fy* (/və/, “my”) (*Orgraff*, 1987, p. 1). In monosyllables which are not proclitics vowel length is most often predictable from the phonological makeup of the syllable. Written accent marks are therefore not universally used to mark vowel length or syllable stress, but rather are called into use in predetermined cases. Some of these cases are identified by a general rule and apply to all or most words which are subject to that rule. These accents tend to be generated when a particular phonological environment regularly produces both monosyllables with long vowels and monosyllables with short vowels, so that the vowel length is not predictable from the written forms without the accent mark. Other contexts for the use of diacritics are triggered when, for example, two different phonological forms would generate the same orthographic form, causing ambiguity around which word is intended. The accent disambiguates these particular pairs, but is not applied to the written forms of all words containing the same kind of phonological environment, since these are generally considered predictable.

Four accent marks are habitually applied to the vowel graphemes in Welsh orthography. The most common is <^> the circumflex, which can be described roughly as a marker of vowel length. The other three accents are <¨>, the diaeresis; <´>, the acute accent; and <`>, the grave accent. We shall consider the circumflex first, since it is by far the most commonly used of the diacritics in Welsh. In the first

type of environment described above, in which a distinguishing diacritic is always triggered, all forms with long vowels are written with the circumflex. In these contexts, the rule is followed indiscriminately, regardless of whether or not there is also a word in use with a short vowel.

Monosyllables ending in [l], [n], or [r], and containing vowels other than those of the <u>~<i> set, vary in vowel length. Vowel length is a reflection of etymology: words which end with an etymological geminate consonant have a short vowel, and those with an etymological single consonant a long (Orgraff, 1987). The latter group should therefore always be written with a circumflex on the vowel³.

The circumflex is used to mark long vowels in words such as <cân> (meaning “song” or “cane”), distinguishing it from <can>, which is both an ILU meaning “white”, now a poetic archaism, and a commonly-used ELU from English “can”, in the sense of a container for food. The etymological geminate consonant in the ILU <can> derives from the sequence [nd] in Latin “candeo”, while the single consonant in <cân> derives from [n] in Latin “cano”. The modern spelling <can> is therefore “megis llaw fer am *cann*”⁴ (Orgraff, 1987, p. 26). Since vowel length in such ILUs is the expression of a consistent etymological principle, Orgraff (1987) prescribes the consistent use of the accent for long vowels in these contexts. Accordingly, words such as <pêl> ([pe:l], “ball”) also take a circumflex, even though there is no attested Welsh word *<pel> (*[pɛl]) which might cause ambiguity (Orgraff, 1987, p. 18).

The authors of Orgraff (1987, p. 22) advise that accent marks should also be used as and when needed, in order to “ochel aneglurder damweiniol”⁵. Outside the universal contexts noted above, therefore, the use of the circumflex is often triggered when some exception to a predictable rule results in a pair of words, one with a long vowel and one with a short, which would be ambiguous in writing without the accent mark. This type of exception may arise when a proclitic shares its phonological form with a normal stressed monosyllable, for example. The noun {tŷ} (“house”) can become <dŷ> after soft treigladd. Without the accent, this could be confused with the proclitic {dy} (singular informal “your”). Orgraff (1987, p. 16) notes that all forms of {tŷ} must be written with the accent, for the sake of

³ The only exceptions permitted by Orgraff (1987, p. 18) are the very common words <hen> ([he:n], “old”) and <dyn> ([di:n], “man”), which are written without accents “o gynefindra hen arfer” (from long-held custom)

⁴ as it were a shorthand for *cann*

⁵ avoid accidental lack of clarity

consistency.

Consistency across all forms of a lexeme is not always required when adding accents to ambiguous forms caused by treiglâd, however. Whereas all forms of {tŷ} require an accent, the same is not true for words such as {cwyn}. After soft treiglâd, {cwyn} ([kɔ̃ɪn], “complaint”) has the form <gŵyn>, and after aspirate treiglâd, the form <chŵyn>. Without the accents, these words could be confused with {gwyn} ([gwɪn], “white”) and {chwyn} ([χwɪn], “weeds”), respectively. However, the base form is written <cwyn>, not *<cŵyn>, since there is no word *[kwɪn] for which it could be confused *Orgraff* (1987, p. 46).

In many other cases, the exceptional forms which trigger the use of the circumflex are ELUs. Monosyllables with final [p] or [k] almost always contain a short vowel, for example; the exceptions, which are marked with an accent, include the ELUs {siâp} (“shape”), {tâp} (“tape”), and {strôc} (“stroke”) (*Orgraff*, 1987, p. 17). A further context for the circumflex is found in consonant-final polysyllables whose final syllable contains a stressed long vowel. Since Welsh words are almost always stressed on the penult, adding the accent mark on the written form helps to indicate the correct stress pattern, in addition to marking vowel length (*Orgraff*, 1987, p. 21). ELUs such as {perswâd} ([pɛr'swa:d], “persuasion”, from English persuade) and {apêl} ([əp'e:l], “appeal”) are in this category, alongside potentially ambiguous ILUs such as {ymlâdd} ([əm'la:ð], “to tire”, as distinct from {ymladd}, ['əmlɛð], “to fight”).

The acute and grave accent marks are occasionally used on graphemes representing short vowels. The grave accent usually indicates that the vowel represented is anomalously short, and loanwords again often provide contexts for its use. The English-origin ELU {mŵg} (“mug”), for example, is pronounced [mʊg], as compared to the ILU {mwg} (“smoke”), which is pronounced [mu:g]. Since monophthongs in Welsh monosyllables ending in [g] are almost always long (*Orgraff*, 1987, p. 17), an accent is placed to indicate the anomalous short vowel. The acute accent marks stress rather than vowel length (*Orgraff*, 1987), and is placed on a grapheme in a stressed final syllable representing either a short vowel or the first vowel of a diphthong: {casét} (“cassette”), {casáu} (“to hate”) (J. E. Hughes, 2011).

A diaeresis is used on the first of a pair of vowel graphemes to indicate that it represents a long vowel, rather than the first element of a diphthong or (in the case of <i>) the palatal approximant /j/ (Brake, 1998; *Orgraff*, 1987). The accent mark in {copïo} (“copying”), for example, shows that it is pronounced [kɔp'i:ɔ], and not *['kɔpjɔ] as the form *<copio> would suggest. In {dŵych}, the diaeresis indicates the

two-syllable pronunciation ['di:ʊχ], whereas *<duwch> would suggest the monosyllabic pronunciation *['diuχ], with a diphthong rather than a sequence of two monophthongs.

This overview highlights two important points about the use of diacritics in Welsh. In the first place, the system has strong links to phonological and prosodic features, and is used as a way to reduce orthographic ambiguity, particularly where ELUs threaten to upset the system's transparency. Secondly, however, the system for applying diacritics is not fully regular, and some of the rules apply only in particular contexts or to particular lexemes. To be sure of using diacritics according to the standard prescribed by *Orgraff* (1987), one must learn both the rules and the exceptions to them. One must also have a good knowledge of existing Welsh vocabulary in order to decide whether a new form needs a diacritic to distinguish it from a form already in the lexicon. Most of the accents are used quite rarely, and Cymrophones may not encounter them frequently enough to extrapolate the norms for their use. Although we may expect diacritics to play a role in expressing the phonology of orthographically-adapted ELUs, therefore, it seems likely that their use will not always be governed by the rules set down in *Orgraff* (1987), particularly in less formal contexts.

4.i.b Print and pronunciation: Welsh grapheme-phoneme correspondences

Both English and Welsh use a form of the Roman alphabet, but the grapheme inventories, orthotactic rules, and grapheme-to-phoneme mappings are very different. Welsh has a shallow orthography (Ball & Williams, 2001), a feature which has been shown to help Cymrophone children in learning to read, as compared to their Anglophone peers (N. C. Ellis & Hooper, 2001; Spencer & Hanley, 2003, 2004). These advantages are of particular benefit to poor readers, who trail much less far behind their peers when learning a shallow orthography (Spencer & Hanley, 2004). Anglophone children with strong reading skills seem to catch up to their Cymrophone counterparts by around age 10. However, this is not the case for those with poor reading skills; Cymrophones of this age with dyslexia, for example, manifest slow but accurate reading, whereas the reading of Anglophones with dyslexia is both slow and inaccurate (Spencer & Hanley, 2004).

In comparing groups of younger (age ~7) and older (age ~10) children, Mayer et al. (2007) found that phonological processes are involved in reading for meaning

by Cymrophones in both groups, but are significantly less important to the older group of Anglophones than to the younger. In addition, around 40% of the English-language spelling errors of Welsh-dominant bilinguals aged between ten and eleven years old are attributable to overgeneralization of Welsh orthographic rules (C. James et al., 1993).

Matched groups of Cymrophone and Anglophone children aged between six and seven years of age were tested on their ability to read words aloud and their comprehension of those words by N. C. Ellis and Hooper (2001). The Anglophone children were able to read aloud significantly fewer words than Cymrophone children from comparable lists of frequent words in their respective languages. However, the Anglophones were able to comprehend more of their wordlist than the Cymrophones were of theirs, as measured by asking them to explain the meaning, offer a synonym, or use the word accurately in a sentence. On average, the Cymrophone children comprehended 59%, and the Anglophones 86%, of the words they could pronounce (N. C. Ellis & Hooper, 2001). N. C. Ellis and Hooper's (2001) research supports the supposition that Cymrophones learn to read by relying largely on sublexical information about the shallow Welsh orthography, while Anglophones rely far more on lexical information about words they already know (Spencer & Hanley, 2003).

These findings for the acquisition of Welsh orthography are in line with scholarship showing that readers in a deep orthography tend to make use of reading strategies differently from skilled readers of a shallow orthography (N. C. Ellis & Hooper, 2001), relying less on grapheme-to-phoneme mapping and more on lexical information about known words (Frost et al., 1987) or on recognizing word morphology (Mousikou et al., 2020). Prior work on Cymrophone literacy strategies therefore motivates the supposition that Welsh readers will be more sensitive to the transgressive nature of ELU orthography in Welsh than Anglophones are in English, because Cymrophones are significantly more reliant on the regularity of grapheme-to-phoneme correspondence than Anglophone readers are.

4.i.c Script and strangeness: orthographic conventions for ELUs in Welsh

The extent to which an English-origin ELU used in Welsh requires orthographic adaptation naturally depends on the phonology and orthography of the original English form. If a given English phoneme can be matched to one in the Welsh

phoneme inventory, then the grapheme(s) representing that phoneme in the English word may or may not need adaptation, depending on whether Welsh permits the same grapheme-phoneme mapping. If the English phoneme is not in the Welsh phoneme inventory, then adaptations will draw on conventional or ad-hoc strategies for approximating the English sound. In order to understand how and why ELUs are orthographically adapted in Welsh, therefore, we must first review the phoneme-grapheme mappings made available in Welsh, and the traditional ways of representing English phonemes in Welsh orthography. The following account describes the orthographic practices of modern Welsh and of modern Standard British English, unless otherwise noted.

4.i.c.1 Consonants in Welsh ILUs

Almost all of the consonant phonemes used in Welsh ILUs can be matched to a single graphemic representation (P. W. Thomas, 1996), as follows:

- /b/
- <c> /k/
- <ch> /x/
- <d> /d/
- <dd> /ð/
- <f> /v/
- <g> /g/
- <ng> /ŋ/
- <h> /h/
- <l> /l/
- <ll> /ɭ/
- <m> /m/
- <n> /n/
- <p> /p/
- <r> /r/

- <rh> /r̥/
- <s> /s/
- <t> /t/
- <th> /θ/

The graphemes <ch>, <ll> and <rh> represent phonemes which are not used in English. Welsh /r/ is generally interpreted as a counterpart to English /ɹ/. Otherwise, the Welsh consonants listed above can be related to English phonemes which are represented with the same IPA symbols in broad transcription.

The phoneme /f/ is usually represented by <ff> in modern Welsh ILUs, except where it begins a word (whose base form has initial /p/) that has been subject to aspirate treigladd, in which case /f/ is represented by <ph> (Orgraff, 1987). The semi-vowel /j/ is usually represented by <i> in Welsh, and the semi-vowel /w/ by <w>. With the exception of <n> and <r>, which may be doubled in certain contexts⁶, consonant graphemes are not doubled.

Apart from the exceptions noted above, the relationship between the Welsh consonant phonemes and graphemes is generally one-to-one; each graphemic representation unambiguously represents a single phoneme. The digraphs <ff>, <ll>, and <dd> are considered units of the Welsh alphabet; they are digraphic graphemes rather than pairs of the monographs <f>, <l>, and <d>. The digraphs <ch>, <ng>, <ph>, <rh>, and <th> are also treated as discrete graphemes.

4.i.c.2 Comparing the vowels of Welsh and English

The vowel phonemes, and their realizations, in Prestige Southern-England English (PSEE), Welsh English (WE), and Welsh are outlined in the guide to transcription conventions (Appendix A.ii). However, knowing the phoneme inventories of the two languages is not, on its own, sufficient to explain how orthographic adaptation of English-origin ELUs proceeds in Welsh, either historically or in contemporary colloquial Welsh. Figure 4.5 illustrates the monophthongs in the phonemic inventories of Welsh, WE, and PSEE. Figure 4.6 shows the diphthongs of Welsh and of PSEE. These figures show that PSEE has rather more phonemic distinctions

⁶ See P. W. Thomas (1996), Appendix II.40–51, for a comprehensive discussion of the doubling of these graphemes

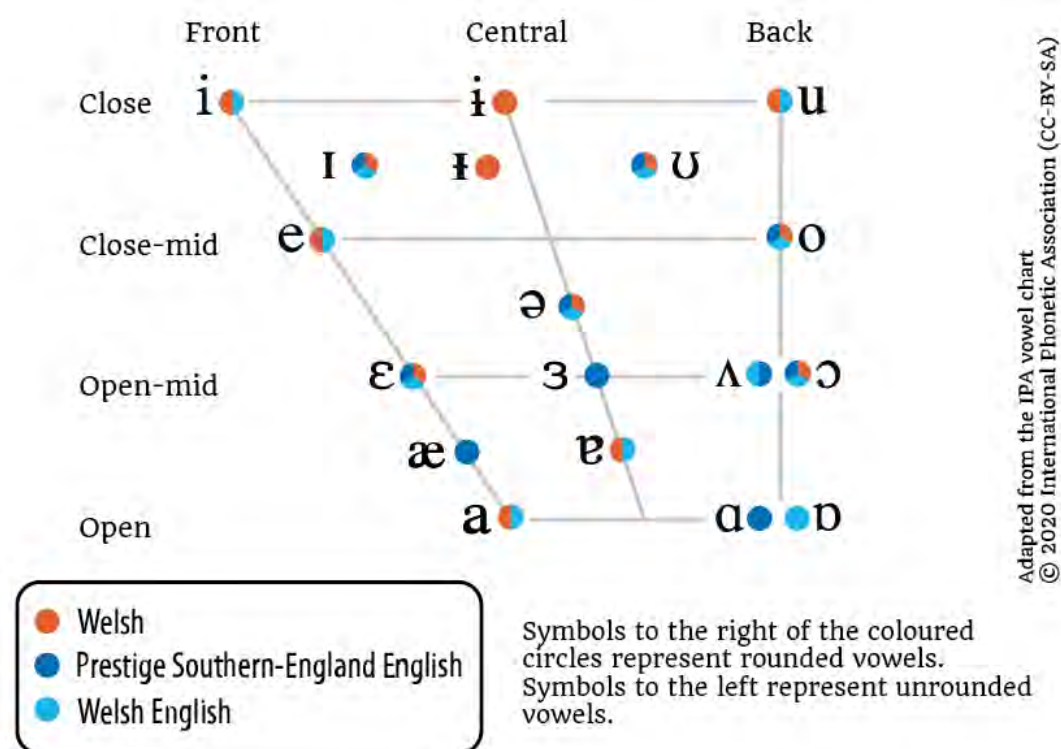


Figure 4.5: The monophthongs of Welsh and English.

amongst its open and open-mid vowels than does Welsh, as well as the diphthongs /eɪ/ and /oʊ/, and a set of diphthongs closing towards schwa (fig. 4.6, number 1), which are missing from Welsh. The transcription of the closing diphthong /oʊ/ reflects gradual changes in the PSEE accent; many older representations of PSEE instead have /əʊ/⁷ (e.g. Connolly, 1981), a diphthong which *is* present in the inventory of Welsh.

Northern Welsh has a number of diphthongs closing towards /ɨ/ (fig. 4.6, number 2); however, since /ɨ/ is not present in PSEE, and has merged with /i/ and /ɪ/ in southern varieties of Welsh, this discussion will mainly disregard /ɨ/, /ɪ/, and their related diphthongs, and assume that /i/ and /ɪ/ are the logical point of comparison for PSEE /ɨ/ and /ɪ/.

⁷ Indeed, Lindsey (2019) argues that this sound still begins with /ə/ in modern SSBE, with an endpoint at /w/ rather than /ʊ/.

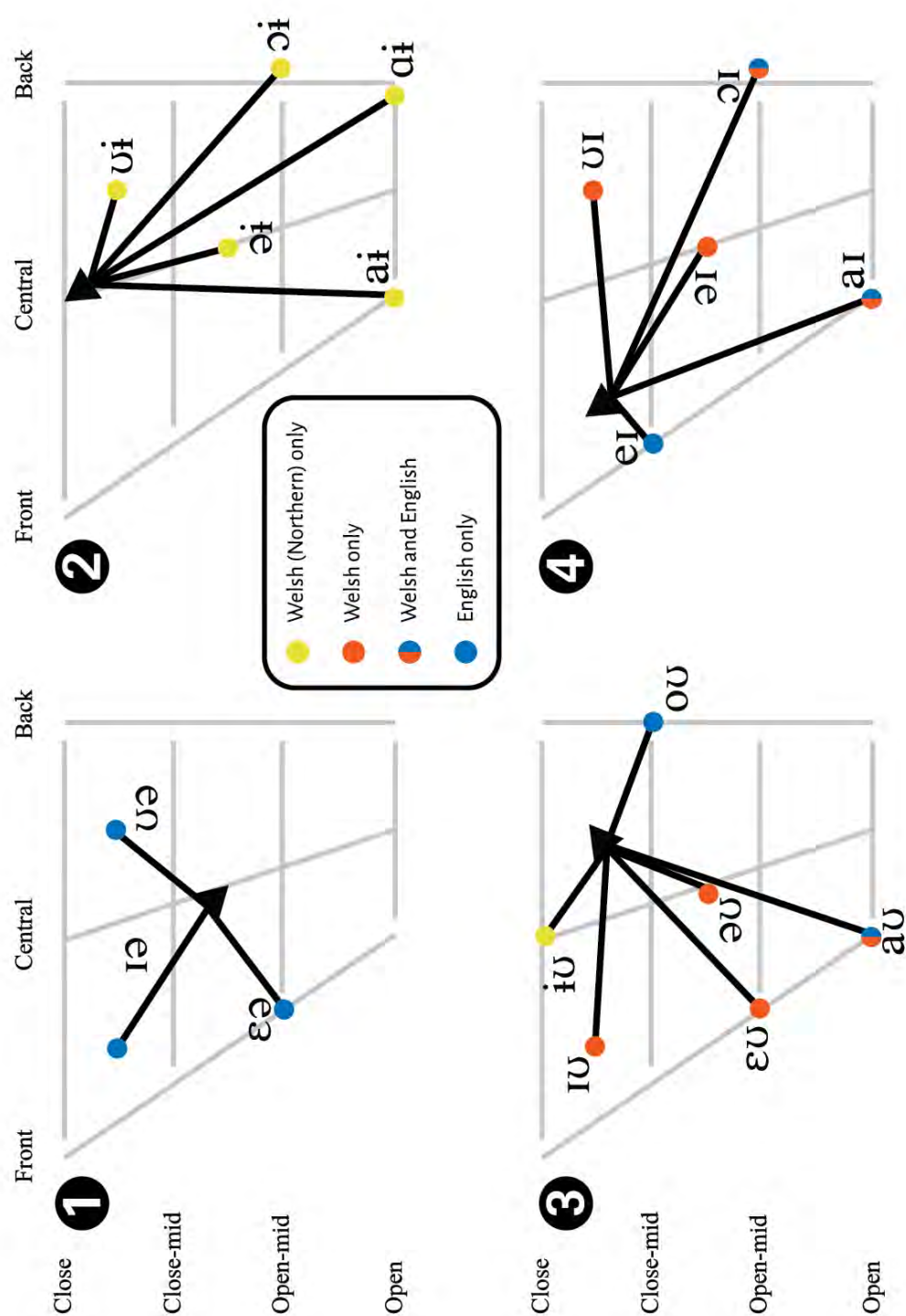


Figure 4.6: The diphthongs of Welsh and English.

4.i.d Schwa and «siwgar»: cowel reduction and vowel epenthesis in Welsh and Welsh English

As noted in Appendix A, schwa has significantly different functions in English and in Welsh, acting as a default unstressed vowel in the former and as a full vowel in the latter (Hannahs, 2013). Welsh schwa differs from the other Welsh vowels in some respects, however, as is indicated by the fact that the <y> set, in contrast to the other six vowel sets, contains only one phoneme. Schwa cannot appear in monosyllables (other than proclitics, which do not behave prosodically like other words). Since monosyllables provide the only context in which vowel length is phonemically distinctive in all varieties, the prohibition on schwa in this context may explain why <y> does not contain two (or more) vowel phonemes distinguished by length, as the other sets do Hannahs (2013, p. 23).

In the Welsh language, word-final [C C_{sonorant}] sequences, especially in colloquial speech, often undergo either the elision of the sonorant, or sonorant cluster epenthesis (Hayes, 2009, p. 238), which inserts an epenthetic vowel before the sonorant (Hannahs, 2013, p. 87). In some colloquial Welsh varieties, therefore, {ffenestr} and {perygl} are ['fe:nɛst] and ['pɛrɪg], while {llwybr} and {rhugl} are ['lʷibɪr] and ['rɪgɪl]. Epenthesis turns the final syllable of the underlying word into a bisyllabic sequence⁸. The elision of the final sonorant is generally restricted to words whose underlying form is already polysyllabic: <llwybr> cannot become *['lʷib], or rhugl *['rɪg], for example. Some polysyllables show a strong preference for the elision of the sonorant (perygl is not normally realized as ?['pɛrɪgɪl]), whereas others are attested with either sonorant elision or sonorant cluster epenthesis: common realizations of <amhosibl> include both [ɛm'hɔsɪbɪl] and [ɛm'hɔsɪb] (GPC *Ar-lein*, 2024, s.v. amhoseibl, amhosib, amhosibil).

In Welsh words with a final consonant + sonorant cluster, the epenthetic vowel articulated in colloquial varieties is not some uniform default vowel, but represents a notional underlying vowel. This is usually the short vowel belonging to the same vowel set as the syllable's existing vowel (Hannahs, 2013). The word <pobl> is therefore realized as ['pɔ:bɔl], <cythrwbl> as [kəθ'ɹɔbɔl], and so on. If the preceding

⁸ As Mayr et al. (2015) point out, words such as rhugl or llwybr are monosyllables in Welsh, whereas English words with similar sequences of consonants treat the sonorant as a syllabic consonant, the nucleus of a second syllable. Where Welsh rhugl has the monosyllabic phonological form /rɪgɪl/, English wriggle has the bisyllabic /'rɪɡl/.

vowel is a diphthong, the epenthetic vowel is of the same set as its second element: <beibl> is realized as [ˈbøibɪl], <llwybr> as [ˈlʷoibr̩], and <powdr> as [ˈpəudər].

Many Welsh English accents resist the reduction or elision of unstressed vowels (Coupland, 1990, p. 246), perhaps as a result of influence from Welsh, which generally does not permit schwa in an unstressed syllable (Wells, 1982b, p. 387). The tendency against vowel reduction in WE has important implications for Welsh loanword adaptation practices. In English forms with word-final [C V C_{sonorant}] or [C V C_{sonorant} C] sequences, PSEE pronunciations often realize the underlying [V C_{sonorant}] sequence with a syllabic sonorant (although there is a contemporary shift towards articulating a schwa before the sonorant in some varieties) (Lindsey, 2019): PSEE squirrel is therefore typically realized as [ˈsgwɪɹ̩] and thousand as [ˈθəuzɹ̩nd].

Typical WE realizations of this type of [V C_{sonorant}] sequence either contain schwa or, in some cases, treat the grapheme(s) representing the syllable's vowel as representative of an underlying short vowel in that position. Parry (1990a, 1990b), for example, records the use of the **DRESS** vowel in the final syllable of {squirrel} and the **TRAP** vowel in the final syllable of {thousand} in parts of southern Wales. WE {squirrel} may therefore be realized as [ˈsgwɪəl] or [ˈsgwɪɛl], and {thousand} as [ˈθəuzənd] or [ˈθəuzɛnd]. The “dark” or velarized allophone of /l/, [ɫ], which appears in many contexts in PSEE, is much less common in WE accents (Walters, 2001). These final syllables resemble the sequences produced by Welsh sonorant cluster epenthesis, but unlike the Welsh examples, the WE pronunciations of English words have a vowel conditioned by the Standard English orthography, rather than by the vowel of the preceding syllable. WE squirrel is therefore not normally realized as *[ˈsgwɪɹ̩], and WE thousand is not realized as *[ˈθəuzɹ̩nd].

In Cymrophone discourse, therefore, the quality of the final vowel in such words may be considered as a sign of an English-origin ELU's level of adaptation to Welsh. An orthographic form such as «siwgar», from the English word sugar, implies the phonetic form [ˈʃʊgər]. This reflects a WE pronunciation, conformant to the WE tendency to resist vowel reduction, where the primary influence on the unstressed vowel is the Standard English orthography. Although «siwgar» is adapted to Welsh orthographic norms, therefore, the form nonetheless suggests an adherence to the phonological norms of WE, rather than adaptation to Welsh phonology. On the other hand, the form «siwgwr» shows both phonological and orthographic adaptation to Welsh norms. It implies the phonetic form [ˈʃʊgʊr], which further implies an underlying phonological form /ˈʃʊgɹ̩/ with sonorant cluster epenthesis to insert the second vowel [ʊ] in the phonetic realization. The variant «siwgr» is also fully

assimilated to Welsh orthographic and phonological norms: the written form implies underlying /'ʃʊgɹ/, with a final cluster of consonant + sonorant which would be illegal in an English word. This orthographic form does not rule out a phonetic realization such as [ʃʊgʊr] with sonorant cluster epenthesis, but represents the standard underlying coda cluster /gr/, rather than the colloquial, epenthetic sequence [gʊr].

4.ii Orthographic adaptation of ELUs: a review of the literature

The brief description of Welsh orthography given above gives an idea of some of the challenges in adapting English-origin ELUs to the Welsh orthographic system. Although both languages use a version of the same alphabet, differences in their phoneme inventories and in their approaches to mapping those phonemes to graphemes, mean that the same graphemic sequence may imply very different phonological realizations in Welsh and in English.

Some of the distinctions identified in Welsh English, but absent from PSEE, may correspond to orthographic differences in the English words. In Welsh Englishes with a **FACE**~**BAIT** split, the monophthongal realization (**FACE**) occurs unless there is <i> or <y> in the orthographic representation, in which case the realization is diphthongal (**BAIT**) (Walters, 1999). Connolly (1981, p. 52) finds a few exceptions to this rule in some southern varieties of WE, such as the common absence of the **FACE** vowel before nasals (so that “chamber” and “danger” contain the **BAIT** vowel) and the consistent use of the **BAIT** vowel in the suffix “-ation”.

KNOW occurs in words spelt with a digraph whose second grapheme is <w> or <u>, or spelt with <o> followed by <l>. Other orthographic forms are usually pronounced with **GOAT** (Connolly, 1981; Walters, 1999). **GOOSE** is produced when the orthographic representation contains an <o> (as in “to”, “too”, “two”, “shoe”) (Connolly, 1981; Walters, 1999). The **FLEW**~**TUNE** vowel is used when the orthography has <ew>, and in many cases where it has <u>, <ue>, or <ui>; but where any of the latter three is preceded by <r> or <l>, either **GOOSE** or **FLEW**~**TUNE** may be used (Connolly, 1981).

It is possible that such variation can ultimately be traced back to phonological differences, rather than being orthographic in origin. An English orthographic form is often a fossilized record of an historical pronunciation which differs, perhaps considerably, from the modern one. Differences in the spelling of pairs which are homophones in PSEE may therefore ultimately reflect differences in historic pronunciation, so that the words were not homophones in English when they were first acquired by Cymrophones. However, Wells (1982b) proposes that at least some of these pairs may instead reflect a stage when Cymrophones were learning English as a second language using written materials. The **FACE**~**BAIT** and **GOAT**~**KNOW**

splits could, in his view, be the result of spelling pronunciations acquired by Cymrophones with English as a second or other language acquired through study Wells (1982b, p. 384). If this is correct, then the influence of orthography on loanword phonology has a long history in Welsh.

Improved access to digital typesetting technologies means that many people now have greater access to scripts and characters other than those traditionally used to represent their language. Whereas typewriters were generally capable of producing only one script family, computers and printers can now theoretically produce the characters of any writing system in the same document. As a result, some languages traditionally represented with non-alphabetic scripts, such as Chinese, are seeing a substantial rise in the use of Roman-alphabet loanwords within texts which are otherwise orthographically Chinese. Many of these Roman-alphabet terms have gained inclusion in Chinese dictionaries (L. Li & Gong, 2018; Xia, 2010).

In cases where ELUs are adapted to an orthography from the same script family, the process may still experience both practical and ideological difficulties. For example, tensions between fidelity to EL pronunciation and fidelity to IL phonological or orthographic requirements may result in variant spellings of established loanwords. The officially preferred versions of loanword spelling as reflected in dictionaries or other standards may therefore not always match actual orthographic practice (Suyuti, 2018; Vasilieva & Ammosova, 2019). In addition, conventions for loanword orthography may differ according to linguistic register, or change over time (Irwin, 2011).

Although it is rarely considered as an influence on LW phonology, the orthographic possibilities of the IL may also affect how an ELU is phonologically adapted when it is borrowed (Daland et al., 2015; Hamdi, 2017; McPeck, 2017; Vendelin & Peperkamp, 2006). McPeck (2017) has shown that the Japanese katakana syllabary is used in innovative ways to represent loanword sounds, using graphemic combinations that are not represented in ILUs to help speakers produce a phonological form that is as close as possible to the ELU's original pronunciation. The original orthographic form of the ELU may also affect how it is produced phonologically, an influence often particularly obvious in unstressed syllables of loanwords (Daland et al., 2015; Hamann & Colombo, 2017). Experimental studies suggest that, in some language pairs, speakers use a combination of phonological perception and EL orthography to produce pronunciations of loanwords (Detey & Nespoulous, 2008; Hamann & Colombo, 2017; Mathieu, 2012).

Concern about linguistic authenticity and the threat from more dominant languages may lead to a desire to represent all ELUs using only the original graphemes of the IL orthography, even if this means changing the form of established loanwords (Imamura, 2017). Orthography becomes enregistered as a symbol of authenticity, part of a broader wish to maintain or reassert cultural identity, and the phonology of the ELU may take second place to such cultural meanings.

As noted above, most of the consonants represented by the Welsh alphabet can be mapped to phonemically similar consonants in English. However, English postalveolar fricative and affricate consonants have no counterparts in the phoneme inventory of Welsh ILUs and, as a result, Welsh has developed novel orthographic

conventions for representing these phonemes. In the colloquial spoken Welsh of some regions, [d̥ʒ] is found as a realization of /dj/, so that the words “diawl” (/ˈd̥jɛul/) and “esgidiau” (/ɛsgɪˈd̥jɑi/) are realized as [d̥ʒaul] and [ɛˈsgɪd̥ʒɛ] (P. W. Thomas, 1996, p. 752). Phonemic /d̥ʒ/ and /t̥ʃ/, along with /ʃ/ and /ʒ/, are found only in loanwords, however. Unlike the consonants of Welsh ILUs, these English consonants may be represented in more than one way, as different conventions have been adopted at various times and in various places (P. W. Thomas, 1996).

4.iii Orthography as a locus of cultural expression

The final section of this chapter gives a brief review of the literature discussing orthography as an individual and social practice. The chapter concludes by discussing a recent study of the enregisterment of Welsh varieties of English on Twitter, which will provide interesting parallels to the Cymrophone data to be discussed in subsequent chapters.

4.iii.a Ghoti and foreignness: orthography and its relationship to phonology

The word “orthography” has its roots in Greek ὀρθο- (correct) and -γραφία (writing), implying that orthography is not merely a system of writing or spelling words, but is the correct system of so doing. The word “heterography”, although significantly less popular, is also attested in the *OED2* (n.d.) to refer to incorrect or irregular spelling. I will argue, however, that a third term is useful for discussing the pressures behind orthographic adaptation of loanwords. Let us take the famous example of the spelling <ghoti> as a representation of the English word pronounced /fɪʃ/, and spelled <fish> in Standard English (Zimmer, 2008). This word is the classic example of the inconsistencies of English spelling, with the <gh> representing /f/ as in <enough>, <o> representing the first vowel of <women>, and <ti> the middle consonant of <nation>.

What this facetious respelling actually does, however, is show that English spelling is rather more systematic than it might appear. No competent user of English would produce /fɪʃ/ as a plausible rendering of <ghoti>, or write <ghoti> even if she was uncertain how to spell /fɪʃ/, since the three sequences <gh>, <o>, and <ti> represent /f/, /ɪ/, /ʃ/ only in very specific contexts (<gh> can only represent /f/ in non-initial position, for example (Borgwaldt et al., 2005)). Rather than an example of inventive orthography, or even of heterography, <ghoti> might be called “scoligraphy” (Greek σκολιός -‘crooked’, ‘bent’); a contortion of the purpose of spelling. It deliberately deforms the conventions of English phoneme-grapheme correspondence, so that they obscure the phonology of the word rather than attempting to convey it. Although they do not represent deliberate departures from the IL’s orthography, loanwords are nonetheless often scoligraphic in IL context. They use the expected graphemes of the IL, but may use them to represent

phonological forms that would never be produced from the given orthographic sequence by a competent user of the IL.

As we have seen in section 1.iii.a, phonological assimilation of loanwords may be an automatic, or even unconscious, process. Phonological features that are not allowable in the IL may, in fact, never be perceived when an ELU is encountered. However, writing is also subject to constraints that are different in kind from those that govern phonology. Importantly, difficulty in producing the graphemes of a given loanword is unlikely to exert significant pressure on written forms (except perhaps in the production of accented letters, which may be hard to access in some contexts of computer-mediated communication), since language users who are literate in a particular script (e.g. Latin or Cyrillic or Arabic) will be able to produce the characters of that script equally well, regardless of whether they are writing IL or EL words. That is to say, two languages that use the Latin alphabet may have significantly different phoneme inventories, but those phonemes are mapped on to graphemes familiar to writers of any other Latin-alphabet language ⁹. In contrast, producing sounds or sound sequences that are not part of the IL may be difficult or impossible for non-EL speakers, leading to a higher probability that those sounds will be adapted when an ELU is assimilated into the IL.

The depth of a language's orthography will affect the extent to which loanwords in the same basic script will be perceived as scoligraphic. Orthographic depth refers to the extent to which a language's writing system uses a one-to-one phoneme to grapheme mapping. A shallow (or "transparent") orthography features numerous one-to-one correspondences, and few cases in which a single phoneme can be represented with numerous different graphemes, or in which a single grapheme can represent more than one phoneme. Once a person knows how the grapheme-to-phoneme mapping of such an orthography is structured, and with practice, they can fairly reliably reconstruct the pronunciation of written words (Ziegler & Goswami, 2005). A deep orthography, in contrast, is irregular, and the spelling of a word does not reliably indicate its pronunciation (Pérez Cañado, 2005). English orthography is characterized by numerous one-to-many and many-to-one correspondences between graphemes and phonemes (Pérez Cañado, 2005), making it one of the most ambiguous alphabetic writing systems in the world (Borgwaldt et al., 2005; Share, 2008). In contrast, Latin-script languages such as Hungarian and Italian have relatively shallow orthographies, although even these languages feature some ambiguities and deviations from one-to-one mapping (Borgwaldt et al., 2005).

Welsh is another Latin-script language with a shallow orthography (Ball & Williams, 2001), a feature which has long been noted as an advantage for its users. T. Edwards (1847, p. 9), for example, argued that "in whatever situation the letters are placed in the composition of words, they have but one invariable sound, which

⁹ There are exceptions, of course, some of which are particularly salient in the Welsh context. These will be discussed below. However, on the whole, the variation in grapheme inventory between any two Latin-alphabet languages is likely to be much smaller than the variation in phoneme inventory between the same two languages.

simplifies the language much”. Although this is something of an exaggeration, certainly as regards modern Welsh, the shallowness of Welsh orthography has been shown to help Cymrophone children in learning to read, as compared to their Anglophone peers (N. C. Ellis & Hooper, 2001; Spencer & Hanley, 2003, 2004). These advantages are of particular benefit to poor readers, who trail far less strongly behind their peers when learning a shallow orthography (Spencer & Hanley, 2004).

In adult readers of alphabetic scripts, orthographic cues to word identification are apparently available more quickly than phonological cues. Readers generate a phonetic representation (or representations, if the word is a homographic heterophone) which lags behind the use of orthographic cues, but nonetheless contributes to word identification (Frost & Kampf, 1992). Experimental data support the conjecture that, for readers of an alphabetic script, both language-specific orthographic rules and the pronunciations of existing words in the language affect how newly-encountered written words are converted into a phonological form (and, indeed, how spoken words are converted into orthographic form). The former is termed “sublexical knowledge” (because the rules generally involve graphemic units smaller than the whole lexeme), and the latter “lexical knowledge” (involving entire lexemes) (Apel et al., 2019).

Effects from the reading aloud of pseudo-words (words that follow IL phonotactic and orthographic rules, but happen not to exist in the IL) suggest that reading in a shallow orthography involves both lexical knowledge (activated by existing words and causing the inconsistent processing of pseudo-words when seen alongside existing words) and sub-lexical knowledge (activated in the lists of pseudo-words) (Job et al., 1998; Marcolini et al., 2009).

The primacy of orthographic cues in reading a word, alongside evidence that bilinguals use these cues to make decisions about a word’s language and, closely thereafter, its pronunciation, explain why EL words may be scoligraphic in an IL context. Surrounding context cues the bilingual reader to expect an IL word (Casaponsa et al., 2015), and if the word has no sublexical cues to indicate language membership, the reader tries to use IL grapheme-to-phoneme mapping rules to produce a phonological model of the word. Depending on how far the two languages’ orthographic systems differ in their grapheme-to-phoneme mappings, the phonological representation produced may be unrecognizable as the target EL word. Although the reader may be able to recover from the error, particularly if the word is frequent in the EL, this is nonetheless a potential cause of confusion and frustration in attempting to read IL texts containing unadapted EL lexemes.

Kolers (1966) found that primary-language French speakers with English as an additional language sometimes mispronounced French words as though they were English, when reading aloud passages in which English and French words were mixed: for example, the word <nuages> in the phrase ‘Plus tard came nuages de moustiques’ was pronounced as /nu'eidʒiz/, and <queue> in ‘A peacock spread its queue sur le gazon’ as /kjūu/ (p. 369). Primary-language English speakers in the same study made similar errors, for example producing the English <moats> as /mo/, or

<warden> as /vaɹdã/ ¹⁰.

This fascinating result suggests that processing of written words may depend in part on the language we *expect* to be reading. The written form <nuages> is not a known English word, and yet the subjects (who, it must be reiterated, were primary-language speakers of French, and not of English) were apparently primed by the preceding English verb ‘came’ to attempt an interpretation of <nuages> according to English grapheme-phoneme correspondence rules. Kolars (1966) suggests that the translation of graphemes into phonemes when reading aloud is a process only begun after first making a categorial decision as to which language a word belongs to.

Bilinguals appear to be very sensitive to the statistical orthographic patterns of their languages (Casaponsa & Duñabeitia, 2016). Language-specific letters and bigrams have been shown to be sufficient to facilitate bilinguals’ lexical decisions, suggesting that sublexical information can determine a word’s language membership without reference to lexical information (van Kesteren et al., 2012). Whatever the source of the effect, it is evident that bilinguals respond differently to graphemic strings that are orthographically legal in only one of their languages, as compared to strings that are legal in both languages (Casaponsa & Duñabeitia, 2016; van Kesteren et al., 2012)).

Lemhöfer and Dijkstra (2004) and van Kesteren et al. (2012) report experiments with highly-proficient bilinguals who are users of two Latin-script primary languages, presented with lists of pseudo-words, some of which contained sublexical information which was specific to one or other of their languages (“sublexically-marked”) and some with no language-specific sublexical information (“sublexically-unmarked”). The subjects were able to reject unmarked pseudo-words more quickly than unmarked ones (Lemhöfer & Dijkstra, 2004), and identify the language of sublexically-marked words more quickly than those without (van Kesteren et al., 2012). Orthography, according to this evidence, is enregistered with languageness, and this languageness manifests itself prior to the recognition of the phonological information conveyed by the orthography. This supports the model of language proposed in Chapter 2, which sees orthography as potentially a site of connotational meaningful in itself, as well as an expression of sublexical meaning through its connection with phonology in the second articulation.

4.iii.b Haitch and humour: orthography as cultural practice

English writing has a long history of representing linguistic difference orthographically. The respelling of words to suggest particular pronunciations or to ascribe particular characteristics to speakers has been used in countless canonical literary works, as well as in other genres such as journalism, comics, and advertising (Gómez-Jiménez, 2015; Malin, 1965; Walpole, 1974). The umbrella term *eye-dialect*

¹⁰ Phonetic transcriptions are my interpretations of Kolars’ information that the words were pronounced like English ‘new ages’ and ‘kew’, like French ‘mots’, and as ‘vahrdaahn’, respectively.

is now sometimes used to describe both standard and non-standard pronunciations represented by non-standard spellings. The original coinage of the term, however, was intended only to cover the non-standard spelling of standard pronunciations, such as <enuff> for <enough> and <pictsher> for <picture> (Krapp, 1925), identified as a shorthand for the speaker's social inferiority Krapp (1925), and now widely understood by readers to index low socioeconomic status (Jaffe & Walton, 2000; Walpole, 1974).

Eye dialect, used narrowly to refer only to forms which represent a standard pronunciation in a non-standard way is contrasted by some linguists with dialect respellings, which reflect non-standard pronunciations orthographically. The distinction is useful for distinguishing between respellings that are motivated by actual pronunciation features of some kind, and those that are alternative spellings of standard pronunciation and are therefore primarily indexical of sociocultural information (Jaffe, 2012).

Applying these categories to Welsh can be challenging. Distinguishing eye dialect and dialect respellings in Welsh is harder than in English, since no one form of pronunciation is treated as a standard in Welsh (Ball & Williams, 2001). As a result, variant spellings might represent more than one 'normal' pronunciation of a word, without falling into the category of dialect respelling. For example, the Welsh plural ending <au> is regularly realised as [a] in Gwynedd and in the eastern Valleys region of southern Wales, but as [e] elsewhere. A non-standard spelling of the word *pethau* ("things") could therefore be either <pethe> or <petha>, depending on the speaker's background (Awbery, 1996). However, unlike an English spelling such as <fillum> to represent a Northern Irish pronunciation of *film*, the two Welsh forms are not primarily intended to index geographical dialect, but rather to indicate the informality of the written register. Although the forms <pethe> or <petha> are not orthographically standard, the pronunciations they encode are, as it were, generically regional, rather than representing a specific local dialect.

Honeybone and Watson (2013) suggest that respellings can be analysed as a form of social practice, in which orthographic choices are conditioned by the relative salience of phonological features. Both phonological and sociolinguistic salience may affect the likelihood that a feature will emerge in the orthography, with stigmatized features such as "h-dropping" particularly likely to emerge because of their widespread familiarity as signals of non-standard English pronunciation (Honeybone & Watson, 2013).

The socio-symbolic nature of respelling in representing linguistic varieties is further indicated by the fact that it is only one of a number of potential strategies that might be employed. Percillier and Paulin (2017) found, for example, that the Anglophone literatures of different regions have marked differences in how, and to what purpose, they represent dialect features. The Scottish literature in their study used respelling far more frequently than either the West African or the Southeast Asian literature, whose preference was for non-standard grammar and code-switching into other languages, respectively, suggesting that respelling, especially in monolingual contexts, may perform some of the same functions as code-switching does in bilingual ones. Percillier and Paulin (2017) theorize that

authors select the features that they feel will best represent the difference of the variety they are representing (Percillier & Paulin, 2017).

While Honeybone and Watson (2013) show that even respellings created with humorous intent can represent subtle phonological features quite accurately, Percillier and Paulin (2017) found that the relative frequency of phonological features observed in the variety itself is not perfectly reflected in literary respellings; rather, features seen as emblematic of the variety are given greater prominence. As Chapter 6 will show, orthographic adaptation of English-origin ELUs by Cymrophones shows a similar tendency, with features strongly enregistered with Welshness becoming particularly salient. Respellings and orthographic adaptations are not direct representations of pronunciation. Rather, they are symbolic of the authors' perception of their own (and/or others') pronunciations, mediated by a range of linguistic and sociolinguistic variables.

Whereas standardized varieties of languages tend to enforce more or less rigid orthographic practices, less-standardized or non-standard varieties may offer their users more freedom. Nonetheless, when using varieties that have not yet developed a standard orthography, users may choose to adhere to the conventions of a dominant language variety. For example, Caribbean Creoles have no standardized orthography, having traditionally been oral languages without the status for use in education, law, or government (Hinrichs, 2004; Sebba, 2000). British Creole speakers tend to choose something close to the orthography of Standard English when they are writing Creole (Sebba, 2000). Writers who code-switch between Standard English and an English-lexified Creole use various approaches to respelling, some directly representing distinctive pronunciation, some indicating that the word is used with Creole grammatical or semantic properties, and some simply indexing "Creoleness" (Hinrichs, 2004). This third type of orthographic adaptation, signalling some aspect of identity as distinct from the authoritative standard, fits in with what Sebba (2003) calls "rebellious spellings", a designation which also encompasses eye dialect forms such as <skool> for Standard English <school>.

Indeed, Sebba (2003) identifies the grapheme <k> as having particular cultural symbolism in a number of languages: its relative rarity in Spanish, for example, exoticizes <k>, making it a prime choice in counter-cultural texts such as graffiti. The tendency to ascribe difference and glamour to <k> in English was identified as far back as 1925, when Pound (1925, p. 44) wrote that "there is no mistaking the kall of 'k' over our kountry, our kurious kontemporary kraving for it, and its konspicious use in the klever koinages of kommerce". Other linguistic commentators of the era noted tendencies towards simplification in the language of commerce and advertising, notably Alexander (1929), who wonders whether the current fad for using <u> to represent the second-person pronoun {you} will die out. This use of <u> has, of course, become characteristic of the language of computer-mediated communication, a result of the pressures of character-counts in early SMS messages as well as of the more generally playful and non-standard style of new media language.

Rare graphemes which are, nonetheless, members of the IL alphabet are not the only possibilities for emblematic orthography. A variety of cultural attributes may be indexed using script mimicry, incorporating graphemes from outside the given

writing system or script (Strandberg, 2020) or by writing the graphemes of one script in a way that resembles the graphemes of another (Sutherland, 2015). The use of so-called ‘metal umlauts’ or ‘rock dots’ by hard rock and heavy metal bands such as Blue Öyster Cult, Mötley Crüe, and Motörhead has been analysed as contributing to extensive use of stereotypical Germanic cultural symbolism within these musical genres, emblematic of militarism and violence (Spitzmüller, 2012). Importantly, Spitzmüller (2012) shows that these diacritics are not usually intended to have any effect on pronunciation.

It is not uncommon to see examples of script mimicry in which English words, particularly those containing the graphemes <r> and <n>, are instead spelled with the Cyrillic graphemes <я> and <и> in order to index Russianness (and associated historical and cultural referents, such as Stalinism). The Cyrillic graphemes are deployed because they are understood to have Russian languageness by Anglophones who have seen written Russian, even (perhaps especially) Anglophones who do not understand Russian. It is possible that Cyrillic «Я» and «И» are particularly salient to users with knowledge of the Latin alphabet, because they look very like mirrored versions of the familiar Latin graphs «R» and «N». They are therefore different enough from the Latin graphs to be distinctive (unlike Cyrillic «У» or «В», which look identical, or nearly so, to standard Latin forms of «Y» and «B»), but are not wholly unfamiliar shapes, like Cyrillic «Ю» or «Ж».

Figures 4.7–4.9 show the cover of a stage adaptation of George Orwell’s *Animal Farm*, as well as two posters for stage versions of the novels. The cover design for the playscript (Wooldridge, 2005) (fig. 4.7) uses only one Cyrillic grapheme, <и>, in the word «АИИМАЛ». The poster for Tell Tale Theatre’s 2014 production (McGuire, 2014) (fig. 4.8) leaves the Latin <n> unaltered, instead using «Я» to replace «R», as well as using a form of capital <a> which is strongly reminiscent of Cyrillic «Д». The most extensive example of script mimicry shown here is in the poster for Blake Theater’s 2018 production (*Blake Theater*, 2018) (fig. 4.9), which uses Cyrillic graphemes extensively in both Orwell’s name and the title of the play. Latin «E», «O», «R», «W», «A», and «N» are replaced by Cyrillic «Э», «Ф»/«Ю», «Я», «Ш», «Д», «И».

The words “George Orwell’s Animal Farm” as represented in fig. 4.9, giving the Latin graphs English phonemic values and Cyrillic graphs Russian ones, and assuming that any graphs shared by both scripts are Latin, might be pronounced /dʒɜfˈɒdʒɛ juʃəʃɛlz dɪrmdl fdʒam/. Knowledge of Russian, or of any other language using the Cyrillic alphabet, is not only unnecessary, but potentially a drawback, in interpreting the text as it is intended to be read. What is needed is recognition of the Cyrillic graphs both as imbued with Russian languageness, and as geometrically similar enough to Latin graphs to be interpretable as stylized versions of those graphs. The Cyrillic graphs are, at the level of the second articulation, to be read as Latin graphs, while retaining their Russian languageness at the connotational level.



Figure 4.7: Cover design for *George Orwell's Animal Farm*, adapted for stage by Ian Wooldridge



Figure 4.8: Poster for *George Orwell's Animal Farm* as performed by TellTale Theatre in 2014.



Figure 4.9: Poster for George Orwell's *Animal Farm* as performed by Howard W. Blake School of the Arts in 2018.

4.iii.c Words and Way-uls: The orthographic enregisterment of Welsh English

Durham (2022) studied dialect respellings representing a Welsh-English accent in a corpus of tweets, concluding that the features of a Welsh accent which were salient to people outside Wales were rarely the same as those salient to speakers of Welsh English themselves. Durham's first dataset included tweets in which a Welsh accent was performed or discussed by people who were not speakers of Welsh English. The second included tweets by Welsh-English speakers, commenting on their own accent.

A restricted set of phonological features emerged as particularly salient to these two groups of tweeters. For the Welsh-English speakers, the set was quite small. The phenomenon of breaking (the insertion of schwa between [j] or [w] and [t], causing the insertion of an epenthetic schwa vowel before the final consonant of words such as {owl} and {real}) was frequently mentioned and performed, as were a number of specific phonemes. The most salient of these were the vowel of **NEAR**, and the **GOOSE** vowel's realization as [ʊ] in words like "tooth". These showed very little overlap with features which were most salient to non-Welsh-English speakers. Breaking was once again a fairly frequently mentioned and performed feature, but the **NEAR** and **GOOSE** vowels were rare. A number of features which were missing from the dataset of Welsh-English speakers were also quite frequent for the non-Welsh-English tweeters: the gemination of consonants after a stressed vowel; "g-dropping" and "h-dropping"; the realization of /ɹ/ as a flap or trill; the realization of the **GOAT** vowel as a monophthong; and the glottalization of /t/ (Durham, 2022).

Durham (2022, p. 209) argues that the non-Welsh-English dataset's preference for using respelling for some features more than others is partly because these features are easier to represent orthographically: the elision of /h/ can easily be shown by omitting <h> from the orthographic form; "g-dropping" by using <in> in place of <ing>; and glottalization of /t/ by replacing <t(t)> with <'>. Nonetheless, the differences between what is salient to insiders and to outsiders cannot be wholly ascribed to ease of representation, and Durham notes that there is quite some difference between the Welsh and non-Welsh linguistic consciousness in what has been enregistered as distinctively Welsh (Durham, 2022).

Early in her discussion, Durham notes that a lack of conventionalization leads to significant variation in how respelling represents Welsh-English accents: "respellings of the word *Wales* on Twitter range from <Way-uls> to <Waaales> to <Wêls>" (Durham, 2022, p. 190). The third example is evidently drawing on Cymrophone orthographic conventions. The use of the circumflex, as we will see, is a clear visual indicator of orthographic Welshness, and the representation of the **FACE** vowel (in its monophthongal WE form, /e:/) as <ê> is distinctively Cymrophone. Nonetheless, Durham makes little mention of the influence of the Welsh language in her study, except to comment on the use of <w> to represent the vowel in words such as {school} and {tooth} (Durham, 2022). Her conclusions about the lack of conventional ways to represent many features of Welsh-English accents might be enhanced by considering the different orthographic resources made available by

Welsh and English.

The linguistic background of the tweeters in Durham's Welsh-English dataset is largely unknown, and so it is impossible to tell the extent to which Cymrophone orthography may have influenced their respelling practices. In forms such as <Wêls>, however, it is impossible not to see the direct influence of the Welsh language. The extent to which Cymrophone orthographic conventions are understood by non-Cymrophones in Wales is not known, and it is possible that at least some of the conventions which are highlighted in stereotypical representations of Welsh, such as the use of <w> to represent vowels, have become enregistered for non-Cymrophones as resources for representing Welsh English. Although the discussion in this study focuses on Cymrophones, further research on the availability of Cymrophone conventions to Welsh non-Cymrophones would provide useful context for Durham's (2022) work.

4.iii.d Gaelic and gobbledigook: the Celtic languages under the Anglophone gaze

As we saw at the outset of this chapter, the visual impact of Welsh, which is significantly different from English in grapheme frequency and orthotactic patterns, is foreign to Anglophones. Examples of social media comments mocking Welsh orthography were presented in Chapter 1.vi.a. Some other modern Celtic languages are treated similarly. Typical Twitter comments include "Irish spelling looks like someone was high" and "Gaelic looks like transliterated chewing noises".

Unlike the orthographies of Irish and Scottish Gaelic, that of modern Manx is based on the conventions of Early Modern English, having been developed by Manx users who were also primary- or additional-language Anglophones (Lewin, 2020; Ó Ciosáin, 2004; Ó hlfearnáin, 2007; Sebba, 2012). As a result, Manx orthography has most prominently been denigrated, not by monoglot Anglophones, but by scholars of Celtic languages working with "puristic nationalist and pan-Gaelic assumptions", who have tended to view it as Anglicized, inconsistent, and as a barrier to access for those who can read Irish and Scottish Gaelic (Lewin, 2020, p. 54). Lewin (2020) convincingly rehabilitates the orthographic standardization of Manx, defending it in particular from the charge of an uncritical reliance on English conventions. He argues that Manx orthography must be assessed on its own merits, as a system which avoids some of the defects of traditional Goidelic orthography, and whose developers prioritized the linguistic needs and sociolinguistic conditions of its users.

While critics of Manx orthography have described it as placing barriers between users of the Goidelic languages (P. Gawne, 2002; Sebba, 2012), Lewin (2020) finds, in turn, that their criticisms have been responsible for harming the profile of the Manx language within Celtic Studies, and led to scholarly neglect of the orthography as evidence for historical phonology. The development of an orthography which differed from the established Goidelic system damaged cultural ties between Manx and its sibling languages on a number of levels, despite the fact that the languages themselves share such strong structural and lexical commonalities. Importantly, it was

responses from users of other Goidelic languages to Manx's apparently English-influenced orthography which caused the rift. "Like other elements of culture, alphabets have both practical and ethnopolitical integrating or differentiating significance" (Landau, 2010, p. 25).

Deliberate deployment of orthography to enhance or reduce cultural closeness has been noted in states and areas under Russian control in the Tsarist, Soviet, and post-Soviet periods. The Turkic vernaculars, with a history of literacy using the Arabic script, were forced by communist rulers to convert first to the Roman script in the late 1920s, and then to the Cyrillic script in the late 1930s (Sebba, 2006). Post-Soviet de-Russification has seen moves to reject Cyrillic (Pavlenko, 2009), including in the six Islamic states which gained independence after the fall of the USSR (Azerbaijan, Uzbekistan, Kazakhstan, Kyrgyzstan, Turkmenistan and Tajikistan) (Landau, 2010). Settling on a script for the national languages of these states has been a difficult process, however, with internal pressures from those who have become accustomed to Cyrillic as well as those who see Latin as symbolic of the break with Russia, alongside external pressures from neighbouring states, particularly Iran, which advocates the Arabic-Persian script, and Turkey, which has encouraged romanization using the same model successfully applied to Turkish (Landau, 2010). In response to developments in these newly independent states, the Russian government passed a law in 2002 enforcing the use of Cyrillic throughout the Russian Federation, preventing semi-autonomous regions, such as Tatarstan, from romanizing their orthographies (Sebba, 2006).

These examples, like those of Manx and the other Celtic languages, show that orthographic systems can enregister strong cultural meanings and are not always merely neutral representations of lexical or phonological content. Irish not only has an orthography that is visibly different from the orthography of English, it was also traditionally printed in *an cló Gaelach* ("Gaelic type"), which is somewhat distinct from the Roman typefaces habitually used to print English (Ó Ciosáin, 2004; Staunton, 2005). Gaelic type is still used, particularly on road signs and other signage (Cisło, 2012). It is also used to write English, an example of script mimicry, with the Gaelic type enregistered with "Celticness" or "Irishness" (Staunton, 2005). Figures 4.10–4.13, for example, show signs on Irish-themed bars and pubs in Europe and the USA, which all use *an cló Gaelach* for English-language text.

Celtic orthographic practices occupy a liminal space, in which the orthography itself is mocked as incomprehensible and random, while its aesthetics are commodified as emblematic of a romanticized Celtic identity. The commodification of the Celtic languages means that they are used to demarcate the Celtic nations as uniquely branded tourist locations, even as visitors are reassured that they need never actually use the languages if they visit (Pritchard & Morgan, 2001). Such attitudes serve to imbue orthography and scripts with social meaning; Welsh orthography is understood not merely as a practicality of the language, but as a site of ideological resistance to English hegemonic linguistic practices. This is the context in which orthographic adaptation of ELUs takes place in Welsh, with both practical pressures from the scoligraphic nature of English words in a Welsh context, and sociocultural



Figure 4.10: Corcoran's Irish Pub – Paris, France. (Image licensed from Adobe Stock Images.)



Figure 4.11: Irish Times – Worcester, MA, USA. (Image licensed from Adobe Stock Images.)



Figure 4.12: Scholars Lounge Irish Pub – Munich, Germany. (Image licensed from Adobe Stock Images.)



Figure 4.13: Irish Pub – Metz, France. (Image licensed from Adobe Stock Images.)

pressure to resist the imposition of English orthographic norms.

Given the tensions surrounding the Welsh language, it is reasonable to conclude that Cymrophones might choose to retranscribe ELUs from English so that they match Welsh orthography as an act of individual identity management or cultural display. As a result of these two complementary sets of motivations for adaptation, the psycholinguistic and the sociolinguistic, the orthography of ELUs offers a rich source for studying the semiotic enscription of Welshness in the language of Cymrophones.

Chapter 5

“For the benefit of Welsh people who had never seen a letter X before”: Designing a study of variation in the orthographic adaptation of English-origin lexical units

They went outside and stood where a sign used to say Taxi and now said Taxi/*Tacsi* for the benefit of Welsh people who had never seen a letter X before.

(Amis, 1986, p. 44)

5.i Introduction

This chapter introduces the three main datasets collected for this study, and describes the methods which will be used to analyse them. The first dataset is a corpus of loanwords automatically extracted from *Geiriadur Prifysgol Cymru* (GPC). The second is a questionnaire, asking Cymrophone respondents to adapt a list of English words to Welsh orthography. The third is the Twitter corpus which provides the largest-scale source of data for linguistic analysis. Having introduced the datasets, this chapter will provide a rationale for using a mixed-methods approach to synthesize and analyze the data.

5.i.a Sources of data for orthographic adaptation

The Twitter corpus described in section 5.iv was the first dataset planned as a source for this study. At the time, Twitter was a particularly well-suited resource for academic research, since it was extremely popular and made the collection of large amounts of naturalistic data fairly easy. The written language encountered on

Twitter tends towards the informal and colloquial, both in English (Durham, 2022) and in Welsh *Willis*2020. However, since the platform is used by numerous professional organizations and official institutions, as well as by private individuals, differences in register might also be expected in accounts tweeting for different purposes and using different kinds of personae.

The survey described in section 5.iii below was designed to complement the naturalistic data collected from the Twitter corpus, but also provides a way of highlighting potential points of interest to be explored further. Both the Twitter corpus and the survey provide data about current orthographic practices. Putting results from analysing these datasets into context also requires an understanding of how Cymrophones have traditionally adapted loanwords to Welsh orthography. To that end, a dataset of grapheme-to-phoneme mappings for established English-origin loanwords was constructed using *GPC*. The construction of this dataset will be described first, since it has implications for both the design of the questionnaire and the processing of the Twitter corpus.

5.ii The orthographic adaptation of established loanwords: building the GPC loanword dataset

At the time of writing, *Geiriadur Prifysgol Cymru* is the only large-scale historical dictionary of the Welsh language. It covers the vocabulary of Welsh from the earliest extant manuscripts up until modern times (*GPC Ar-lein*, 2024). As a result, it is the most comprehensive source available for studying established loanwords in Welsh.

Parry-Williams (1923, p. 16) observed that early-twentieth-century “learned” borrowings into Welsh tended to be deliberately altered “to invest them with more naturalness” by replicating patterns of phonological and orthographic change resulting from natural assimilation processes in older loanwords. This implies an instinctive awareness amongst Cymrophones of traditional orthographic practice regarding loanwords: the aim could not be achieved if the patterns in question did not convey an immediate “naturalness” to the Cymrophone eye. In order to understand patterns of orthographic adaptation of ELUs by Cymrophones, it is therefore necessary to establish something of a baseline, which will quantify features of traditional and conventional loanword orthography in Welsh. The loanwords gathered from *GPC* are an important resource with which to compare the data collected in the Survey and Twitter datasets, as well as providing potentially

interesting insights in their own right.

5.ii.a Identifying loanwords

The size of the dictionary makes manual searching for loanwords prohibitively difficult. As a result, I opted instead to crawl the dictionary’s website using a Python script, and to identify loanwords automatically using contextual clues from the dictionary content. Entries usually indicate the headword’s part of speech, offer a definition in Welsh as well as possible translation equivalents in English, and give one or more illustrative quotations showing the headword in use. In some cases, an entry may list variant spellings of the headword.

A Python script was created which checks entries in the online version of the dictionary, looking for an etymology including the string “bnth.” (short for “benthyciad” – *borrowing*). This process should identify most words which are acknowledged by GPC as loanwords (from any language). As with any automatic process, it is likely that some loanword entries will not be identified correctly; any entry whose etymology is not structured using one of the dictionary’s most common formats for loanword etymologies will not be found, for example. Nonetheless, this initial process successfully found 12,635 entries. Since the output included all loanwords, not just those of English origin, further processing was necessary. This was accomplished by analyzing the text of the etymology and rejecting entries not assigned confidently to an English etymon. The automation of this step is, once more, likely to have resulted in the erroneous removal of some entries; however, 10,018 English-origin ELUs remained at this stage.

5.ii.b Adding phonetic information

The graphemes of the Welsh dictionary headwords were next aligned to the phonemes of their English etymons. Basic phonetic transcriptions were initially sourced from the freely available *CMUdict* (2014) (Carnegie Mellon Pronouncing Dictionary), and automatically converted to a representation of Prestige Southern-England English (PSEE) matching the transcription conventions described in Appendix A. The resulting transcriptions were checked for errors and corrected where necessary, in preparation for the alignment process.

The loanwords were stored along with their English etymon, which could then be looked up in the phonetic dictionary. The task of aligning graphemes and phonemes in this dataset might therefore have been accomplished using some kind of

basic sequence alignment procedure, such as the popular Needleman and Wunsch (1970) algorithm. However, the algorithm would also later be used to predict phonetic representations of unrecognized items in the Twitter corpus. Since the English etymon for these items is not known, the phonetic representation is part of the target to be sought by the algorithm rather than part of its input. It therefore made sense to develop an independent algorithm which, given only the orthographically-adapted string, could generate plausible candidates for the PSEE phonetic sequence underlying it, and subsequently return a candidate Standard English orthographic form or forms.

The algorithm was tested on pairs of loanwords and their English etymons collected from four sources. The words collected from *GPC*, as described above, provided the majority of the training data. To these were added responses to the survey described in the following section, and Wenglish words collected by searching the CorCenCC corpus (Knight et al., 2020). Finally, I checked for phonemes and phoneme sequences which were poorly represented in the data from these three sources, and hand-created orthographically-adapted representations of five hundred additional English words. Altogether, these four sources provided 9,328 unique pairs of words. Some English words in the list have multiple adapted representations, and some adapted strings are matched to more than one English lexeme. The Wenglish form <siwt> is found as a representation of English {suit} and {shoot}, for example, while the English {cabbage} has numerous representations, including <cabatsh>, <cabbait>, <cabej>, <cabets>, and <cabetsh>.

Having obtained matches for all the words in the dataset, the final step was to align the Standard English orthographic forms with the grapheme-phoneme pairs. This was done using the Needleman-Wunsch algorithm, which selects optimal sequences of alignments according to a defined score given to each individual alignment in the sequence. The alignments were then checked by hand and fixed where necessary. The final dataset consists of lists of Wenglish (graphemes) - RP (phonemes) - BrEn (graphemes) triplets. The list for the Wenglish form <ciwcymer>, for example, is [(‘c’, ‘k’, ‘c’), (‘iw’, ‘jʊu’, ‘u’), (‘c’, ‘k’, ‘c’), (‘y’, ‘ʌ’, ‘u’), (‘m’, ‘m’, ‘m’), (‘Ø’, ‘b’, ‘b’) (‘er’, ‘ə’, ‘er’)]. The algorithm was later used to create similar sets of triplets for the wordforms collected from the survey of orthographic adaptation and the Twitter corpus, described below.

5.iii A survey of orthographic adaptation practices

A survey to elicit orthographically adapted forms was designed to complement the naturalistic data collected from the Twitter corpus, and also as a way of highlighting potential points of interest to be explored further. The survey instructions include this passage:

I understand that there are words you would not choose to use in your Welsh, but try your best to adapt them anyway. Remember that it's fine not to adapt a word if the result, in your estimation, would be incomprehensible. If so, respond with the original English word.

The Survey therefore deliberately provides an artificial context in which respondents are focused on the question of orthographic adaptation, rather than using a word in a form that is natural to them, in the hopes of eliciting extreme and less usual patterns of variation which will prove salient for the quantitative analysis. Whereas a Twitter user may simply deploy an English-origin ELU without any orthographic adaptation if that seems most appropriate to them, the survey respondents are explicitly requested not to use the original spelling unless a word needs no adaptation, or cannot be adapted comprehensibly. If any patterns of (non-)adaptation in the survey responses are found to correlate with traditional measures of loanword status, such as a word's listedness or frequency of use in other corpora, this may therefore be considered as potential evidence of the word's enregisterment as Welsh for the survey respondents.

The prompt words presented to respondents were chosen with the phoneme inventory of British English in mind, and the range of phonemes commonly used by speakers of PSEE is represented in the adapted words produced by the respondents. Although the survey data is not naturalistic, it offers a wide-ranging point of reference for Cymrophones' orthographic practices in adapting a range of English phonemes. Unfortunately, however, the survey was designed before my decision to add a number of phonemes to the standard list used by Wells (1982a), as discussed in Appendix A.i. As a result, the vowels of **TOWER** and **CURE** are absent from the survey data.

5.iii.a Survey respondents' language use and background

The full text of the survey can be found as Appendix E.i, with an English translation of the introduction and instructions as Appendix E.ii. Respondents were asked to provide demographic data before continuing to the loanword prompts. For all

questions, there was the option to respond with “Rather not answer”, but in the event, all respondents whose data were used in the final dataset chose to provide the requested demographic information.

Charts showing the demographic profile of respondents may be found in Appendix E.v. Fig. E.1 shows the number of respondents who associate their spoken Welsh with each of the dialect areas shown on the map of Wales, and the percentage of total respondents these numbers represent. The largest group of respondents speaks a variety associated with the north-western area of the country, but there are also a fair number of respondents who speak southern and south-western varieties. Overall, northern Wales is somewhat over-represented in the data, but the southern part of the country nonetheless provides nearly 40% of respondents.

From fig. E.2, we see that the majority of respondents are adults aged between 30 and 50. A third of respondents are aged between 30 and 39, and another third between 40 and 49. A further fifth are aged over 50. Teenagers and younger adults, aged up to 25, form the smallest proportion of respondents.

Respondents were asked about the geographical area with which they associate their Welsh, and their current age group, in order to control for any extreme skew in demographics, but these categories were not used in the data analysis itself. Geography and age are, of course, likely triggers of certain types of linguistic difference. D. Willis (2017), for example, used Geographically Weighted Regression to analyse the diffusion of dialect changes in spoken Welsh. However, geography is likely to correlate with features ascribed to dialect, more than with features traditionally noted as register markers, like formality and use of standard language, which are the direct focus of this study. The data analysis therefore makes direct use only of responses in which participants reported their use of or exposure to Welsh in a variety of contexts, both currently and in childhood.

The vast majority of respondents, nearly 80%, report acquiring Welsh from birth (fig. E.3), with fewer than 10% saying that they acquired the language as a teenager or adult. In the terms defined for this study, most of the survey respondents can be classified as primary-language speakers of Welsh. As might be expected, therefore, more than half of respondents received all of their compulsory education through the medium of Welsh (fig. E.4), and nearly 80% of respondents received at least some primary and some secondary Welsh-medium education.

More than half of respondents report that Welsh was the main language of their childhood home, while seventeen percent report using only English at home in their youth (fig. E.5). These proportions are mirrored in respondents’ current reported

home language: once more, just over half report using Welsh as their only language at home, while 14.5% use no Welsh (fig. E.5).

Finally, respondents were asked whether they used Welsh in five contexts (figs. E.7–E.11): on social media; in writing, informally (e.g. notes to friends, emails, or text messages); in writing, formally (e.g. essays or reports); in social settings (e.g. in the pub, chapel, or social clubs); and in professional or educational settings¹. Respondents could indicate that they used Welsh in each context regularly, occasionally, or not at all. In addition, the social media context allowed respondents to say that they had no social media accounts, or that they used social media in English or some other language but not in Welsh.

In informal contexts, over 70% of respondents report using Welsh regularly: 73.5% often write text messages, informal emails, or notes in Welsh (fig. E.10), and 71.1% often use spoken Welsh in social contexts (fig. E.8). Somewhat fewer respondents (63.9%) use Welsh frequently on social media (fig. E.7), a reduction which may, in part, be explained by the tendency towards linguistic homogeneity in online contexts. The same proportion of respondents report using spoken Welsh frequently in professional or academic settings (fig. E.9). The lowest proportion of “Often” answers is found for formal written contexts, in which fewer than half of respondents report using Welsh frequently (fig. E.11). As might be expected, the number of “Never” answers is also highest in the formal written context, representing nearly a third of responses. In contrast, only around ten percent of respondents report never using spoken Welsh in professional or educational settings. Only 3.6% of respondents, or three individuals, report never using spoken Welsh in social settings, and only one person reports never using Welsh for informal writing. Of the respondents who never use Welsh on social media, all have social media accounts and post in English or another language (fig. E.7).

Overall Welsh usage by respondents is summarized in fig E.12, by scoring each response to the five contexts of use and summing each respondent’s scores. Answering “Never” scores 0 for a given context, “Sometimes” scores 1, and “Often” scores 2. Over a third of respondents score a full ten points, showing that they report using Welsh often in all five contexts. Only one respondent scores 0, indicating that

¹ Respondents should have been offered a “does not apply” option for this last context, since respondents who are not currently in work or education would have no professional or academic context for which to answer. This limitation may have artificially inflated the number of ‘Never’ responses.

they never use Welsh in any of the contexts outlined for the questionnaire. Over eighty-five percent of respondents have a score of at least 5, meaning that they report using Welsh at least occasionally in at least three contexts. Differences between respondents, both in their education and home language, and in their reported use of Welsh, will later be used as variables in a Principal Components Analysis, helping to identify linguistic features of interest for further qualitative analysis.

5.iii.a.1 Survey loanword prompts

The survey was designed to elicit Welsh orthographic adaptations of seventy-five English words. Respondents were asked to adapt each prompt word orthographically in order to fit it into a Welsh context sentence. The seventy-five prompt words were chosen in order to represent each of the phonemes commonly used in PSEE. Where possible, examples were sought to illustrate each phoneme in word-initial position and in medial position. Word-initial phonemes which are potentially subject to treigladd in Welsh were, where possible, included with at least one context which should trigger some form of treigladd, and at least one which should not. The following examples show some of the prompts for words which have an initial /k/ in English:

9. cable Dwi 'di colli fy _____ i chargio'r iPhone. [*I've lost my _____ to charge the iPhone.*]

62. camouflage Mae angen _____ ar filwyr. [*Soldiers need _____.*]

70. karma Fydd _____ yn dal lan gyda hi rhyw ddydd! [*_____ will catch up with her some day!*]

Prompt 9 should trigger nasal treigladd, causing initial /k/ to be realized as [ŋ]. If treigladd were applied, one would therefore expect any Welsh orthographic adaptation of the word to begin with <ngh>. In the analysis, nonetheless, treigladd was only very inconsistently applied, and by very few respondents. Data about rates of treigladd were therefore not used in the quantitative analysis procedure.

I considered it possible that not only the phonology of the English prompt word, but also its orthography, might affect decisions about adaptation into Welsh. Prompts were therefore also selected with a subsidiary aim of representing a variety of English spelling conventions. For example, prompts 62 and 70 represent the two most common orthographic representations of /k/ in English: <c> and <k>. There were also prompts where /k/ is represented in the English orthography by <q> ("equal"), by

<ck> (“breadsticks”), and by <ch> (“technology”).

Finally, I also included some words which are the etymons of well-established loanwords in Welsh, such as “cardigan”, “goal”, “lounge”, “dozen”, and “onion”. Including these words would, I hoped, give some indication of the extent to which the existence of established loanwords affects variation in individual adaptation practices. For some of these, the form(s) found in *GPC* are either identical to the Standard English forms, or modified very little: “cardigan” and “tractor” appear unmodified, for example, while “zero” exhibits only the replacement of initial <z> with <s>, to create <sero>. Others, in their usual Welsh forms, are more substantially adapted: “onion” gives rise to a loanword with a variety of possible spellings, including <wynwyn> and <winion>, and “bizarre” is adapted as <bisâr> in the standard spelling given by *GPC*.

A list offering exhaustive coverage of all the considerations laid out above would have run to many hundreds of words. A set of prompts of such length might have made it difficult to find willing respondents for the survey, or to keep their interest once they had begun participating. I therefore limited the number of words to seventy-four, keeping the survey size manageable, providing good coverage of a range of English phonemes, and offering at least some examples of established loanwords, English orthographic variation, and contexts for *treigl*. The full list of items is given in Appendix [E.iii](#).

5.iii.a.2 Data collection and filtering

The survey was constructed using SurveyMonkey ([n.d.](#)), a digital tool for designing and implementing surveys and questionnaires. The service handles data collection and aggregation, and provides the responses in a spreadsheet. This allows the creation of a totally anonymous survey, since the participants need never contact the researcher directly. A link to the survey was published on social media, with a request that Welsh speakers pass the link on to their contacts. This yielded 101 completed surveys.

In total, eighteen sets of responses were totally discarded. Five respondents translated the prompt words into Welsh, rather than adapting the English words to Welsh orthography. Another thirteen surveys were submitted incomplete, containing responses for fewer than 15% of prompts. Of the remaining 83 respondents, ten had omitted to provide responses for either one or two of the prompts. These 83 responses were therefore classed as substantially complete and form the basis for the dataset.

Responses to individual prompt words were removed from the dataset in three general cases. Some respondents had evidently misread a prompt. The most commonly misread prompt was “untied”, which was interpreted as “united” by sixteen respondents². With some prompt words, particularly those which the context sentences required be interpreted as verbnouns, respondents made morphological adaptations, in addition to (or instead of) adapting the orthography. For example, twenty-three respondents adapted the prompt “overtaking” by replacing the final morpheme *-ing* with the Welsh verbnoun suffix *-io*, producing forms such as <ofertakio>. This is interesting evidence that unadapted English nominal *-ing* forms seems to be particularly hard to fit into a position where a verbnoun would be used in Welsh, whether because they are instinctively felt to be of the wrong category, because the replacement of *-ing* with a Welsh verbnoun morpheme is so natural to Cymrophones, or for some other reason. In this context, however, the form could not be considered a direct orthographic adaptation of the ELU provided and had to be discarded.

Finally, in cases where the prompt word is the etymon of an established loanword in Welsh, some participants responded with a form of the loanword which has been so adapted over time that it no longer matches the phonological form of its English etymon. Ten participants responded to the prompt “enjoy” with some variant of the Welsh verbnoun “joio”, for example. In all cases, therefore, responses were removed if the form offered as a response clearly did not represent the prompt word, whether because the prompt had been misinterpreted, or because adaptations other than the strictly orthographic were evident in the offered form. After these removals were completed, each of the 83 surveys still contained responses to the majority of the prompt words. The total of missing and removed responses came to 149, leaving 6,076 valid responses in the dataset.

² The accompanying context sentence, which translates as “Your laces are all ____.”, should have prompted a correct reading. That it did not may suggest that participants were not expecting to be guided by the meaning of the sentences, instead interpreting them purely as syntactic frames for the prompt words. It may also, or alternatively, suggest that some respondents were not paying much attention to the context sentences at all, simply adapting the prompt words out of context.

5.iv Orthographic adaptation in naturalistic data: building the Twitter corpus

5.iv.a Ethical considerations

ESRC guidance notes that “Information provided in forums, social media or spaces on the internet that are intentionally public would be considered ‘in the public domain’”, and its use should therefore not require the seeking of consent (ESRC, 2019). Nonetheless, as the same publication goes on to say, “the public nature of any communication or information on the internet or through social media should always be critically examined”. This section will therefore briefly make such an examination, as well as discussing other ethical issues which might arise from the collection of Twitter data for this project.

Townsend and Wallace (2016) have synthesized research on social media data ethics into a useful set of guidelines for researchers. In addition, Fiesler and Proferes (2018) offer the perspectives of social media users themselves on the use of their posts for academic research purposes. The following discussion, addressing how this study proactively sought to reduce various types of potential harm to participants, is indebted throughout to these two articles.

The final dataset of tweets features words from 840 Twitter³ accounts. Social media sites differ in the extent to which users might reasonably expect their posts to be seen by strangers. Unlike many sites, Twitter is built around public engagement between strangers, as well as between networks of friends, colleagues, and acquaintances. Accounts of celebrities, politicians, and organizations feature prominently on the platform, and are known to interact with non-celebrity accounts. The phenomenon of a tweet by a non-celebrity becoming unexpectedly popular, far beyond the usual reach of the user’s immediate group of followers, is common enough to have developed its own terminology: “going viral”. Unless a user chooses to make their tweets private, therefore, the structure of interaction on the site makes it unlikely that they are unaware that strangers may view their tweets (the methodology for this study specifically excluded the collection of data from protected

³ Although Twitter had undergone its transformation into X by the time this study was complete, the relevant data collection had all been completed before that time. The study will therefore continue to call the source of the data “Twitter”, and use the present tense to refer to it as a platform.

accounts).

This does not cover cases where a user chose to make their account private or delete one or more tweets after data collection was complete, however, and this is an unavoidable drawback and can only be mitigated by ensuring that the data are treated as sensitively as possible in other ways, for example by ensuring that the data are robustly anonymized. In this case, demographic data such as tweet location (where present) were discarded before storing the tweets, and usernames were converted into numerical pseudonyms. Profile information, photographs, and so on are not present, even in the raw data.

Tweets from the dataset will never be quoted with a username attached. I have chosen to attribute tweets from institutional or otherwise public Twitter accounts elsewhere in this study, but in all cases these are tweets found by browsing Twitter, rather than ones collected as part of the dataset, in which all usernames are anonymized. Nonetheless, there needs to be particular sensitivity regarding the quotation of tweets containing personal, sensitive, or identifying information, as they could ultimately be traced back to a specific user. Personal names, usernames, locations, and other sensitive information will therefore, where necessary, be changed in ways that leave the linguistic aspects of tweets intact. If a tweet cannot be anonymized or sensitive information cannot be removed without damaging the linguistic content, I will not quote that tweet at all.

5.iv.a.1 Access to the dataset

According to Twitter's terms of use, the dataset may not be shared with others, except as a list of tweet IDs (Twitter, 2019). These tweet IDs could, in theory, be used to reconstruct the dataset, although by the time this study is concluded it seems likely that academic access to Twitter will have been suspended, making reconstruction much harder. Although the replicability and accessibility of my research are, of course, a concern, I will nonetheless err on the side of caution and decline to make the dataset available even in the form of tweet IDs, since it was collected without informed consent from users.

5.iv.b Twitter corpus collection procedure

In 2013, around a third of tweets were written in the English language, with a further 16% in Japanese (Orcutt, 2013). These two languages therefore accounted for around half of Twitter posts. 12% of tweets were in Spanish and 8% in Malay. Arabic and

Portuguese tied for the position of fifth most-used language, with 6% of tweets apiece. French, Turkish, Thai, and Korean were represented in one or two percent of tweets each, leaving only 12% of tweets written in one of the world's hundreds of other languages (Orcutt, 2013). Majority languages with large numbers of speakers such as Hindi, Russian, and German are represented in this comparatively small slice of tweets, alongside minoritized languages such as Welsh, Cherokee, and Māori.

While linguists interested in the development of English in online spaces may be faced with the problem of tailoring their dataset to a manageable size, those interested in a language like Welsh are instead faced with the problem of finding sufficient examples of the language in use on the platform. Nonetheless, data from Twitter have been used successfully in studies of Welsh (R. J. Jones et al., 2013; D. Willis, 2020) and of Welsh English (Durham, 2022).

5.iv.b.1 Pilot study: setup

The main aims of the pilot study were to assess the feasibility of compiling a Welsh-language Twitter corpus, and to identify users of Welsh on Twitter whose tweets could be collected. At this stage, I was not concerned with finding official, organizational, or professional Welsh-language Twitter accounts. These are comparatively easy to identify. They also tend to have a stable language choice for tweeting: whether their tweets are exclusively in Welsh, or give an English translation alongside the Welsh, this practice tends to be consistent across all of the account's tweets. Personal accounts, on the other hand, are generally not subject to any official language policy. Even someone who tweets primarily in Welsh is likely also to tweet in English at times, for example when engaging with non-Cymrophone Twitter users. Conversely, some Cymrophones may tweet almost exclusively in English, but switch to Welsh in order to reply to Welsh-language tweets. Since I did not wish to exclude those who tweet infrequently in Welsh, it was not reasonable to select accounts by the language used in the Twitter bio (the user's self-description, displayed on the main page of their Twitter feed). Evidently, even those who tweet regularly in Welsh might choose to write their bio in English, in order not to exclude the huge audience of non-Cymrophone Twitter users. I therefore decided to try searching for frequent Welsh words across the entire Twitter archive, in hopes that it would be possible to extract from the results a list of accounts which had tweeted in Welsh.

At the start of this research project, the Twitter API's Academic Research product (Tornes & Trujillo, 2021) was not yet available. I therefore used the Python

library GetOldTweets3 to collect tweets (Mottl & Henrique, 2018), which allowed a user to perform a keyword search of the Twitter archive. In order to select useful search terms, I used a rough frequency list of Welsh generated from two online corpora of Welsh: the Bangor *Siarad* corpus (Deuchar et al., 2018), and Cronfa Electroneg o Gymraeg (CEG) (N. C. Ellis et al., n.d.)⁴ Since CEG is composed entirely of written texts and *Siarad* of transcripts of spoken language, using their data in combination provides some balance between these two linguistic modalities.

Many of the commonest Welsh words are homographs of words in one or more majority languages. The Welsh feminine personal pronoun {hi}, for example, is essentially unusable as a search term because the English greeting {hi} overwhelms it in results; Welsh {hon} (“this one” (fem.)) is confusable with the English endearment {hon} (short for “honey”); Welsh {yw} (a form of the verb {bod}, “to be”) is confusable with the English initialism {yw} (standing for “you’re welcome”). When combined with other common Welsh words, however, these terms are far less likely to return English-language search results. I therefore curated a list of frequent two-word, and occasionally three-word, Welsh sequences, listed here:

arall o'r, beth odd, beth oedd, dim ond, ddim ond, deud wrth, dweud wrth, fath ma, fath yma, fathau ma, fathe ma, gwneud hynny, gneud hynny, neud hynny, wneud hynny, hon yn, hon yw, hwn yn, hwn yw, math yma, mathau ma, mathe ma, mathau yma, mathe yma, nad odd, nad oedd, neb arall, nid chdi, nid chi, nid fe, nid fi, nid fo, nid hi, nid nhw, nid ni, nid ti, oddi wrth, ddi wrth, ond roedd, ond rodd, ond nid, pan oedd, pan odd, rhywun arall, roedd e'n, rodd e'n, roedd hi'n, rodd hi'n, sai wedi, sai di, smo fi wedi, smo fi di, so fi wedi, so fi di, so i di, so i wedi, sydd wedi, sy wedi, un arall, wrth fynd, wrth gefn

The terms include colloquial and standardized variants (e.g. standardized {oedd} and its colloquial counterpart {odd}) as well as some regional variants such as the second-person singular pronoun {chdi} (found mainly in the north-west of Wales (D. Willis, 2017)). These forms were included so that formal and standard language was not privileged in the data collection process.

⁴ The Bangor *Siarad* corpus is available under open licence at <http://bangortalk.org.uk/> and CEG can be freely downloaded for non-commercial use at <https://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en>.

A few of the seed terms produced very noisy output, containing few Welsh tweets relative to the number of non-Welsh tweets. The phrase {nid fi} (meaning “not me” in Welsh), for example, is also used in some varieties of English-lexified Creole, such as Jamaican Patois. The word {fi} in this context can mean “to” before an infinitive verb; {nid fi} is therefore a way to write “need to”:

- (10) limie upload a pic,u nid fi see mi likkle one
let me upload a pic, you need to see my little one

Search terms resulting in a high number of false hits were rejected, as there was no efficient way to retrieve the Welsh data from them. The data files retrieved using the remaining search terms were compiled into a single dataset.

5.iv.b.2 Cleaning the dataset

Before anonymizing the data, I manually checked the list of usernames and removed those that clearly belonged to corporate, organizational, or professional accounts (e.g. broadcasters, universities, charities, shops), as well as parody and political accounts. I produced a list of usernames whose tweets were automatically removed from the dataset using a Python script. The data was then anonymized: usernames were identified and replaced with unique numeric identifiers, again using a Python script.

Sorting the data according to how frequently a given username’s tweets were represented made it clear that many of the infrequent authors are not, in fact, writing genuine tweets in Welsh. In some cases they are tweeting coherent Welsh-language phrases; they are, however, not genuine or original productions of the tweet authors. For example, some of the most frequent results for usernames with five or fewer tweets included quotations from famous Welsh songs, and Welsh song titles with links to videos of musical performances. The sentence “Mae’r rhaglen hon yn cynnwys iaith gref a golygfeydd o natur rywiol” (*This programme contains strong language and scenes of a sexual nature*), which is posted as a warning before television shows with adult content, was also frequently tweeted. These and similar quotations and formulae account for hundreds of the tweets in the pilot study’s dataset. They are also fairly easy to locate and remove, by searching for frequently repeated strings in the dataset.

As a whole, I classify the tweets described in the preceding paragraphs as “non-original”, because although they may feature genuine examples of the Welsh language, some substantial portion of the tweet’s text is not the original production of the tweet author. In contrast to these genuine Welsh phrases, the dataset also features

numerous tweets apparently generated by auto-translation services such as Google Translate. I classify these as “pseudo-Welsh”, since they are generated by a translation algorithm rather than by a human with actual knowledge of the Welsh language. Pseudo-Welsh tweets are harder to identify automatically than non-original tweets; instead, they are best identified by a human reader, as they have extremely unusual (if not nonsensical) syntax and word choices, along with one or more of the features noted in the following outline.

Many such tweets state (explicitly or in a hashtag) that a translator has been used⁵:

- (11) Eich dim ond yn gerddwr araf wedyn!! Galwch darsi!! (Chi ddim yn gwybod im yn gwneud hyn i gyd ar google gyfieithu!)

You're just a slow walker then!! Call a taxi!! (You don't know I'm doing all this on google translate!)

[Your only a slow walker afterwards!! Call a taxi!! (You don't know im doing all this on translation google!)]

- (12) Da iawn, dwi ddim ond ffugio gyda chyfieithu google oherwydd im yng Nghymru

Well done, I'm just faking with Google Translate because I'm in Wales

[Well done, I'm just to fake with translating google because im in Wales]

- (13) Mae'r ddedfryd hon yn Gymraeg oherwydd digwyddiad chwaraeon heddiw. #GoogleTranslate

This sentence is in Welsh because of today's sporting event. #GoogleTranslate

[This verdict is Welsh because of a sporting occurrence today. #GoogleTranslate]

Pseudo-Welsh tweets are often marked by the use of apparently nonsensical words, which result from a translation algorithm making the wrong choice amongst English homographs:

- (14) Nid fi oedd yn gwneud unrhyw beth * ystlumod lashes *

⁵ In this and subsequent examples of auto-translated “Welsh” tweets, the first English translation given is an attempt to reconstruct the author’s intended meaning. The second, in square brackets, gives an English version that aims to convey the errors and infelicities of the pseudo-Welsh tweet.

*I wasn't doing anything * bats lashes **

[It wasn't me doing any thing * nocturnal-flying-mammals lashes *]

- (15) *nid fi yn deall eich iaith. Siarad mwynghlawdd neu nad ydynt yn trafferthu.*

I don't understand your language. Speak mine or don't bother.

[not me understanding your language. Speaking mineshaft or they are not bothering.]

Numerous pseudo-Welsh tweets are attempts to render idiomatic English formulae, and are particularly commonly addressed to national sporting teams before important fixtures:

- (16) *gallwch wneud hynny wales*

You can do it, Wales!

[you are able to do that thing wales]

- (17) *mynd yn mynd Cymru, gallwch wneud hynny!*

Go, go Wales, you can do it!

[to go going Wales, you are able to do that thing!]

- (18) *Yn dod ar fechgyn, gallwn wneud hynny.*

Come on, boys, we can do it.

[Coming onto boys, we are able to do that thing.]

In some other cases, auto-translate is used to insult or belittle the Welsh language:

- (19) *Mae hwn yn un iaith yn hyll-edrych*

This is one ugly-looking language

[This is one language is ugly-looking]

- (20) *fi dont adnabod sut mae hyn yn cael ei ystyried yn iaith dim ond cyfres o lythyrau taflu at ei gilydd ac yn cael eu hystyried geiriau*

I don't know how this is considered a language; just a series of letters thrown together and considered as words.

[me dont recognize how this is being considered a language only a series of letters throwing together and being considered words]

- (21) *yr iaith hon yn rhoi dyslecsia mi*

This language gives me dyslexia.

[this language giving dyslexia me]

- (22) yr iaith hon yn warth
This language is a disgrace
 [this language a disgrace]

Finally, strings of pseudo-Welsh are tweeted by spammers alongside links to product sites:

- (23) Dim ond y galon yn wahanol o ddyn i ddyn ...
<http://solidwooddiningroomtable.zdiningtablesforsale.info>
Only the heart differs from person to person ... http://solidwooddiningroomtable.diningtablesforsale.info
 [Only the heart different from man to man ...
<http://solidwooddiningroomtable.diningtablesforsale.info>]

It is unclear why these spam tweets use Welsh for their text content.

The volume of pseudo-Welsh data of all types tweeted by accounts with few tweets in the dataset meant that this section of the dataset would have required significant manual checking in order to extract the few genuine Welsh tweets. Therefore, accounts with fewer than six tweets were excluded from the final dataset and were not processed further. The remaining tweets, with non-original tweets already having been automatically identified and removed as far as possible, were manually checked for further non-original or pseudo-Welsh tweets, which were also removed. At this point, 2,470 authors remained in the dataset, and their usernames were stored for the next stage of data collection.

5.iv.b.3 Selecting a balanced list of Twitter accounts

By this stage in the project, Twitter had introduced its Academic Research API (Tornes & Trujillo, 2021). Successfully applying for a Research API account allowed me to access Twitter's full archive. The API has the significant advantage that it allows searching for tweets in a specific language. Tweets in Twitter's archive are tagged with a language code (or with "und" if the language cannot be detected), automatically assigned by a language detection algorithm ('@kurrik', 2013). Of course, relying on this algorithm risks missing some tweets whose language is misidentified. However, I tested the language tagging by requesting English and Welsh tweets from a selection of accounts, and found no obviously mis-labelled tweets.

To explore the enregisterment of orthographic features in ELUs, it was necessary to construct a dataset which could be expected to exhibit some variation on the basis of register. While it is clearly possible to treat tweets as a discrete register (Hodgkin, 2017), it seems likely that there are also sub-registers within this overarching category. I therefore hypothesized that tweets from the category I will label *official* could be identified as a separate register from those I will label *personal*. The former category includes accounts representing organizations, institutions, political entities, and businesses, as well as those which represent a single individual who is explicitly tweeting on a professional basis. A list of prospective personal accounts was constructed as part of the pilot programme described above. It was therefore necessary to collect a list of usernames representing official Twitter accounts, as a counterpart.

I identified a list of official accounts by searching Twitter bios for words such as “swyddogol” (*official*), “cyfrif Cymraeg” (*Welsh-language account*), “ysgol Gymraeg” (*Welsh school*), “siop” (*shop*), “cymdeithas” (*society*), “undeb” (*union*), and “plaid” (*political party*). For each account, Twitter’s algorithm would recommend related or similar accounts; I used these recommendations to identify further usernames. The institutions, organizations, and individuals whose accounts were identified during this process fit into one of the categories listed below. In lieu of usernames, which are suppressed to preserve anonymity, I have given a brief description of a few of the accounts identified in each category:

- Campaigning, Political, and Official Services
 - A LGBTQ+ campaigning organization
 - A member of the Senedd (Welsh parliament)
 - A city council
 - A hospital board
 - A police force
- Media and Publishing
 - A national (U.K.) news company
 - A publishing company
 - A journalist and television presenter
- Educational and Professional
 - A secondary school’s mathematics department

- A students' union at a Welsh university
- An examination board
- A company providing professional training for educators
- A professional organization for medical personnel
- Cultural and Commercial
 - A local Eisteddfod
 - A popular music magazine
 - A chapel
 - A children's television show
 - A record label
 - A choir
 - A literary podcast
 - A bookshop
 - A tourist board

As this list indicates, the *official* category of Twitter accounts covers a broad range of Twitter users. Some are prominent national institutions whose accounts have tens of thousands of followers on Twitter. Others are small local societies with tens of followers. What they have in common is that their tweets are made for professional or organizational purposes. Only around 1% of accounts in the official list belong to individuals; the vast majority represent organizations or institutions of some sort. The individual accounts included in the final official list were accounts with a Twitter bio explicitly identifying their professional role, and lacking any disclaimer asserting that it is a personal account (such as "Tweets are my own and do not reflect the views of my employer"). Accounts of this type tend to be very clearly used for specific professional purposes, for example for a politician to distribute campaign news, or for a television presenter to advertise details of their next appearance.

Having established a list of official usernames, I began testing the Twitter API with requests for the archived tweets of three accounts. It quickly became clear that some of the accounts in the official list, notably those for major media and news outlets, would return hundreds of thousands of tweets. This was significantly more data than could be expected from the average account, and would risk skewing the dataset if the entire archive of such an account were to be included. As a result, I modified the code to request only a selection of tweets for these larger accounts. If an account returned a high volume of data in response to the initial request, the code

switched to requesting tweets a month at a time, starting in the year 2007. Requests were limited to 500 tweets per month, choosing a random day of each month on which to begin the request, in order to avoid any bias towards month beginnings or ends. For most official accounts, however, all tweets were collected. The same code was used to collect tweets from the personal accounts, except that the code for large volumes of data was removed, since none of the personal accounts was anywhere near as prolific as the most prolific official accounts.

At this stage, the list of official accounts contained 380 usernames. The personal account list constructed during the pilot was significantly larger, and would need to be cut down in order to balance the two halves of the final corpus of tweets. I achieved this by collecting data from all the prospective personal accounts, and then pairing each official account with a personal account yielding a similar number of tweets. Having collected all the tweets from the personal account list using the procedure already described, I generated a list of personal account identifiers and the number of valid Welsh-language tweets collected from each.

At this stage, there were 380 accounts in the official list, with a total of 689,162 tweets, while the list of personal accounts contained over 2,400 individual authors. Fortunately, this meant that it was possible to find personal accounts which matched well to the official accounts in terms of the number of tweets they had yielded. The final selection of 380 personal accounts yielded 689,149 tweets in total. The match between individual accounts could not be exact: the most prolific official account yielded 49,789 tweets, while the most prolific personal account yielded only 27,746 tweets. At the other end of the frequency scale, there were thirty-two official accounts, but only eight personal accounts, yielding fewer than one hundred tweets each.

Some accounts (especially those in the official dataset) tend to repeat tweets with identical text, for example when advertising a product, making a public service announcement, or reporting a news story. Personal accounts occasionally repeat the same text numerous times, often when making political statements or repeating a campaign slogan. I debated retaining repeated tweets, since they are part of a genuine representation of the use of Welsh on Twitter. However, although the repetition of the tweets may possibly have made their contents more visible to their audience, I decided that this was not sufficient reason to include every repetition in this dataset, since it risked skewing the dataset towards particular types of language without good evidence that these repetitions were significantly more visible or salient for Twitter users. A python script filtered duplicated tweets, retaining only one copy of any

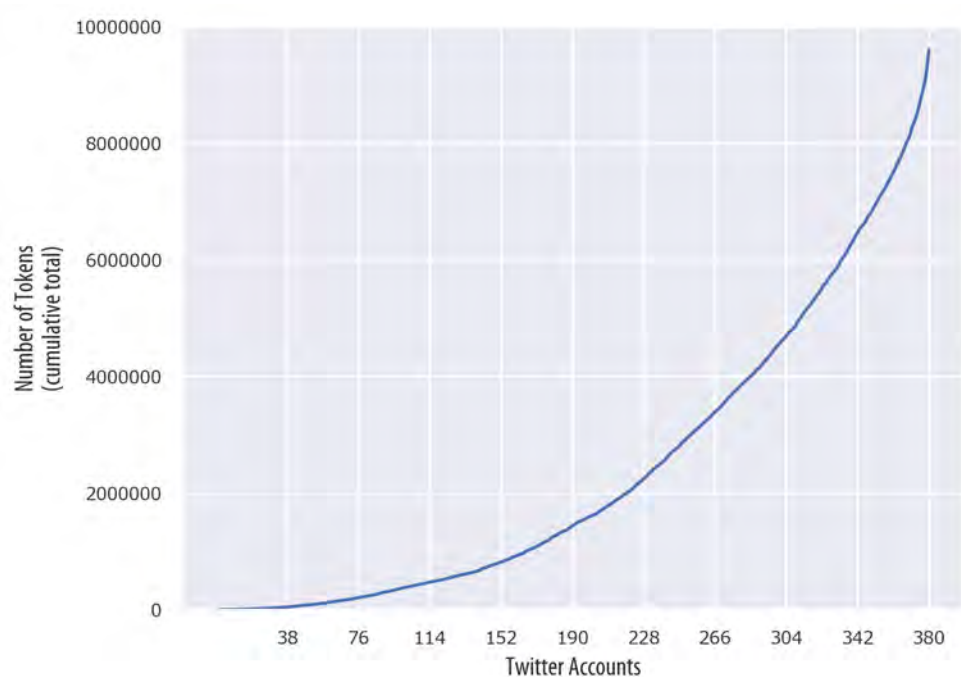
duplicated text. This did not include very short repeated texts, which are common on social media, including greetings, messages of congratulation, questions and answers, and responses expressing thanks.

The number of tweets per account provided an initial metric for balancing the contributors chosen for each of the two groups, so that the language of those who tweeted in Welsh frequently and those who tweeted infrequently was fairly equally represented. However, a naive tokenization of the tweets resulting from the collection process described above, splitting the text on whitespace to give a rough estimate of the number of words in each tweet, showed that official accounts tended to write longer messages on average. The average length of a personal tweet was around 14 words, as against 21 words for an average official tweet. This meant that the final corpus might be expected to have a word-count imbalance, with around 60% of the words coming from official accounts and only 40% from personal accounts.

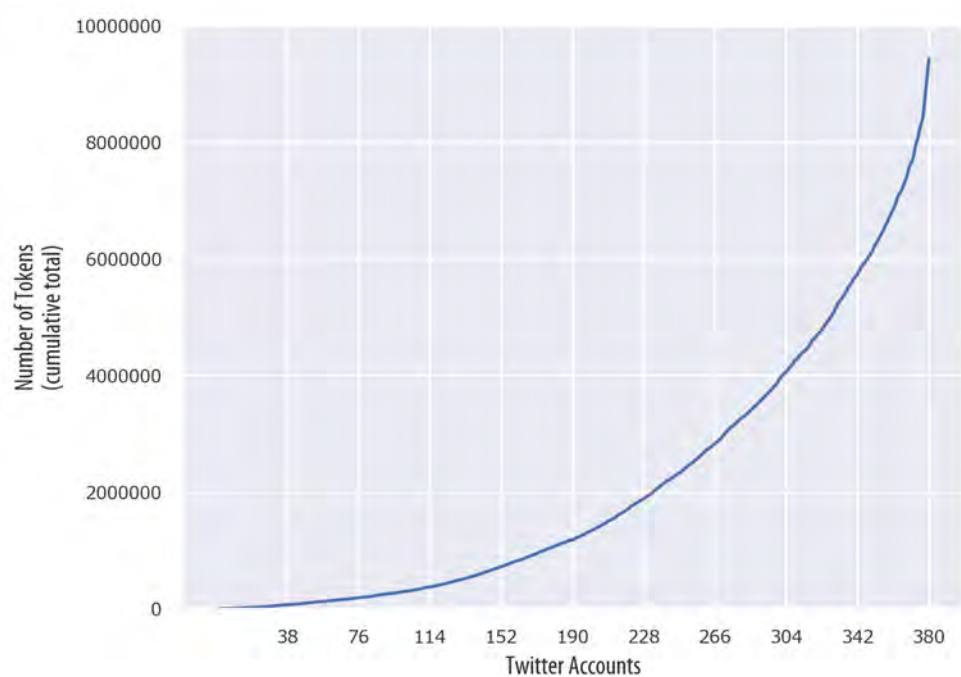
To finalize the corpus, therefore, tweets were filtered out from the official accounts to align their predicted token counts more closely to those of the personal accounts. This was achieved simply by removing every n^{th} tweet from each of the most prolific official accounts: every other tweet from the highest-frequency account; every third tweet from the next two hundred accounts; and every fourth tweet from the next one hundred. Data from the remaining (least prolific) accounts was not filtered. This procedure incidentally reduced the discrepancy between the most prolific official account and the most prolific personal account, which now accounted for roughly 313,000 and 383,000 tokens of the total, respectively. The graphs in [5.1a](#) and [5.1b](#) show the distributional similarity of the cumulative token counts achieved across the 380 accounts in each category, sorted from least to most prolific.

5.iv.b.4 Selecting lexemes

The Twitter dataset, at over 15.7 million words, represents a large corpus of modern Welsh. In order to analyze the use of loanwords, and of orthographic adaptation of ELUs, in detail, it was necessary to focus on a subsection of this corpus. In particular, the discussion to follow in [Chapter 6](#) makes use of detailed information about lexemes' listedness in seven different dictionaries. It also compares ELUs to semantically similar ILUs, inspired by Myers-Scotton's (1993) use of similar comparative measures, and notes whether the ELU has an obvious synonym in the IL. Compiling information about listedness, semantics, and synonymy for every ELU in the corpus would be a daunting task, and could not be accomplished within the



(a) Cumulative distribution of tokens across 380 accounts classed as “official”.



(b) Cumulative distribution of tokens across 380 accounts classed as “personal”.

Figure 5.1: Token distributions in official and personal Twitter subcorpora

confines of the current study. Instead, the analysis focuses on a subset of ELUs. The final lists of ELUs selected are given in Appendix F.iv.

The ELUs selected for close analysis, like the list of lexemes used as prompts for the survey of orthographic adaptation, were compiled with a number of aims in mind. As with the survey lexemes, the Twitter lexemes needed to represent the phonemic inventory of modern English well. They should also offer numerous sites for morphological adaptation and triglad. In addition, the lexemes should ideally be a mix of established loanwords and neologisms, and belong to a range of semantic categories. Initially, data from the CorCenCC corpus was analyzed to identify ELUs in modern Welsh and to categorize them by their frequency. Since frequency of use is used as a measure of the widespreadness of a lexeme by some scholars, it was important not simply to select the most frequent lexemes, but rather to aim for a range of different frequencies. The CorCenCC data offered a useful external source for frequency counts, so that the selected lexemes could be chosen without any prior knowledge of their distribution in this study's Twitter dataset.

Based on the ELUs represented in CorCenCC, I established six broad categories of ELU, and selected forty lexemes for each category. The categories were selected on semantic grounds, drawn from common categories of ELU in the CorCenCC data, and refined with a view to representing various kinds of both formal and informal language. The 'technical' category includes terms of art in science and technology, as well as the names of medicines and illnesses, and words relating to of academic and educational fields. The 'sociopolitical' lexemes cover geopolitics, economics, business, careers, and civil rights, and include the names of countries and political movements. In the 'media' category, lexemes relate to information technology, social media, and traditional broadcast media, as well as to the arts. The 'domestic' lexemes represent fields such as clothing, hygiene, food, housing, shopping, tourism, personal transport, and family. The 'colloquial' category encompasses lexemes which are frequently used for interpersonal talk and to manage discourse, including terms expressing emotional state, discourse markers, and taboo words.

In addition to these five semantic categories, an "uncategorized" label is used for lexemes that do not clearly belong to any of the other five categories, or which might belong to more than one category. The lexeme <power>, for example, might equally be called a technical or a sociopolitical lexeme, while <friend> might be used interpersonally (and thus be a 'colloquial' lexeme), but is also frequently used in discussions of social media (thus a 'media' lexeme).

Based on the lists of ELUs, I created lists of semantically- or pragmatically-similar Welsh ILUs. This was not always easy, and the lists of ILUs do not contain strong matches for all of the ELUs. On the whole, however, the lists of ILUs offer an approximate counterpart to the lists of ELUs. A few proper nouns and trade names are included in the list, since there is potential for these to behave somewhat differently from common nouns. The EL proper nouns were matched with IL proper nouns from similar fields, where possible; for example, the name of the political party UKIP is matched with the Welsh political party Plaid Cymru, and the music competition Eurovision is matched with the Welsh music festival Tafwyl. It was rather more difficult to match names of software products and social media companies. The Welsh spell-checking software Cysill was included as a counterpart to the software package Photoshop, and the colloquial terms {gweplyfr} and {gwyneblyfr} (both literally “face book”) were included as matches for Facebook. These are, evidently, not perfect matches. Alongside referring to the Adobe software package, {photoshop>} is now used as a generic term for altering an image, whereas {cysill} is unlikely to be used except as a direct reference to the Welsh-language spelling checker. Nonetheless, the inclusion of a specifically Welsh software package, like the inclusion of the names of a Welsh political party and television program, should enable a comparison of proper nouns in the two languages.

The final list of ELUs in each category was, in a few cases, expanded to more than forty, in order to ensure full coverage of unusual phoneme sequences. In a very few cases, in contrast, ELUs identified from CorCenCC were not present at all in the Twitter data, and these were therefore removed from the final list. The lists of ELUs in each semantic category therefore range from thirty-eight to forty-five lexemes, while the lists of ILUs range from thirty-two to forty lexemes.

Alongside assigning a semantic category, ELUs are also categorized according to whether they have a Welsh synonym (that is, a Welsh ILU with a closely related meaning). The coding for this measure was established by looking up each ELU in a number of bilingual dictionaries, as well as searching *GPC* for the use of the English word in question as a translation of an ILU. In cases where the dictionaries consistently pair the ELU with the same ILU as a translation equivalent, it is categorized as having a strong Welsh synonym. ELUs in this category include {freelance}, paired with the ILU {llawrydd}, and {stereotype}, paired with {ystyrydeb}. Where the ELU is used by the dictionaries to define numerous different ILUs, or where the ILU given is a multi-word unit or a periphrasis, the ELU is categorized as having a weak Welsh synonym. Examples include the ELU

{apparently}, which matches with the IL multi-word lexeme {yn ôl pob sôn}; and the ELU {operation}, which may be translated by the ILUs {gweithrediad}, {ymgyrch}, and {llawdriniaeth}, representing various different senses of the ELU. If no accepted modern translation equivalent which is an ILU can be found, as with the ELUs {smoothie} and {camera}, the lexeme is categorized as having no Welsh synonym.

Because the semantic and synonymy categories were established by a single researcher, they are not as robust as might be wished. Further research in this area would benefit enormously from having multiple data coders agree on semantic and synonym category labels for items. Although these category labels may not be entirely robust, however, we will see that there are promising correlations between them and other established measures used in previous research, such as listedness, corpus frequency, and morphological assimilation. Despite the methodological limitations of the categories as they have been established for this project, the results discussed here strongly suggest that attention to such aspects of lexemes can improve the depth and precision of research on code-switching and borrowing.

5.v Limitations in the data

The data collected from the three sources described above will have numerous limitations, particularly as regards the effectiveness of making a quantitative analysis. Some of these limitations are intrinsic to the nature of linguistic data, notably the inherent non-randomness of language in use and the tendency towards skewed distributions of features, which pose challenges for traditional statistical techniques (Dunning, 1993; Gries, 2005; Koplenig, 2019; Lijffijt et al., 2016). Other limitations are specific to the collection procedures adopted and the nature of the datasets themselves. I have attempted to acknowledge the major, dataset-specific limitations already, in the description of each dataset: the survey suffers from a small number of participants; the use of automatic procedures to identify ELUs of interest in the dictionary and in the Twitter data is subject to the fallibility of the algorithms involved; the small selection of loanwords as a focus in the Survey and Twitter data means that any conclusions may not generalize to all English-origin ELUs in Cymrophone discourse.

The collection of three parallel datasets is an attempt to compensate for some of these limitations: the dictionary data introduces a large set of different loanwords; the

Twitter data provides a large sample of language in use; the Survey data gives an insight into demographic variables. Each dataset also offers a different context for the production of ELUs. In *GPC*, loanwords are listed in a standardized form, and are selected because they have a recorded history in Cymrophone discourse. The Survey lexemes are elicited in response to an artificial, experimental task, which focuses respondents' attention on the process of loanword adaptation itself. The comparatively controlled situations of the *GPC* loanwords and the Survey loanwords offer control over some variables, yet we may expect rather different types of linguistic production from the lexicographers of *GPC* and the survey participants. The former are recording scholarly objects, whose orthography belongs to a tradition of loanword adaptation in Welsh. The lexicographers select an orthographic form (or forms) reflecting the usage of the word, but it must also be one which conforms to certain rules and conventions of Welsh orthography. (As mentioned above, *GPC* does not normally list variant spellings with any of the non-Welsh alphabetic graphemes <k>, <q>, <v>, or <z>.)

The Survey data, in particular, may be affected by the fact that human behaviour often changes when the individual knows she is being observed, or finds herself in an unfamiliar situation (Cohen et al., 2007). The Twitter data, in contrast, is produced in a naturalistic context by hundreds of users whose motivations are largely unknown at the stage of data collection. The only categorization which is attempted at this stage is the post-hoc labelling of Twitter accounts as either "official" or "personal". The analysis in the following chapter will therefore need to begin by establishing whether any features of these two subcorpora support treating them as discrete types of language, or whether the use of language by these Cymrophones on Twitter is either too uniform or too diffuse to identify regularities in the two user groups.

5.vi Data analysis procedure

The varied nature of the data to be analysed here, the methodological difficulties of robust statistical analysis of linguistic data, and the importance of context and subjectivity in the research questions themselves motivated the selection of a mixed-methods analytic approach, combining qualitative and quantitative perspectives (Segumpan & Abu Zahari, 2023). Using multiple methods of data collection to observe linguistic behaviour from more than one standpoint, ideally making use of qualitative data as well as quantitative, can increase the likelihood that

conclusions are not simply artefacts of the collection and analysis methods (Cohen et al., 2007). Accordingly, an exclusively quantitative approach seemed inadequate here. Instead, quantitative analyses of various aspects of the data will be used to provide large-scale overviews of the datasets, and will guide qualitative analysis by establishing sites of potential interest in the Twitter subcorpora. Since the quantitative results reported here are not used as the basis of inferences about use of language by Cymrophones in general, following Gries (2005) and Koplenig (2019) I therefore do not report measures of statistical significance (i.e. p-values).

Qualitative analysis will consist of exploring the areas highlighted by quantitative analysis in more detail, seeking to understand what lies behind the quantitative patterns. This will, in particular, involve seeking examples in the Twitter data of linguistic features identified as being of interest, and considering them in the context of the full texts to which they belong. Situating individual data points in context in this way will allow us to consider how the context of the message may affect even apparently inconsequential linguistic choices, and to explore the possibility that changes at the level of orthography may form an essential part of the connotative meaning communicated. This analysis will build on the theoretical discussion in chapter 2, challenging the traditional view of sub-morphemic linguistic units as non-meaningful and explaining the observed patterns in the data by proposing a model of semiosis which integrates even individual written characters as potential vehicles of meaning.

5.vi.a Quantitative methods

The quantitative methods employed will, at their simplest, include the use of tables, plots, and charts to visualize the distribution of various linguistic features in the datasets. Factor analysis (Van Haneghan, 2021) will also be used as a more complex quantitative approach to analyzing the data. This (and similar) methods can aid qualitative research by illuminating relationships and themes in categorical data. Factor analysis refers to a group of related methods including principal components analysis (PCA), which will be the method used in this study.

PCA involves taking a large set of observed variables in a dataset and reducing them to a smaller set of derived components, which capture underlying patterns in the interaction of the directly recorded variables (Van Haneghan, 2021). Instead of reducing the complexity of the data by restricting the analysis to a small subset of variables, PCA permits us to compress the variables by recognizing redundancy; that

is, when variables behave in similar ways, showing correlated patterns in certain contexts, a new component can be created which combines those variables. The components can then be used in subsequent analysis in place of the original variables, retaining a large proportion of the original data's variance, with the advantage of reduced dimensionality. Principal components analysis for this study was undertaken using the Python library *Prince* (Halford, [n.d.](#)).

PCA permitted me to draw together a wide range of variables, reducing them to a manageable set of derived components which still captured over 75% of the total variance in the data. The variables used as inputs included the following (variables for the Twitter corpus were generally repeated at the level of the subcorpora, as well):

- For Survey and Twitter lexemes: listedness, date of attestation (if listed), frequency in CorCenCC, frequency in Twitter/Survey dataset, number of unique users in Twitter/Survey dataset,
- For Twitter lexemes: rate of treiglad, rate of English inflection, rate of Welsh inflection, ILU synonyms, semantic category.
- For variants: listedness, date of attestation (if listed), frequency in CorCenCC, frequency in Twitter/Survey dataset, number of unique users in Twitter/Survey dataset.
- For phoneme-grapheme correspondences⁶: frequency in the GPC data, frequency in the Twitter and Survey datasets, number of lexemes in the GPC data, number of unique users in Twitter and Survey datasets, standardness in English and in Welsh (based on presence of the phoneme-grapheme pair in a variety of resources for teaching phonics and similar reading strategies).
- For grapheme-grapheme correspondences⁷: frequency in the GPC data, frequency in the Twitter and Survey datasets, number of lexemes in the GPC data, number of unique users in Twitter and Survey datasets, similarity between Standard English and adapted grapheme.

⁶ Including phoneme-Standard English grapheme and phoneme-adapted ELU grapheme

⁷ That is, correspondences between the Standard English grapheme used in spelling the English etymon and the grapheme used in spelling the ELU.

- For phonemes: type (vowel or consonant), features of articulation (vowel length, height, and frontness; consonant voicing, and place and manner of articulation)
- For graphemes: presence in the standard alphabets of modern Welsh and modern English, accent type.
- For Twitter users: frequency of use of emoji and hashtags, average length of tweet.
- For Survey users: age of acquisition, home use of Welsh in childhood and currently, amount of Welsh education, current usage of Welsh in formal/informal written/spoken and online contexts.

Evidently, these sources of information represent a complex web of variables, some of which apply to all three datasets, others to only one or two. There would be no clear and efficient way to include every variable directly in an analysis of the data. PCA instead allowed them to be considered together, calculating how each of the variables correlates with others, and drawing out patterns which I used to highlight individual lexemes with particular salience, to be the focus of qualitative analysis.

5.vii Analytic categories

5.vii.a Objects of analysis: lexemes, variants, graphemes, and phonemes

Most of the quantitative data which will be presented in this chapter is broken down at the level of the lexeme, the wordform, the phoneme, or the grapheme, terms which are defined in the Glossary which precedes the main text of this study. Lexemes are demarcated by curly braces. The standard form of the lexeme {colour} as a singular noun is the variant <colour> in Standard British English, and <color> in Standard U.S. English. In the case of the English-origin ELUs in this study, the lexeme is labelled using the standard British English spelling, even if a different standardized form has been established for the loanword in Welsh: the lexeme would be labelled {sugar} despite being an established loanword with a standard spelling {siwgr}. This is to enable a uniformity of representation across all of the lexemes, some of which have no established Welsh orthographic forms.

The lexemes listed for close analysis in the Twitter corpus in this study are not ascribed a specific part of speech. If the lexeme can have a given part of speech in English, examples from the corpus with that part of speech are included in the analysis. A word in use is considered to belong to a lexeme if it has acquired inflectional morphemes from Welsh or English, but not, on the whole, if it has acquired a derivational morpheme. The one exception comprises verbs which acquire a past or present participle morpheme, potentially turning them into adjectives or nouns. It is difficult to establish the part of speech of such words without analysing the context in which they are used, and constraints of time made this impossible. The words <depressed> and <depressing> are therefore considered to belong to the lexeme {depress} in this study, regardless of whether the suffixes are behaving as inflectional or as derivational morphemes in any specific case. For words which might be of more than one part of speech, inflectional morphemes appropriate to any of those parts of speech are acceptable. Corpus entries for a lexeme which can be construed as either noun or verb in English, or as a verbnoun in Welsh, may therefore include words with plural inflection, verbal inflection, or a Welsh verbnoun morpheme.

A written wordform is the exact form of the lexeme as it appears in the corpus, including any inflectional morphemes or treigladd, and is represented here between guillemets («»). In the Twitter data, variants are derived from wordforms by neutralizing inflections, treigladd, and incidentals of orthographic or typographic representation. This includes reducing multiple repeated letters to a single grapheme. The lexeme {psychic} might therefore have a variant, <seicig>, to which is ascribed all wordforms such as «SEICIGIAID», «seicigs», «ssssseicig», and so on, which can all be normalized to produce that variant. Reducing repetitions evidently entails some loss of information, but since the focus here is particularly on the expression of grapheme-phoneme relationships, I took the decision that representing both «ssssseicig» and «seicig» as the variant <seicig> was likely to be more useful than keeping them separate. In each case, the grapheme <s> is used as a counterpart to the phoneme /s/. In the context of electronic communication, I consider such multiple repetition as an emphatic typographic feature, similar to using capitalization or an italic typeface, rather than an orthographic choice, like that between using <s> and <ps> to represent the first phoneme of {psychic}.

In some cases it is not entirely clear how to normalize a form, for example when inflectional morphemes are to be removed. The wordform «manifestoes» might be normalized as <manifesto> or as <manifestoe>. Where there is ambiguity, decisions have been taken with a bias towards standard orthography, in order not to exaggerate

orthographic deviance in the dataset. The orthographic patterns exhibited in the wordform are used to guide a decision about which language's standards to follow. On the basis of standard English forms such as «mosquitoes» (<mosquito>), «tomatoes» (<tomato>), and «potatoes» (<potato>), therefore, «manifestoes» would be normalized to the singular variant <manifesto>. On the same basis, multiple repetition in a wordform whose standard orthography has a double grapheme is reduced to two graphemes for the variant, rather than one: «prosesssssss» would become <prosess>, following Standard English <process>.

5.vii.b Quantitative categories of metadata

5.vii.b.1 Listedness: background

As already discussed (1.iv.d), Deuchar et al. (2018) define listedness in terms of “the community dictionary” which, they claim, “may be expected to be a guide to listedness”. The identity of “the community dictionary” is not made clear in Deuchar et al. (2018). The monograph references two standard dictionaries of Welsh (*Geiriadur yr Academi* (1995) and *Geiriadur Prifysgol Cymru* (2002)) as well as the electronic dictionary suite *Cysgeir*, which may therefore be assumed to be the source of listedness evaluations. Elsewhere, Deuchar is somewhat more explicit about the criterion: “listedness is reflected in established dictionaries of the language, in particular the comprehensive *Geiriadur Prifysgol Cymru* (GPC) for Welsh. If an English-origin item is found in this Welsh dictionary we shall assume that it is a borrowing, whereas if it is absent we shall assume that it is a switch.” (Deuchar, 2005b). More recently, Deuchar (2020) asserts that an ELU used by Cymrophones may be called a borrowing “because it is listed as such in the authoritative Welsh dictionary *Geiriadur Prifysgol Cymru*”. *Geiriadur Prifysgol Cymru* is therefore apparently the “community dictionary” generally relied upon by Deuchar to establish listedness.

Nonetheless, there is still some room for uncertainty about what exactly is being measured under the name “listedness” in work on English-origin ELUs in Welsh. In particular, articles in this field sometimes fail to clarify whether there should be a single source of authority for listedness, or whether more than one dictionary might be consulted. In (Stammers & Deuchar, 2012, p. 631), for example, the criterion is described both as “whether or not a word is included in **the** Welsh dictionary”, and a few sentences later as whether the word is “listed in **a** Welsh dictionary” ((emphasis mine) Stammers & Deuchar, 2012, p. 631). The former description seems closer to

that of Deuchar (2005b, 2020), emphasizing the importance of *Geiriadur Prifysgol Cymru* as the source of listedness judgements, whereas the latter is closer to the criterion put forward by Deuchar and Stammers (2016), who put forward “the fact that it is not listed in any dictionary of Welsh” as the definition of an unlisted ELU.

Standard reference works are often popularly understood as works of unbiased authority, eliding the fact that they are the product of human decision-making in specific cultural contexts. The *Oxford English Dictionary* (OED), for example, has obvious biases towards the Anglo-American literary canon and the British English geographical core in all but very recent OED decisions on word inclusion. Technical and scientific terminology may spread directly from specialized wordlists into general dictionaries, even if it is rarely or never used in reality (Burchfield, 1973), in response to the perceived benefits of comprehensiveness of coverage, especially in prestigious fields such as technology and the sciences. On the other hand, lexicographers may respond to sociocultural pressures by defining categories of lexeme which are unacceptable for inclusion, regardless of their widespreadness in the linguistic community. In older dictionaries, or those aimed at children, for example, swearwords, slurs, and other taboo terms may be omitted (Burchfield, 1973). Standard lexicography may therefore tend to overrepresent the language of socially powerful groups and prestigious fields of knowledge, while underrepresenting that of marginalized people and contexts.

In addition to these general patterns of cultural bias, the lexicography of minoritized languages is often characterised by conservative, prescriptivist approaches to vocabulary inclusion (Paton, 1995). Rather than specifying unacceptable lexemes in terms of cultural sensitivity, however, this prescriptivism instead excludes lexemes based on their etymology, choosing to define ELUs as outside the scope of the minoritized language’s vocabulary.

Stammers and Deuchar (2012, p. 631) acknowledge some of the difficulties in relying on dictionary data to establish a word’s status in the linguistic community: “Even where a sufficiently comprehensive dictionary does exist, it will tend to represent the more formal, standardised variety” (Stammers & Deuchar, 2012, p. 631). A factor which emerges from this quotation, but to which little further attention is paid, is the necessity that the dictionary be “sufficiently comprehensive”. Quotations above from Deuchar and colleagues emphasize *GPC*’s status as “comprehensive” and “authoritative”, implying that this is a quality which suits it particularly well for establishing listedness.

As Burchfield (1975) notes, however, historical dictionaries generally aim for

something approaching completeness of coverage of even rare items, particularly those which are used by canonical authors, and of items which faded from use decades or centuries ago. While a dictionary such as *GPC* is invaluable for establishing the history of an ELU in use by Cymrophones, its very comprehensiveness means that it is likely to list forms which have fallen out of use, forms which have been superseded by newly-coined ILUs, and/or forms which were never widespread amongst Cymrophones.

GPC is a “community dictionary” for Cymrophones in the same sense that *OED* is a “community dictionary” for Anglophones. It would be misguided to use the mere fact of listedness in the *OED* to establish whether a word is part of the community lexicon of modern British Anglophones. Of course, a recently updated *OED* entry would allow one to establish whether evidence of current usage had been found or not, but the procedure established by Deuchar and colleagues uses only the presence or absence of the item in *GPC* as the criterion to establish listedness.

To address some of these concerns, the criteria used in this study for establishing listedness are inspired by the procedure described in Poplack et al.’s (1988) work on Canadian French, in which they analyse an English-origin ELU with a diachronic lens, consulting Canadian French reference works dating from 1880 to 1985. They also consider the question of the geographical spread of ELUs across Francophone communities, by consulting two standard dictionaries of European French. This procedure asks finer-grained questions of the dictionary data: it not only establishes whether or not language authorities consider the word to belong in the lexicon at the time of writing; it also seeks to establish a timeline for the appearance of ELUs in Canadian French, and to distinguish ELUs which belong to the Canadian French linguistic community from those which belong to the global Francophone linguistic community.

The emphasis on global varieties of language would be impossible in the context of the present study. The only major geographic variety of Welsh, other than that spoken in Wales itself, is that of Patagonia, for which there exists no authoritative dictionary. The major dataset collected here, from Twitter, does not presume that contributors come from a particular geographic area; nonetheless, it seems very likely that the majority of the data are produced by Cymrophones from Wales. The other aspects of Poplack et al.’s (1988) approach can easily be adapted to the Welsh context, however.

Since the focus of this study is on modern Welsh, I chose to use dictionaries published in the twentieth century as the source of listedness judgements. A word’s

listedness is scored from 0 to 6, according to how many of the dictionaries feature the word. For bilingual dictionaries, I searched for the ELUs both as lexemes in the Welsh–English section and as translations of the lexemes in the English–Welsh section. For the one English–Welsh dictionary consulted, *Geiriadur yr Academi* (1995), words were judged to be listed if they appear as translations of an English lexeme.

Rather than aiming for absolute comprehensiveness, I instead selected a set of six dictionaries of different sizes and with different intended audiences. Judging listedness as a scalar variable, and using a set of quite different dictionaries, should give a broader range of listedness scores. This will, I hypothesize, reveal more detailed patterns of word usage than the criterion of listedness in *GPC*.

The two most comprehensive dictionaries chosen are the English–Welsh *Geiriadur yr Academi* (1995) and the monolingual Welsh⁸ *Geiriadur Cymraeg Gomer* (2016). Two large bilingual dictionaries, *Geiriadur Mawr* (1983) and *Collins–Spurrell* (1991), are added, representing the kind of affordable, general-purpose dictionary which is perhaps more likely to be encountered in the school or home. The final two dictionaries are rather more specialized: *Geiriadur Termau* (1973) and *Modern Welsh Dictionary* (2007).

No published account of the number of lexemes in most of these dictionaries is available. Previous scholarly works on Welsh lexicography have tended to compare page counts instead of directly giving headword counts (e.g. Dodd, 1998). The printed first edition of *Geiriadur yr Academi* (1995) has some 17,000 pages, and that of *Geiriadur Cymraeg Gomer* (2016) just over 12,000. The Welsh–English portions of *Geiriadur Mawr* (1983), *Geiriadur Termau* (1973), *Modern Welsh Dictionary* (2007), and *Collins–Spurrell* (1991) contain, respectively, around 500 pages, 280 pages, 250 pages, and 175 pages.

Estimates of headword coverage are also given for each of the dictionaries by Heinz (2003), and these seem likely to be rather more useful than the page counts given above. According to Heinz’s calculations, *Geiriadur yr Academi* has over 100,000 headwords; *Geiriadur Mawr* around 46,000; *Collins–Spurrell* around 26,000; and *Modern Welsh Dictionary* perhaps 10,000. *Geiriadur Termau* is not included in Heinz’s study, but seems on visual inspection to be similar in scope to a dictionary for

⁸ Although *Geiriadur Cymraeg Gomer* does offer English translation equivalents for many of its lexemes, its primary content (lexemes and definition texts) is in Welsh.

which Heinz estimates around 37,000 headwords. *Geiriadur Cymraeg Gomer*, which was published after Heinz's study, claims 43,000 entries. (Since it is monolingual, this would put it at around twice the size of the larger bilingual dictionaries considered here, in terms of Welsh lexemes covered.)

Geiriadur Mawr and *Collins-Spurrell* represent two separate strands of twentieth-century bilingual Welsh–English lexicography. Spurrell (later Collins Spurrell) had been publishing and revising a bilingual Welsh–English–Welsh dictionary since the 1916 first edition edited by J. Bodvan Annwyl, while *Geiriadur Mawr* was an expansion of the *Geiriadur Newydd* of 1953, a collaboration between H. Meurig Evans and W. O. Thomas (Dodd, 1998). Both dictionaries may notionally be used by either Cymrophones or Anglophones, although Dodd (1998, p. 56) notes that the English–Welsh portion of *Geiriadur Mawr* is less comprehensive and less thorough than the Welsh–English portion, and that the dictionary as a whole seems to assume that its readers are “bilingual speakers of Welsh, wanting from the dictionary only confirmation of what they know in English”.

The two largest dictionaries complement each other in terms of the audience they assume. The apparatus of *Geiriadur yr Academi* is largely in English, and guides to features of Welsh pronunciation and morphology are offered to the reader. In *Geiriadur Cymraeg Gomer*, in contrast, the apparatus is entirely in Welsh, and the editors' foreword makes explicit their desire to create a suitable dictionary for adult Cymrophones, including those pursuing further and higher education through the medium of Welsh (*Geiriadur Cymraeg Gomer*, 2016, p.xiv).

The final pair of dictionaries both have a more specialized audience in mind, although in very different ways. *Geiriadur Termau* is a dictionary of terminology, intended particularly for use in Welsh-medium education (Andrews & Prys, 2016). As such, it fills in some of the gaps in works such as *Geiriadur Mawr* and *Collins-Spurrell*, whose coverage does not generally extend to the specialist vocabulary of most academic fields. At the other end of the scale, *Modern Welsh Dictionary* is, its Preface asserts, “intended for the growing numbers of learners of modern Welsh in all walks of life”. In Heinz's (2003) estimation, it is extremely limited, even given its intended use by learners. Whereas *Geiriadur Termau* was chosen to ensure coverage of less-common vocabulary, *Modern Welsh Dictionary* was therefore selected to ensure that any word achieving a full listedness score of 6 could be considered as quite central to the language, at least according to the judgement of lexicographers. A word included in both a dictionary for learners and a dictionary of terminology aimed at Cymrophones in Welsh-medium education may reasonably be

judged as part of the core vocabulary of modern Welsh.

Furthermore, even achieving a listedness score of 5 would require presence in at least one of the two more specialized dictionaries, indicating that the word is considered either a necessary item of Welsh technical vocabulary, or a central item for those acquiring the language. It is also possible that a word with a score of 5 is in both of the specialized dictionaries, of course. A word which is fairly well established may happen to be missing from one of the larger dictionaries as a result of differences of editorial approach, or of the date of compilation of the dictionary. It is an advantage that one of the larger dictionaries, *Geiriadur Cymraeg Gomer* (2016), is also the most recently compiled and published, since this offers the best chance that recently coined or adopted Welsh vocabulary will be listed in at least one dictionary.

5.vii.b.2 Listedness of variants

Although I will challenge the utility of *GPC* as the source of listedness ratings for lexemes, I hypothesize that it may prove to be a valuable source for evaluating the listedness of variants. The six dictionaries described above all present words in what they consider to be a standard orthographic form (or, very rarely, forms). *GPC*, in contrast, often records at least some competing orthographic forms, rather than insisting on a single standard headword orthography. As a result, a lexeme may appear in five or six of the listedness dictionaries, but most of the variants found in the Twitter and Survey data will appear in none. Usually, a lexeme with a listedness score greater than one has a single listed variant, which shares its listedness score. However, a few lexemes have more than one listed variant. The lexeme {screen} is listed in the form <sgrin> by *Geiriadur Mawr* and *Geiriadur Cymraeg Gomer*, and as <sgrîn> by the other four dictionaries. *Geiriadur Prifysgol Cymru* gives both variants, as well as <ysgrîn> and <ysgrin>. The same is true of some of the ILUs in the dataset: the lexeme {morthwyl} (“hammer”) appears as the variant <morthwyl> in all six of the dictionaries, three of which also list the variant <mwrthwl>. *GPC* lists both variants, along with <mwrthwyl>, <mwrthol>, and <myrthwyl>.

GPC is not fully descriptive, and the variants it lists are still standardized. Unlike the sets of variant forms given in the OED, which aim to record “all the ways in which a word has been written, over time and in different regions” (*OED terminology*, n.d.), *GPC*’s variants are generally limited to a small list featuring representatives of major patterns of orthographic representation. If a headword is particularly varied in orthographic expression, the list may have “&c.”. Very rarely, the entry itself may contain a section, headed “Amr.” (“Amrywiad(au)”: *variant(s)*)

listing a variant or variants along with examples of the form in use. It is rare to see variants listed in *GPC* which do not conform to Welsh orthographic norms, unless they are proper nouns. Headwords derived from “Zwingli”, “Zulu”, and “Kant” are the only ones listed with an initial <z> or <k>, and there are currently none with an initial <v>, <q>, or <x>. Variants with <sh> for [ʃ] and <tsh> for [tʃ] are generally included only when using the conventional <(t)si> or <(t)s> would cause confusion, such as when the phoneme begins a consonant cluster, as in <shrapnel>, or is in word-final position, as in <bitsh> (“bitch”).

On the whole, *GPC* is a fairly conservative guide to the orthographic variants of both ELUs and ILUs. It does not, for example, generally list common colloquial variant spellings, such as the northern Welsh substitution of <a> for <e> in representing the vowels of some unstressed syllables. In some cases, southern colloquial spellings are represented. The substitution of <ae> with <e> (in southern varieties) or <a> (in northern varieties) in representing final unstressed syllables, which is extremely common in modern Welsh, is rarely represented in *GPC*’s lexeme variants. In some cases, however, the medieval form of a word has <e> where modern standard form has <ae>, for example in <chware>, which is listed as a variant of <chwarae> (“to play”). As a result, modern southern colloquial forms may sometimes be listed, but only because their orthography happens to match that of much older variants⁹. Despite its general conservativeness, nonetheless, *GPC* is still a more comprehensive guide to the more frequently encountered orthographic variants of the lexemes in the dataset, and may therefore be of use in measuring the listedness of variants.

5.vii.b.3 Age of attestation, frequency, and widespreadness of lexemes and variants

For words listed in *GPC*, the date of earliest attestation date is taken from the examples listed in the dictionary. For words listed in other dictionaries, but not in *GPC*, the date of publication of the earliest dictionary is used as the date of attestation. All other words are treated as having no date of attestation.

Following Poplack et al. (1988) and others, I also chose to include information about how frequently used and how widespread the lexemes under consideration are

⁹ I am grateful to David Willis for his assistance in clarifying this aspect of *GPC*’s policy. Any errors in this section, as elsewhere, remain entirely my own.

in the Cymrophone language community. Accordingly, I searched for frequencies of both lexemes and variants in the CorCenCC corpus (adjusted as frequency per million words in the corpus as a whole). This gives an indication of frequency of use across spoken, electronic, and written contemporary Welsh, from a variety of genres and registers, and produced by numerous authors and speakers. The CorCenCC data also have the advantage of being independent of the datasets gathered for this study, so frequency counts from CorCenCC offer a useful comparison for the same words in the Twitter data.

Frequency is measured in the datasets themselves, as well as in CorCenCC. In the Twitter corpus, this permits measuring whether certain words, or certain types of word, are more frequently used by personal or official accounts, for example, and the frequencies can also be compared with those extracted from CorCenCC. Although frequency of use of lexemes is not relevant for the Survey data, the relative frequency of the specific variant forms produced may prove to be of interest. This, too, can be measured against frequency of use in CorCenCC, to explore questions such as whether words that are more widespread in the language community tend to have more stable orthographic forms. The number of individual users a lexeme or variant has is recorded, to supplement measures of frequency, so that a distinction is possible between a variant used twenty times, twice each by ten different users, and a variant used twenty times by a single user. The lexemes have the same overall frequency in each case, but the former is more widespread whereas the latter looks potentially as though it is an idiosyncrasy of the user in question.

5.vii.b.4 Standardness of orthographic adaptations

The data from GPC is, on the whole, used as a benchmark for the other two sources. As described above (5.ii), the ELUs collected from GPC were broken down into sets of grapheme-phoneme-grapheme triplets. This made it possible to extract statistical profiles of grapheme-phoneme pairs in ELU spellings, which indicate how standard a particular pairing is amongst established English-origin ELUs in Welsh. Comparisons of the grapheme-phoneme mappings in *GPC* with those found in the Survey and Twitter datasets was included amongst the variables used as input to the Principal Components Analysis, which identified items of interest for qualitative analysis.

Appendix C.iii contains charts illustrating the range of graphemes used to

represent each PSEE phoneme. Four charts are shown for each phoneme¹⁰, one for each dataset, showing the distribution of different graphemic representations of that phoneme. These charts show in detail the significant differences in grapheme–phoneme correspondences between phonemes and between datasets. Figure C.54, for example, shows the phoneme /h/ to be quite uniform across all four datasets, represented almost exclusively by <h>, although with some use of a zero graphetic variant <ipafont □>, particularly in *GPC*. Figure C.47, in contrast, shows the much greater variability of the affricate /d͡ʒ/. The grapheme «j», borrowed into the Welsh alphabet from English, representing the vast majority of instances of this phoneme in the Survey, and is the most commonly used grapheme in *GPC*. In the personal Twitter data, however, it is almost exactly as frequent as the graphetic representation <si>, conventionally used in Welsh to represent alveolar fricatives and affricates. In the official Twitter data, this more conventionally Welsh representation dominates, with <j> representing only 7.5% of instances.

Different adaptation strategies are characteristic of the different corpora, a fact which, in the following chapter, will be analysed to show the semiotic potential of orthographic adaptation. As the example of /d͡ʒ/ alone suggests, traditional loanword adaptation may offer numerous ways to represent English phonemes. Some of these have strong Welsh languageness, being scoligraphic to a monoglot Anglophone. Others are either conventions which happen to match English ones, or conventions directly drawn from English orthography. Of the latter group, <j> is the most prominent, as a grapheme only fairly recently accepted into the Welsh alphabet. It may therefore retain an English languageness which makes it less congenial in what I have hypothesized is a more formal register of Twitter texts. The next chapter will analyse the data in more detail, showing how orthographic adaptation intersects with other kinds of adaptation, as well as factors such as lexeme frequency, semantics, synonymy, and dictionary listedness. This will show that Welsh orthographic practices are enregistered with Welsh languageness, providing resources for meaning-making both through their representation of phonology and beyond it.

¹⁰ Except in cases where the Survey dataset is omitted because the phoneme in question is not represented in its data.

Chapter 6

“Ai hop ddat yew can ryd ddys”: Data analysis and discussion

Ai hop ddat yew can ryd ddys and ddat yt meics sens tw yew. Iff yew can ryd ddys, dden yew ar probybli a Welsh spicyr and ar reit at h m widd owar ffonetic langwey. In  l laiclihd, efriwyn els wil bi wyndyring wy ther’s no Gwgl Transl t byton, so yt’s basicyli a sictet langwey!¹

Twitter user 16 (personal subcorpus)

6.i Dataset descriptions

6.i.a Distributional patterns of variants and graphemes across datasets

Having established, in the previous chapter, the motivations for collecting three complementary datasets for this study, we can now move on to summarize the characteristics of those datasets, in preparation for analyzing the orthographic data they contain. This discussion will largely concentrate on analysis of variant forms of the lexemes, and of phoneme-grapheme correspondences. Before beginning the discussion of the individual datasets, it will be useful to have a broad comparative overview of these aspects of the datasets. Table 6.1 shows counts of unique headwords and unique variants in each dataset. Fig. 6.1 summarizes the number of

¹ *I hope that you can read this and that it makes sense to you. If you can read this, then you are probably a Welsh speaker and are right at home with our phonetic language. In all likelihood, everyone else will be wondering why there’s no Google Translate button, so it’s basically a secret language!*

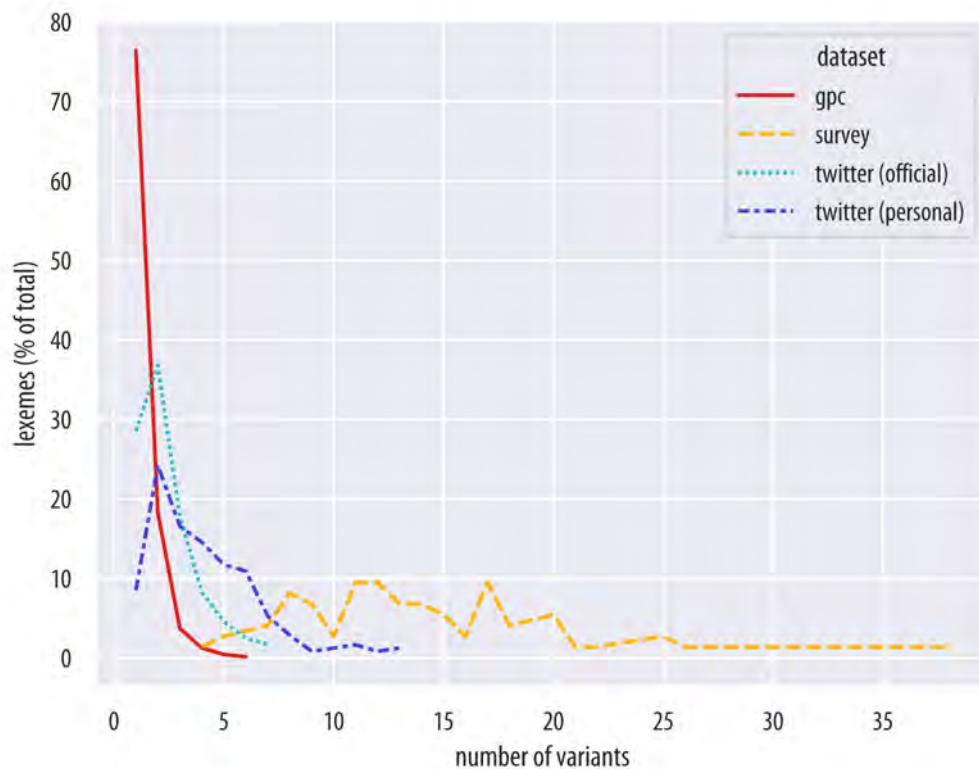


Figure 6.1: Proportion of total lexemes by number of variant forms recorded for those lexemes, separated by dataset.

variant forms recorded for each lexeme in a dataset, showing the proportion of total lexemes which have a given number of variants.

Table 6.1: Counts of unique lexemes and variants in each dataset

Dataset	Unique Lexemes	Unique Variants
Survey	74	1108
Twitter (official)	242	578
Twitter (personal)	248	1023
Twitter (all)	250	1089
GPC	6551	8558

The datasets differ markedly from one another, although the two Twitter subcorpora are rather more alike than any other pair of unrelated datasets. The *GPC* data strongly favour listing very few variant orthographic forms for lexemes, with over 76% of lexemes having only a single variant, and no lexemes having more than six variants. In total, nearly 95% of lexemes in the *GPC* dataset have only one or two variants.

The Twitter subcorpora both show a spike at two variants, although the proportion of lexemes with two variants is much higher for the official subcorpus, at nearly 37%, than for the personal, at just over 24%. The official subcorpus drops off sharply from this peak, with a limit of seven variants. The personal subcorpus, in contrast, has a gentler slope from the peak at two variant forms, trailing off to a limit of thirteen variants. Around 83% of the official subcorpus lexemes have between one and three variants. To reach the same coverage in the personal subcorpus requires combining lexemes with between one and six variants, representing just over 86% of lexemes. Only 49% of the lexemes in the personal subcorpus have between one and three variant forms.

Unlike the other datasets, the Survey dataset shows no strong peak in the number of variants. Instead, the proportion of lexemes fluctuates between around 1% and 9% for variant counts between four and twenty, before trailing off with variant counts between 21 and 38 having between zero and two percent of total lexemes. To reach 80% coverage of the Survey data, lexemes with between one and 21 variant forms must be included. Whereas *GPC* almost overwhelmingly favours listing a single orthographic form, and the Twitter users tend to produce a fairly small number of variant forms for the majority of lexemes, the Survey respondents produce a far more varied range of orthographic forms.

Similar patterns are reflected when comparing the number of orthographic variants produced for each individual phoneme in the four datasets. Figs. ?? and ?? each contain two charts. The first shows the number of orthographic variants produced for each phoneme in each dataset. The second records the standard deviation of the frequencies with which the orthographic variants are recorded, normalized as a percentage of the total samples for each phoneme. For example, a phoneme with four variants, each appearing 25 times, would have a standard deviation of zero: the mean of the four frequencies is 25, which is identical to the frequencies themselves. For a phoneme with four variants, one of which appears 97 times, with the other three appearing once each, the mean of the four values is still 25, but all of the actual recorded values differ markedly from that mean, so the

standard deviation is much higher at 41.57.

Comparing figs. C.1 and C.3 (Appendix C.i), it is clear that vowel phonemes on the whole tend to attract a greater number of orthographic variants than do consonant phonemes. Amongst the vowels, the unstressed **COMMA** and **LETTER** vowels are overall the most varied, reflecting the Welsh and WE tendency to produce phonemes which reflect the written forms here (as discussed in Chapter 4.i.d) rather than a default unstressed vowel.

There is also greater variability between datasets for the vowels than for the consonants. *GPC* has the most variant forms for most vowels, likely reflecting the fact that the *GPC* data contains a far greater range of lexemes, many of which reflect historical trends in pronunciation and orthography. For the consonants, nonetheless, the *GPC* data is less clearly dominant, with the Survey and the Twitter personal subcorpus showing the most orthographic variants for many phonemes. This probably reflects the fact that English and Welsh have the same or similar orthographic representations for many consonant phonemes, unlike vowel phonemes.

The Twitter official subcorpus tends to be the dataset showing the least variation for both vowel and consonant phonemes. This seems likely to reflect the relatively conservative nature of this subcorpus which, as we shall see, is characterized by a preference for standard linguistic forms. The Survey data, comparatively, shows far less variation for consonants than for vowels. Whereas the Survey is generally the most varied dataset, other than *GPC*, for the vowel phonemes, it is often the least varied dataset for the consonants.

The **NEAR~BEER** phoneme shows particularly broad variation between datasets. The survey records twenty-five different variants for this phoneme, whereas the Twitter subcorpora only have four or five. *GPC* sits roughly halfway between these extremes, with fourteen orthographic variants. It is worth noting that the lexemes in which the **NEAR~BEER** phoneme appears in the Twitter subcorpus, {queer} and {chandelier}, are both likely to have the **BEER** phoneme in those WE varieties which have a **NEAR~BEER** split. The lexemes in the Survey data, on the other hand, represent both **NEAR** ({earwig}, and often {cheer}), and **BEER** ({hemisphere}, and sometimes {cheer}). The greater number of orthographic variants in the Survey dataset may therefore partly result from the fact that some respondents were providing orthographic representations of what, for them, are two separate phonemes.

The orthographic sequences <iър> and <iър>, which are produced in the Survey data only in variants of {earwig}, may represent attempts to convey the [jœ:]

pronunciation of **NEAR**, with <i> representing the initial glide and a combination of accented vowel grapheme + <r> representing the open-mid [œ:]. Although Welsh <ŷ> would normally imply a vowel in the region of [i:], we shall see below that Welsh diacritics are often used in non-standard ways by the Survey participants, indicating features such as stress rather than the quality or quantity of the vowel itself. The accent may therefore be interpreted as an indication that this is a rising sequence of glide and vowel, with the vowel having greater prominence, rather than the falling diphthong /ɪə/ which has a more prominent first element. The choice of <ŷ> might therefore be understood as a combination of the accent as a stress marker and the Welsh grapheme <y> representing the Welsh **YN** vowel pronounced long ([ə:]), which is probably the closest phoneme to [œ:] in the Welsh vowel inventory. Similarly, <êr> can be interpreted as an combination of the Welsh diacritic, again used only for stress, and English orthographic rules, which allow <er> to be interpreted as the **NURSE** vowel /ɜ:/, again possibly the closest phoneme to [œ:], this time in the English vowel inventory.

Contrast these graphemic variants with the variant <îar> (recorded in the Survey for both cheer and earwig), which implies a more prominent first element, and thus either a PSEE-type [ɪə] pronunciation, or a WE-type pronunciation with the diphthong broken into two vowels with an intermediate glide: [ijə]. The latter type of pronunciation is also implied by the variants <ïar> (produced for cheer) and <ïyr> (for hemisphere). These show a standard use of the Welsh diaeresis, rather than the non-standard circumflex, to indicate that the two successive vowel graphemes should be pronounced separately, separated by a glide, and not as a diphthong.

Full visualizations of the graphemic representations of phonemes are given in the charts in Appendix C.iii. These show, in most cases, a fairly uniform pattern of behaviour across the four corpora. GPC often has the widest range of graphemic representations, particularly for vowels. This is not surprising, given that the number of lexemes represented in each case is much higher than in any other dataset. The Survey dataset, nonetheless, often shows a level of graphemic variation similar to that of GPC. The range of actual phonemes in use in the Survey often diverges from the phonemes represented in GPC, reflecting the non-standard and non-traditional adaptation strategies used by Survey participants. In most cases, the graphemic variants used in the Official Twitter subcorpus are a subset of those used in GPC, reflecting the Official accounts' preference for standard orthography, one of the features that characterizes these tweets as belonging to a more formal register than those of the Personal accounts.

Having made a brief overview of some of the comparative features of the four datasets, we may now turn to analysing them individually, beginning with GPC, which provides information about standard and traditional loanword adaptations.

6.i.b The GPC dataset

The *GPC* dataset consists of 6,811 unique English lexemes, with 8,557 total Welsh graphemic variants. As shown in table D.1 (Appendix D.i), around three-quarters of the English-origin lexemes are listed with a single variant form, and most of the remainder have only two variants listed. Given that *GPC* is one of the most comprehensive lexicographical sources of variant spellings available in Welsh, it seems reasonable to say that the authoritative account of the language shows a preference for representing ELUs with a limited range of orthographic forms.

The lexemes with the highest number of listed variants include {counsellor} (listed as <cownselar>, <cownseler>, <cownselor>, <cownsilor>, <cownsler>, and <cwnsler>) and {waiter} (listed as <gweitiwr>, <weitar>, <weiter>, <weitiwr>, <weityr> and <wetar>)². In the former, we see an example of the variability in representation of schwa between a consonant and a sonorant (discussed in section 4.i.d), as the sequence /s(ə)l/ is represented as <sel>, <sil>, and <sl>. It also illustrates the possibility that diphthongs which close on /u/ may be reduced to a high back rounded monophthong, represented by <w> in <cwnsler>.

Variants of the lexeme {waiter} illustrate a commonly conventionalized features of English-origin ELU orthography in Welsh: the use of <gw> for initial /w/. This pattern is frequently seen in ELUs. Some of these are originally derived from some variety of French, such as “gwarden” (*warden*, from Old French *wardein*, a variant of *guarden* OED2 (n.d.)), which may have had a pronunciation beginning with /gw/ when they were borrowed. Other ELUs which frequently show this feature, such as variants of the etymon {watch} (<gwats>, etc.), however, have no obvious historical source for the initial <g>. This feature may, therefore, have arisen by analogy with the French-derived ELUs, showing the influence of traditional orthographic

² *GPC* lists two separate headwords derived from {waiter}, treating those with final <wr> separately. This is presumably because “-wr” is productive as a suffix in Welsh, meaning that <gweitiwr> etc. might be analysed as derived from “wait” + “-wr”. Nonetheless, *GPC*’s etymology states simply “bnth. S. *waiter*” (*a borrowing of En. “waiter”*), and it seems likely that this is correct, with the final English morpheme “-er” of the borrowed lexeme being treated as though it were Welsh “-wr” because of the phonological similarity. I have therefore chosen to treat both types of spelling as examples of {waiter}.

adaptations on new ones, even when the trigger for a particular historical adaptation is missing in newer lexemes. The spelling with initial <g> might alternatively be the result of generalization from patterns of treiglad: Welsh ILUs with initial /w/ are rare, and initial [w] in an ILU's surface form is likely to be an example of soft treiglad from a word with underlying initial /gw/. This kind of reanalysis is common, particularly in earlier loanwords.

Table D.2 (Appendix D.i) shows how the phonemes in the *GPC* dataset are distributed, and how many unique graphemic sequences are used to represent each phoneme. The number of different English graphemes refers to the standard English orthographic forms of each lexeme, and the number of Welsh graphemes counts the various graphemic sequences which correspond to a given phoneme in any variant form listed in the dictionary. In each case, the pronunciation of the standard modern English form of the lexeme was used, even where the etymon had a strikingly different pronunciation when borrowed. This means that patterns of conventional representation of phonemes can be taken into account; for example, the use of the grapheme <cn> to represent /n/ is rarer than the use of <n>, but is conventionalized in ELUs such as <cnoc> (from {knock}) and <cnec> (“fart”, from {knack}). The older pronunciation with initial /k/ is reflected in the Welsh form, but a modern Cymrophone can only have direct experience of the English phoneme-grapheme pair /n/ - <kn>, without reference to the obsolete phonetic form.

Table D.2 illustrates the inevitable unevenness of the dataset: some phonemes are far more frequent than others, and this must be taken into account in considering how much graphemic variation they exhibit. Nonetheless, some points of interest may already be identified in these raw numbers, including the tendency for consonants to be more graphemically stable in Welsh orthography than they are in the original English etymons (/t/, for example, has thirteen different English graphemic forms, and only six in Welsh). The most notable exceptions are the alveolar affricates /tʃ/ and /dʒ/, with seven English variants apiece but twelve and fifteen, respectively, in Welsh. Indeed, the alveolar affricates appear as points of interest in the raw counts of all three datasets.

6.i.c The survey of orthographic adaptation

In sharp contrast to the *GPC* lexemes, most of which are represented by a single orthographic form in the dataset, all of the lexemes in the survey dataset have a minimum of four variants. Variant counts for the survey lexemes are given in

Appendix E.vi, Table E.1, and variant forms themselves are listed in Appendix E.iv. The least varied headword was {drink}, with four variants. Appropriately, the lexeme with the largest number of variants was {variation}, with thirty-eight variants.

The distribution here is obviously quite different from that found in the other datasets. The majority of lexemes in the *GPC* data have only one or two variants, as might be anticipated in a context which privileges linguistic standardization. Even on Twitter, nonetheless, there is a strong trend for headwords to be realized with a fairly small set of variants. This trend is stronger in texts produced by the official accounts, but is nonetheless still noticeable in the personal subcorpus. In the survey, in contrast, the number of variants shows far less of an obvious trend.

There are three factors which seem likely to have contributed to this difference between the datasets. Firstly, the lexemes selected for the survey are not necessarily ones that Cymrophones would ever choose to adapt, in a naturalistic context of use. The choice of lexemes was driven by numerous factors, as already described, which are unrelated to any measure of whether the lexeme is actually likely to be adopted as an ELU in normal Cymrophone discourse. Secondly, most of the survey's lexemes have little formal history of use in Welsh. Around 54% of the survey lexemes are listed neither in any of the reference dictionaries nor in *GPC*, whereas 73% of Twitter lexemes are in one or more dictionaries, and 79% are in *GPC*. Any adaptations users choose to make to these lexemes will therefore not be aided by any prior experience of standardized spellings.

The third possibility is that, as I have already suggested, the experimental context itself may have affected the orthographic choices made by the participants. The differences between the survey and the other datasets offer some initial support for the hypothesis that the survey's focus on adaptation as a goal in itself may have encouraged respondents to adopt more extreme patterns of adaptation than are normally found in Cymrophone language. The survey's request that respondents not refer to dictionaries or other linguistic references, and the explicit explanation that the goal is to gather personal responses, was designed to elicit idiosyncratic and creative orthographic forms.

In Appendix E.vi, Table E.2, the number of graphemic variants associated with each phoneme is given. Once again, there is quite some variability in how often each phoneme is found in the data, although this is not as extreme as the variability in the Twitter dataset, since each word is collected exactly once from each respondent. The most represented phoneme (/I/) is therefore around seventeen times as frequent as the least (/aə/). In the Twitter data, for reference, the most frequent phoneme (/l/) is

more than three thousand times as frequent as the least (/ɪə/).

Here, again, the postalveolar affricates emerge as sites of great variability, as do a number of the vowels. The distribution of grapheme variation is not identical to that found in the Twitter data. Notably, /ɔ/ is one of the least varied of the short vowels in the survey, and was the most varied in Twitter. The cluster of /w/, /ʊu/, and /jʊu/ was also a potentially interesting feature in the Twitter data, with the consonant /w/ showing significantly less variation than the vowels, both of which are highly varied. The same pattern is repeated here, with even less variation in graphemic representation of /w/.

There is, again, large amounts of variation in the representation of diphthongs. In the Twitter data, /eɪ/ and /aɪ/ are particularly varied in comparison with other diphthongs; the Survey data also shows much greater variability in /au/, as well as a modest increase in the variability of the other diphthongs, as compared to Twitter. The diphthongs closing on schwa are also much more varied here than in the Twitter data.

Of course, since the sets of lexemes represented in the Twitter and Survey datasets are not the same, it is possible that some, or even most, of the difference in phoneme-grapheme mappings can be explained by the difference in lexemes. The comparisons made between the Twitter subcorpora, which are both based on the same set of lexemes, will therefore always be more valid than those made between the Survey and the Twitter corpus. (The Principal Components Analysis, described below, attempted to account for this by including the lexeme as a variable in the process of identifying salient forms for further analysis.)

6.i.d The Twitter corpus

6.i.d.1 The Twitter subcorpora and register variation

As Biber and Egbert (2018, p. 9) point out, it is often not obvious how registers of online content can be identified. Traditional registers, based largely in analysis of published print media, are very often identified by external characteristics. These may be concrete, including the type of physical publication to which the text belongs (a book, or newspaper, or magazine, for example), or abstract, comprising explicit metadata about the publication, such as whether it is classified as fiction or non-fiction, as journalism or as academic research, and so on. Online texts, especially those which are authored by individuals rather than by institutions, often lack many or all such traditional indicators of register categorization, and Biber and Egbert

(2018, p. 9) argue that much of the publicly-accessible content on the web should be thought of as belonging to hybrid registers, combining multiple communicative purposes.

Despite their existence as digitally mediated samples of written language, many uses of language online share features which are otherwise characteristic of speech (McCulloch, 2019). Biber and Egbert (2018, p. 178) identify online interactive discussions, for example, as a type of “written dialogue”, whose popularity as a way for individuals to interact spontaneously in writing has been made possible by the development of the internet. The notably oral features of online discussions include limited planning and editing, and a preference for shorter words. Many of the keywords identified which characterized online discussions include typically oral discourse markers such as “haha”, “just”, “anyway”, and “yeah”, as well as characteristically digital abbreviated discourse markers like “tbh”, “lol”, “tho”, and “imo” (Biber & Egbert, 2018, p. 188).

Although they, too, share many oral features, social media posts form a more distinct and internet-specific register, rather than a hybrid of one or more of the register categories of traditional print media (Biber & Egbert, 2018, p. 4). Social media can also be further sub-categorized by using the platform that hosts the text as an indicator analogous to the publication-related indicators of register in traditional print media. Messages posted on Facebook, Twitter, Reddit, or Tumblr are created under quite different technological and cultural conditions; the sites’ affordances for authoring and publishing, content policies, and the audience an author can expect to reach, are different in each case. For Hodgkin (2017, p. 20), Twitter is “an institution largely constituted from the use of digital language”. From authoring original messages, to replying, retweeting, quoting, and even adding a “like” to a tweet, Twitter users perform a variety of speech acts by which not only Twitter’s vast corpus of text content, but also its relationship networks and information flow, are defined.

Tweets are strongly influenced by the technologies involved in their creation and reception. The way in which language circulates on Twitter is both socially driven and inherently digital (Hodgkin, 2017); the latter characteristic makes the discursive patterns of social media difficult for human users to predict and comprehend, because the scale at which Twitter’s content algorithms manipulate data is utterly divorced from the scale at which human relationship networks operate. These algorithms are also driven by priorities and imperatives which are not, or not always, aligned to those of the platform’s users. Although the original organization of

the timeline of tweets seen by a user was chronological (Zappavigna, 2012), Twitter later introduced a non-chronological algorithm (Pick, n.d.) which, it emerged, prioritized not only more popular tweets (Oremus, n.d.) but also conservative, extremist, unreliable, and revenue-generating content (Fiske, 2022; Huszára et al., 2022). Users who believe that they are being presented with content on the basis of features which have salience in human cultural terms, such as recency or communal approval of the message, risk being manipulated (Fiske, 2022), and may come to believe that the state of grassroots discourse in their communities is truly reflected by content which is deemed advantageous by a complex algorithm designed to optimize Twitter's commercial and political interests.

Tweets are used by individuals for socializing and informal discussion, but also by professionals from many fields as a way to network and engage in professional development (e.g. Forte et al., 2012; Madrigal et al., 2017). The platform is particularly popular amongst journalists and news organizations (Engesser & Humprecht, 2015; Revers, 2014). For some news organizations and other institutions, Twitter may simply be the equivalent of an automated news feed (Engesser & Humprecht, 2015), providing a way to disseminate headlines or links to content hosted elsewhere, rather than being used as an interactive tool for virtual conversation with other Twitter users. Nonetheless, many professionals and institutions use Twitter to establish online personae or brand identities through interactive tweeting. For some, their Twitter profile becomes a kind of digital portfolio (Ottovordemgentschenfelde, 2017), a way to curate and display a professional identity.

Individuals who have adopted Twitter in a professional capacity do not, in general, relinquish professional standards and adapt entirely to the informal norms of personal social media use (Enli & Simonsen, 2018; Olausson, 2017; Revers, 2014). Rather, they use the particular affordances offered by Twitter to build new professional conventions, for example using hashtags to separate editorial commentary from factual content (Revers, 2014); retweeting content as a means of engaging with a political constituency (Enli & Simonsen, 2018); or including apparently spontaneous photographs of day-to-day activities to create an image of transparency and honesty (Olausson, 2017).

This brief overview of some of the notable features of online communication, and of Twitter as a specific context for that communication, provide clear grounds for considering tweets as a discrete register of computer-mediated communication. There is also justification for the hypothesis that tweets from the two types of Twitter account selected for this study are likely to constitute distinct sub-registers, marked

by different linguistic patterns. Professionals and institutions generally use Twitter to build or promote a curated identity, to network, and for professional development. Institutions may use Twitter to interact with their audience, but these interactions are more likely to be governed by standards of professional conduct and to use generic templates for communication (Burton & Soboleva, 2011; Lutzky, 2024).

The constraints of using a written modality for what is essentially conversational discourse (McCulloch, 2019), as well as the extra pressure imposed by Twitter's limit on message length, make it desirable for users to find compact digital methods of expressing paralinguistic meaning, using features such as memes, slang, and humour as potential vehicles (Zappavigna, 2012). This chapter will focus primarily on loanwords and their orthographic adaptation, analyzing how patterns of spelling variation function as markers of meaning above and beyond the denotative meaning of the lexical tokens to which they belong.

This kind of meaningful orthographic variation is not exclusive to loanwords. Tatman (2015), for example, found that Twitter users used novel respellings to express phonetic dialect differences. Furthermore, Tatman suggests that, like spoken phonetic variation, this digital representation of accent may vary according to register. Eisenstein (2015) also studied the use of dialect orthography on Twitter, finding that respellings tended to acquire at least some of the register variables associated with the phonetic patterns they represent. Non-standard respellings were more common in informal tweets, and particularly in messages addressed to single individuals rather than to the general Twitter audience.

Given that message length is one of the most basic distinctive characteristics of tweets as a social medium, we will begin by looking at counts of tweet length across the Twitter corpus. Table F.1 (Appendix F.i) shows that tweets from official accounts are, on average, nearly six words longer than those from personal accounts. The main source of this difference is in original tweets, rather than quote-tweets or replies, as illustrated by Table F.2 (Appendix F.i). The original messages authored by the official accounts average over twenty words per message, whereas those authored by the personal accounts average only thirteen. The difference between the quote-tweets and replies in the two subcorpora is far smaller, at less than 1.5 words for both types of message. The tweets by official accounts differ significantly from those by personal accounts in both length and the distribution of message types. This provides initial support for suggesting that the use of Twitter by the two groups is rather different.

Examples of longer messages from the official accounts (exx. 24–25) suggest that the overall difference in tweet length may be related to the official accounts'

preference for highly informational tweets which convey information related to the organization's business or political aims. These examples avoid the abbreviations and slang which are characteristic of informal online texts (McCulloch, 2019). The language would not be out of place in a traditional printed medium such as an advertisement or a newspaper article. The combination of formal stylistic features with the structuring of information as a sequence of logically connected units, each represented by a new sentence, leads overall to longer texts.

- (24) Ydych chi eisiau gweithio i gyflogwr hyblyg gyda chyflog cystadleuol, cynllun pensiwn gwych a 30 diwrnod o wyliau blynyddol? Rydych yn y lle iawn – rydym yn chwilio am Swyddog Cyfathrebu i ymuno â'n Tîm Marchnata cyfeillgar, yn #EnwDinas. Gwnewch gais yma: <https://t.co/...>
Do you want to work for a flexible employer with a competitive salary, excellent pension scheme and 30 days of annual leave? You are in the right place – we are searching for a Communications Officer to join our friendly Marketing Team, in #CityName. Apply here: <https://t.co/...>
- (25) Mae'n #WythnosHinsawddCymru. Heddiw cafwyd ddigwyddiad i drafod pwysigrwydd ieuenctid Cymru ar yr #ArgyfwngHinsawdd. Plant a pobl Cymru – daliwch ati i ddefnyddio eich llais. 🌍 #WalesClimateWeek @LlywodraethCym @YCAWales @futuregencymru @sizeofwales <https://t.co/...>
It's #WalesClimateWeek. Today there was an event to discuss the importance of Welsh youth for the #ClimateEmergency. Children and young people of Wales – keep using your voice. #WalesClimateWeek @LlywodraethCym @YCAWales @futuregencymru @sizeofwales <https://t.co/...>

As already noted, professional Twitter users often make use of platform-specific features in ways which further their professional aims and purposes. Since emoji are a notable feature of social-media communication, an analysis of their use in the two subcorpora may offer further grounds for recognizing a difference in register. Emoji function in multiple ways, both to enhance and supplement the themes of the text and to express attitudes and pragmatic meanings (L. Gawne & McCulloch, 2019; Parkwell, 2019; Wagner et al., 2020), which means that they fulfil Zappavigna's (2012) criteria for a systemic-functional analysis of online register variables in terms of interpersonal functions. The semantics of emoji are flexible, and the meanings attached to an individual emoji may be conventionalized differently in different language communities (Wijeratnem et al., 2016). As a result, the interpretation of

emoji has become a particularly fraught matter in legal contexts, with decisions about cases of fraud, breach of contract, and defamation hinging on the meaning implied by the use of emoji (Kauffman, 2018).

Emoji are traditionally associated with informal texts (Parkwell, 2019), and used to imply pragmatic context which may otherwise be difficult to convey in text alone, such as sarcasm and humour (Kauffman, 2018; Parkwell, 2019). This being the case, we might expect to find fewer emoji in the official tweets than in the personal. Fig. F.5 is therefore somewhat surprising, since it shows that official accounts use significantly *more* emoji per tweet than personal accounts.

To throw light on this distribution of emoji, each emoji was scored based on how frequently it was used in each of the two subcorpora. Only emoji used at least ten times across both subcorpora, and by at least two users, were counted. For each emoji, the score in the Personal subcorpus was subtracted from the score in the Official subcorpus. The resulting number was divided by the overall frequency of the emoji. Emoji used much more frequently in the Official subcorpus would therefore have a score near 1, and emoji used much more frequently in the Personal subcorpus would have a score near -1. Those used equally often in both subcorpora would have a score near 0. The twenty emoji with the highest positive and negative scores are listed in Appendix ??, along with a list of the twenty emoji scoring closest to zero.

These lists suggest that official accounts and personal accounts use emoji very differently, and to perform different functions. The personal users' emoji are almost all facial expressions, implying the author's emotional state, ascribing a particular tone (such as humorous or playful) to the message they accompany, or performing interpersonal functions in relation to the addressee. The remaining emoji either represent gestures, with broadly the same range of functions as the facial expressions, or (in one case) represent an alcoholic drink, a symbol of casual social interaction. In contrast, the official accounts' most characteristic emoji feature no facial expressions. The hand gestures characteristic of official accounts are either deictic or celebratory (a pointing finger or clapping hands), in contrast to the interpersonal gestures of the personal subcorpus (a thumbs-up, or the "okay" gesture³).

The functions performed by the official accounts' emoji are those of guiding the user's attention (pointing fingers, arrow symbols, the warning triangle, and the

³ This gesture has, more recently, been used as a white supremacist symbol. This does not appear to be the meaning of any of the instances in the current corpus.

megaphone symbol), establishing communal identity (celebration, often paired with sporting symbols and/or the Welsh flag), and brand or organizational promotion (for example, media symbols of various kinds used to draw attention to organizations' publications and broadcasts). The fact that official accounts use emoji more commonly than personal accounts may be therefore be attributed to the desire to establish a consistent brand identity and a consistent manner of communicating with the audience. Whereas personal accounts use emoji on a per-message basis, to reflect immediate emotional states, to respond directly to others, and to indicate the tone of individual texts, official accounts often have a more deliberate and strategic approach, perhaps with a kind of "house style" which prescribes the use of particular emoji for every message which contains a certain type of content.

Further analysis of the corpus might well reveal more points of difference between the two sets of accounts identified here, but such analysis is beyond the scope of this study. The differences in message length, message type distribution, and emoji use established here, taken alongside prior scholarship which highlights the distinctiveness of professional Twitter use, provide sufficient grounds for treating the subcorpora as discrete sub-registers of the Twitter register.

6.i.d.2 Twitter lexemes: summary statistics

Table F.3 (Appendix F.i) shows that the official Twitter accounts tend to produce fewer variant forms of each lexeme than the personal accounts do. Across the entire Twitter dataset, most of the lexemes have at least two variants, however, which indicates that, in some cases, lexemes which show only one variant in a given dataset are realized with at least one different variant in the other dataset. The differences between the two subcorpora suggest that greater orthographic variation overall may be a feature of an informal register; this is predictable, as both orthographic inconsistency (in the form of typographic or spelling errors) and orthographic playfulness seem, on their face, more likely to be enregistered as informal.

For the lexemes used as the basis of this study, unlisted variants (those which have no entry in any of the dictionaries consulted, including *GPC*) are much less common in the official dataset than in the personal. Counting every instance of every variant form separately, 16.6% of all the variants in the personal subcorpus are unlisted, as compared with 3.9% of variants in the official subcorpus. If we count only variants of lexemes which have at least one listed form, and consider the number of unique variants which are unlisted in each subcorpus, we find that 31% of unique variants in the official corpus are unlisted variants of listed lexemes, compared with

56% of unique variants in the personal subcorpus.

Looking more closely at the unlisted variants of listed lexemes, one of the patterns emerging from the official subcorpus is that many of these variants differ from their listed counterparts only by a single grapheme. In some cases, they are northern dialect variants, such as «chwartar» (listed <chwarter>, from {quarter}) and «potal» (listed <potel>, from {bottle}). Many others differ only in the presence or absence of a diacritic, such as «smŵddi» (listed <smwddi>, from {smoothie}), «syffrajet» (listed <syffrajét>, from {suffragette}), and «wardrôb» (listed <wardrob>, from {wardrobe}). In the personal subcorpus, unlisted variants of this type are also present, but many more of the variants are quite different from listed forms of the same lexeme, like «opereshyn» (listed <operasiwn>, from {operation}), «iw craen» (listed <wcráin>, from {ukraine}), and «morgij» (listed <morgais> or <morgaets>, from {mortgage}).

The most varied lexemes include {emotional}, {watch}, {desperate}, and {actually}, with thirteen variants each in the complete dataset. In the official subcorpus, the most varied word is {exclusive}, with variants ranging from the Standard English spelling <exclusive> to adaptations such as <ecslwif> and <egsglwsif>. The most varied lexemes in the personal dataset are a subset of those in the full dataset: {watch}, {desperate}, and {actually}.

The distribution of phonemes in the Twitter data is represented in Table F.4 (Appendix F.i). As with the GPC data, there is significant variation in how frequently each of the phonemes is represented in the data. Despite that fact, a few obvious items of interest emerge here, as some phonemes exhibit far greater numbers of graphemic variants than others, and these are not simply the phonemes which are more frequent overall. The phoneme /ɪə/, for example, which is only represented 38 times in the data, nonetheless has five distinct Welsh graphemic counterparts. In contrast, /h/, which appears on 39,808 occasions, has only one graphemic form.

The postalveolar affricates once again exhibit large variability, as already predicted, reflecting the variability seen in the GPC data. In this case, the Welsh graphemic variability is much greater than can be accounted for by the variability in English graphemic representations of the phonemes.

Some of the vowels are also particularly varied. In some cases, the number of English graphemic counterparts is again quite large: the **FLEECE** vowel has thirteen different Welsh graphemic counterparts, and seven English. In others, the difference between the two languages is more extreme: **LOT~WANT** has twelve distinct Welsh graphemes, and only four English. On the whole, the diphthongs formed from a

front vowel plus a close-front second element show more variability than the others. The diphthongs showing more variability are ones whose mapping to graphemes in Welsh is unusually ambiguous, at least in southern varieties: the English **FACE**~**BAIT** vowel (corresponding closely to Welsh **EIN**~**CREU**) may be <ei> or <eu>, and the **PRICE** vowel (close to Welsh **SAIN**~**GWAUN**~**HAEN**) may be <ai>, <au>, or <ae>. If this were the only influence, nonetheless, we might expect to see a similar amount of variation in the representation of the **CHOICE** vowel (Welsh **TROI**~**POEN**), which can be <oi>, <ou>, or <oe> in (southern) Welsh. For these vowels, Welsh and English both use graphemic representations formed of a vowel grapheme plus <i>, although Welsh **SAIN**~**GWAUN**~**HAEN** is <ai> where English **PRICE** is more likely to be <ei>, and Welsh **EIN**~**CREU** has <ei> where English **FACE**~**BAIT** is more likely to have <ai>.

It is also interesting to note that some of the diphthongs showing less variability are particularly graphemically different across the two languages: **GOAT**~**KNOW** and **MOUTH** both tend to correspond to English graphemic sequences including <o> (such as <o>, <ow>, <oa>, <oe>, and <oh> for the former, and <ou>, <ow>, <ough> for the latter). In Welsh, the corresponding vowels of **BYWYD** and **GLAW** are written <yw> and <aw>.

As the patterns of variant distribution shown above predict, a larger number of graphemic variants is usually used in the personal subcorpus than in the official for most of the phonemes. The one exception is /d͡ʒ/, which has 10 Welsh graphemic variants in the official data and 9 in the personal data. The difference is not large, and may simply reflect the more general variability of /d͡ʒ/ in Welsh ELUs, but it is nonetheless another point in favour of closer analysis of the postalveolar affricates.

In contrast with *GPC* data, there are no phonemes here which have more English graphemic variants than Welsh, either overall or within the individual subcorpora. There are some for which the numbers are equal, nonetheless, leading to the possibility (which will be explored later) that the Welsh variation is conditioned by the English. Phonemes with equal numbers of English and Welsh graphemic variants are more frequent in the official data than in the personal.

6.ii Quantitative analysis of diacritic use

Chapter 4 established that orthographic adaptation is not always merely a matter of representing the lexeme's phonology, but may also be used as a vehicle for creative and humorous expression, as well as to indicate cultural Welshness. The

non-Cymrophone reader is not easily able to interpret Welsh orthography as a representation of English phonology; instead, it is a source of obscurity or ridiculousness in its deviation from an Anglophone standard. The Cymrophone reader, on the other hand, can bring the resources of either language, or both, to bear in reading a word. The orthography can therefore be at once a transparent window through which phonological meaning is seen, and an opaque screen upon which cultural meaning is directly displayed.

We have already seen how certain graphemes are interpreted as having particularly strong semiotic Welshness from the outside perspective: vintage postcards focus on the graphemes <w>, <y>, <dd>, <ch>, <ll>, and <ff>, in particular, to create a visual impact on the Anglophone reader. The discussion above hints that <iw> may also be salient in this regard; it is the only grapheme which distinguishes the English “Portugal” from the Welsh “Portiwgal”, which was a source of humour for an official BBC twitter account. These graphemes may identify an orthographic string as semiotically Welsh without any need that the string be otherwise comprehensible to the reader. As Durham’s (2022) study of Welsh-English dialect respelling showed, nonetheless, what is enregistered as a sign of languageness can differ markedly for those inside and outside a language community. The graphemic signals of Welshness for non-Cymrophones are largely those which are foreign to Anglophone convention, particularly the use of <w> and <y> to represent vowels. For a Cymrophone, however, salience is based on experience of Welsh orthography as part of a functioning linguistic system, not as a visual shibboleth.

The comparative frequency of certain graphemes and grapheme sequences offers one obvious locus of difference between English and Welsh orthography. The presence in English of graphemes which are absent from Welsh is another. A third is the use of diacritics to mark vowel graphemes. This practice is rare in English, and almost always marks a word as an ELU. The role of diacritic marks in standard Welsh orthography has been outlined above (section 4.i.a). Diacritics have a role to play in ensuring that the phonological form of ELUs can be retrieved from the orthography, but, unlike in English, diacritics are also regularly used to write ILUs.

The charts in fig. C.5 show how often characters using different types of diacritics correspond to seven classes of vowel sound, in each dataset collected for this study. It is important to remember that the survey and Twitter data use a restricted set of lexemes, and that this is likely to affect which diacritics are used and which vowels they represent. Whereas the GPC data has only one example of any pair of lexeme and variant, the survey and Twitter datasets may contain multiple copies of

the same pair, and represent fewer different lexemes overall. Nonetheless, some interesting patterns emerge from the data which suggest that diacritics are not always used simply as direct representations of phonological features, or in response to standard Cymrophone spelling practices.

The diaeresis is the least frequent diacritic in the standardized ELU spellings in *GPC*, appearing only sixteen times, and is correspondingly rare in the other datasets. The role of this diacritic is very specific, applying to a restricted phonological environment. It is therefore comparatively rarely seen in ILUs, and may not be salient enough to be a target of enregisterment. The grave accent is also quite rare in *GPC* (33 occurrences). As the chart shows, it is used mainly for short vowels. The few examples of long vowels are representations of lexemes with the **BATH** vowel, such as {bath} and {glass}, which are realized with short vowels in some varieties of Welsh English, a realization which has been carried over into the Welsh ELU pronunciation.

The grave accent is proportionally far more frequent in the survey responses than the diaeresis, whereas it is even less frequent than the diaeresis in the Twitter corpus. For survey respondents, the grave accent seems to be used largely to mark the schwa vowel, as in the wordforms «armÿr» ({armour}), «orijinÿl» ({original}), and «ôrffÿn» ({orphan}). It marks the **STRUT** vowel in addition to the **COMMA** and **LETTER** vowels, in words such as «rÿff» ({rough}) and «ÿddyrwais» ({otherwise}), suggesting that it is marking the mid-central vowel quality rather than lack of stress. For one user, this extends to the long mid-central **NURSE** vowel in the words «ÿrth» ({earth}) and «conosÿr» ({connoisseur}). Whereas the standard use of the grave accent is to mark shortness, therefore, it appears to have become associated for many survey participants specifically with the schwa vowel.

The acute accent is used with the same frequency as the grave accent in the survey data, but is comparatively more frequent in the other datasets. The *GPC* data, as always, reflects Welsh orthographic standards, using the acute accent to mark short vowels, or the first element in a diphthong, in a stressed final syllable. A few examples of long vowels marked with an acute accent are also present; as with the grave accent, these are generally phonemes which are long in PSEE but realized short in the Welsh pronunciation of the ELU. Often, the phoneme is an r-linking vowel which is realized in Welsh with a short vowel + [r], as in «absÿrd» ({absurd}), «distsiárs» ({discharge}), and «repórt» ({report}).

In the survey data, there are twenty-two examples of the acute accent. Fifteen of these come from a single user, and these will be discussed separately below. The

remaining seven suggest the same association with the **NURSE** and **STRUT** vowel seen for the grave accent, with three examples of «conosýr» ({connoisseur}), and one of «sôdýst» ({sawdust}). The other uses show no particular pattern, and only one is used on an anomalously stressed syllable, in «résyfwar» ({reservoir}). The survey data do not, on the whole, suggest that the standard role of the acute accent in Welsh orthography is salient for Cymrophones when adapting ELUs.

In the Twitter data, the acute accent is found more frequently in the official dataset than in the personal. Almost all of the variants, in both datasets, are forms which are listed in one or more dictionaries, or in *GPC*: «swffragét» and «syffrajét» ({suffragette}), «carafán» ({caravan}), «wcráin» ({ukraine}), and «olréit» ({alright}). Of these, only the first three lexemes appear in the official data. The only other use of the acute accent in the official data is in the form «wcráin» ({ukraine}), which is likely to be an error for «wcráin» rather than a creative respelling. In the personal subcorpus, three unlisted variants appear with acute accents: «ffésbwc» ({facebook}), «plís» ({please}), and «eniwé» ({anyway}), all three are either discourse markers or brand names, and belong the groups of lexemes highlighted as particularly interesting by the quantitative procedure described later in this chapter. The variants come from three different user accounts. In each case, the diacritic marks a long vowel or narrow diphthong, rather than a short vowel or the first element in a closing diphthong (as is its standard use). The scantness of examples of the acute accent outside established variant spellings confirms the non-salience of this diacritic for Cymrophones as anything other than a symbol prescribed in certain conventional spellings.

The final diacritic is the most frequently used, and the most varied: the circumflex. The patterns of use of the circumflex in *GPC* largely reflect the conventions established in *Orgraff* (1987). Examples which are anomalously matched to a PSEE short vowel generally represent one of two categories. The first includes forms whose pronunciation at the time they were borrowed was different from the PSEE form, such as «parasôl» ({parasol}) which had the pronunciation /pærə'səʊl/ in the 1800s (*OED3*, [n.d.](#), s.v. parasol). The second includes forms with either **KIT** or **FOOT** in PSEE, usually with an attestation date before 1700, for which the Welsh spelling with a circumflex is usually only one of a number of variants listed by *GPC*. This category includes «cŵc» ({cook}, also listed as «cwc» and «cwg»), and «ysgîl» ({skill}, also listed as «sgîl», «ysgîl», and «sgîl»). Apart from these forms, the use of the circumflex in *GPC* is overwhelmingly to mark long vowels or diphthongs, particularly closing and narrow diphthongs.

The circumflex has a more varied use in the survey, and is employed much more

frequently by survey respondents than its frequency in *GPC* would predict. Of 6,600 variant forms in *GPC*, only 455 feature a circumflex accent, representing just under 6.9% of the total. In the survey, 1,108 of 5,980 variants feature at least one circumflex accent, making up 18.5% of the total. The most anomalous uses of the circumflex, to represent unstressed vowels, is also comparatively rare, and occurs in one of three general circumstances. The first comes before sonorants, where PSEE often reduces the vowel to zero. Examples include «eicwâl» ({equal}), «orffŷn» ({orphan}), and «câbêl» ({cable}). The second context includes word-final examples of the **LETTER** and **COMMA** phonemes, for example «oprâ» ({opera}), «carmâ» ({karma}), «boîlêr» ({boiler}), and «armŷr» ({armour}). The third context comes in the representations of unstressed central syllables in trisyllabic lexemes, such as «camwfflaj» ({camouflage}), «conîsŷr» ({connoisseur}), and «enfilôp» ({envelope}). In both these cases, the circumflex may be interpreted as highlighting the Welsh-English avoidance of producing a zero-phoneme or schwa in unstressed contexts.

The circumflex is used in representing short vowels slightly less frequently than in representing unstressed vowels. The most common use of the circumflex for a short vowel is in the lexeme {goodness}, which has six examples of the variant «gŵdnes», and one each of «gŵdnis» and «gŵdnys». These forms are part of a pattern of using the circumflex on short vowels in stressed syllables, found also in forms such as «brêdstics» ({breadsticks}), «êniow» ({anyhow}), and «tecnôloji» ({technology}). In each of these cases, the addition of the circumflex forces a less-faithful phonological interpretation of the orthography by implying the presence of a long vowel. The same is true when the circumflex is used on <y> or <w> in a monosyllable with the **STRUT** vowel, such as «rŷff» or «rŵff» ({rough}), which can only be interpreted as /rɪɪf/ and /rɯɪf/, respectively. A circumflex is also used on four occasions for the short first vowel of {bizarre}, twice in the variant «bîsâr», and once each for «bîsar» and «bîzar». Here, the syllable is not stressed, and the purpose of the circumflex is even less clear. It may reflect a non-standard pronunciation of this lexeme by the respondents in question, or possibly an attempt to ensure that the orthographic form is understood to represent {bizarre} and not its near-homophone {bazaar}. If the latter is the case, the circumflex might be understood to indicate not length per se, but rather the specific quality of an unstressed vowel. Whereas graphemes without diacritics may be interpreted using English phoneme-to-grapheme mappings, those with a circumflex are more obviously Welsh, and this may serve to steer the reader away from the ambiguous English schwa in unstressed syllables.

In all these cases, it seems that the survey respondents have internalized the standard use of diacritics to indicate stress and to disambiguate vowel length in monosyllables, but have chosen a diacritic which, given its standard interpretation, is at odds with the lexeme's PSEE phonology. It may be that the familiarity of the circumflex as the most frequently used accent in Welsh orthography prompts Cymrophones to choose it as a default in unfamiliar linguistic situations, such as when attempting to adapt ELUs which they might normally never use in Cymrophone contexts. This suggests that the circumflex may be enregistered as the generic "Welsh diacritic" for some Cymrophones, gesturing towards some phonological difference from the norm, rather than specifying a fixed phonological property such as length, or simply imbuing wordforms with semiotic languageness.

Survey respondents use the circumflex much more often on centering diphthongs and, particularly, on long monophthongs than they do on short and unstressed vowels. These categories are far more prominent in the survey data than in either *GPC* or the Twitter corpus, and give an insight into some of the distinctive Welsh-English pronunciations highlighted by Durham (2022). One of these is the **NEAR** vowel, most characteristically represented in the homophony, or near homophony, of "ear" and "year" in some varieties of Welsh English (Durham, 2022; Mayr, 2010; Walters, 2001). Durham found this feature to be particularly salient for Welsh-English speakers, although it was not well recognized by speakers of other varieties. The survey respondents who deployed a circumflex in their variant of the lexeme {earwig} always did so in representing the initial phoneme, PSEE /Iə/. In many cases, the circumflex appears to be used to clarify the syllabification of this phoneme: forms such as «îyrwig» and «îarwig» point towards pronunciations in the region of ['ij,ərwig] or ['ij,erwig]. In these cases, the circumflex appears to be playing the role something like that usually played by the diaeresis in standard Welsh orthography. In some cases, the circumflex is instead placed on the second element representing the **NEAR** vowel: «iêrwig» and «iÿrwig», for example, which suggest pronunciations roughly like ['jɛ:rwig] or ['jɜ:rwig]. These representations seem closer to the distinctive Welsh-English /jœ:/. In a rather different use of the circumflex, this lexeme is also realized with the variants «êrwig» and «îrwyg», which suggest a specifically Cymrophone pronunciation, with the PSEE centering diphthong /Iə/ realized as a monophthong plus /r/: ['i:rwig] or ['e:rwig].

The relatively greater frequency of circumflexes in representing centering diphthongs and long monophthongs in the survey data, compared with the ELUs in *GPC*, is largely a result of the survey respondents' comparatively high faithfulness to

the phonological form of the English lexeme. *GPC* contains numerous loanwords whose pronunciation has been adapted to Welsh conventions over time; their orthographic form is therefore often a representation of a phonologically adapted ELU, not of the ELU with its English phonology. As a result, phonemes such as the diphthongs of **PRICE** and **SQUARE** are frequently encountered in established Welsh loanwords as monophthongs, for example <pris> (/pri:s/, from {price}) and <sgwâr> (/sgwa:r/, from {square}). These forms often do not require a circumflex accent according to standard orthographic rules. PSEE long monophthongs, meanwhile, are all r-linking vowels, and these are regularly adapted to become short vowels plus /r/ in established loanwords, such as <sbort> (/sbɔrt/, from {sport}) and <car> (/kær/, from {car}). These short vowels never trigger the use of a circumflex.

The survey respondents, in contrast, most frequently represent a fairly faithful PSEE or Welsh-English pronunciation, rather than one which has been phonologically adapted to conform to Welsh conventions. The representation of PSEE diphthongs as long monophthongs, a common feature of Welsh-English varieties, is a frequent target for circumflexes in the survey data. A few examples include {tourist} (e.g. «tôrisd» and «tôrist»); {aircraft} (e.g. «êcraffd» and «êcrafft»); {envelope} (e.g. «enfylôp» and «onfelôp»); and {major} (e.g. «mêjy» and «mêsior»). In these examples, the circumflex is performing its standard role as a marker of length. As these examples show, even when the orthographic form retains an <r>, suggesting a rhotic pronunciation, the vowel is often explicitly marked as long, as in «êcrafft». The form «êcraffd» without the <r> implies a strongly WE pronunciation, associated particularly with the Cardiff area of the **SQUARE~LAIR** vowel as [e:] (B. Collins & Mees, 1990). The form with the <r>, in contrast, implies a rhotic pronunciation, which may either reflect a WE accent or the influence of Welsh phonology.

Lexemes which are well established as loanwords in Welsh are more likely, on the whole, to be represented by variants which imply a conventionalized pronunciation. In representations of {hemisphere}, for example, the final vowel is represented with <êr> 42 times and with <er> 54 times, matching the pronunciation of the established loanword <hemisffer> (/hemisfer/) rather than the PSEE /hemisfiə/ or WE /hemisfjə/. This lexeme is also one of the few in the dataset which prompted the use of the diaeresis, in the form «hemisffiyr», where it makes clear that this is to be read as a Welsh-English pronunciation.

Turning to the Twitter data, we see patterns which are not straightforwardly like either the survey data or the *GPC* data. Almost 80% of circumflexes in the personal subcorpus, and over 60% in the official, are used on one of the narrow

diphthongs (the **GOOSE**~**FLEW** or **FLEECE** vowels). The closing diphthongs are very rarely marked with a circumflex in the official subcorpus, but the centering diphthongs are represented with a circumflex much more frequently here than in any other dataset. When we look closely at the data, however, this is revealed to be an effect of the high frequency of two lexemes, {square} and {power}, established ELUs in Welsh whose usual orthographic forms are <sgwâr> and <pŵer>.

To draw together the data about accent use, we will take a closer look at a survey respondent whose responses represented fifteen of the twenty-two uses of the acute accent in the dataset. This respondent regularly uses the acute accent to mark anomalous stress (i.e. stress on a syllable other than the penult), as in «syfáif» ({survive}), «enjói» ({enjoy}), «cámoſfflâzh» ({camouflage}), and «orújunyl» ({original}).

The respondent has above-average scores on the questions asking how often Welsh is used in professional and social situations, in informal writing, and online. The only usage category for which their score is very slightly below average is formal written usage. Interestingly, compared with the survey average, the respondent was also far less likely to produce variants which are listed in one of the six dictionaries or in *GPC*. This raises the tantalizing possibility that, like skilful code-switching, performative orthographic mixing may also reflect a high degree of competence in both languages, a subject which seems likely to reward further research. Of course, a single data point such as this must be interpreted with caution, but this user's behaviour supports the argument that performative orthographic adaptations of ELUs are a type of bilingual behaviour associated with the deliberate performance of linguistic Welshness. This respondent's use of the acute accent implies familiarity with standard Welsh orthographic practice, combined with a willingness to stretch the standards in order to inscribe Welshness on the wordforms they produce.

6.iii Quantitative analysis of lexemes and variants

6.iii.a Listedness, semantics and synonymy

The lexemes chosen for the Twitter subcorpus have all been assigned to a broad semantic category, and assigned a status denoting whether they have a clear Welsh-origin synonym. It is therefore possible to compare the distribution of these categories across the subcorpora, and to examine how the semantics and synonym

status interact with other features, such as listedness. Table F.5 shows the counts of individual lexemes in each category for each language of origin, and table F.6 shows how the English-origin lexemes are distributed by synonym status and semantic category.

The charts in figure ?? allow us to compare the distribution of English-origin ELUs and Welsh-origin ILUs according to their semantic categories. Firstly, we can see that the personal accounts use the English-origin ELUs selected for this study slightly more often overall than the official accounts do. Nonetheless, the Welsh-origin lexemes are used overall nearly twice as often as the English-origin lexemes by both types of accounts. Although the difference between the two subcorpora on this measure is comparatively small, the two types of account have a very different distribution of semantic categories. Personal accounts use colloquial words far more often than the official accounts do, reflecting the register differences already established between the two subcorpora. This category includes swearwords as well as colloquial discourse markers such as {okay} and {yn ôl pob sôn} (*apparently*, lit. “according to every report”), and terms of evaluation such as {lovely} and {penigamp} (*awesome*).

The official subcorpus contains comparatively more sociopolitical and technical lexemes than the personal, regardless of the language from which the lexeme is taken. The difference is more marked for the English-origin ELUs, however. These two categories make up around a third of the overall ELUs used by official accounts, but only around one-tenth of those used by personal accounts. Technical and sociopolitical lexemes account for around 15% of the Welsh-origin lexemes used in the personal subcorpus, and just over 20% of the lexemes in the official subcorpus.

Having seen how the semantic category of the lexemes in this study correlates with the official and personal Twitter registers, it will also be useful to consider how semantic category and listedness interact. Figure F.1 shows, for each language of origin, the proportion of lexemes in each semantic category which are listed in *GPC*. As we have already seen in examining the data for *treigl* and inflection, using *GPC* listedness alone does not provide a particularly informative view of the data. Nonetheless, it is notable that the English-origin colloquial lexemes are much less likely to be listed than the other lexemes. For the Welsh-origin lexemes, meanwhile, it is the media category which is least likely to be covered by *GPC*, perhaps reflecting how comparatively fast-moving lexical development is in this area.

A more detailed insight into the semantic categories is provided by looking at listedness across a range of dictionaries, as shown in fig. F.2. One of the most

noticeable differences between these charts and those showing listedness in *GPC* is that 80% of colloquial English-origin lexemes are totally unlisted in the six selected dictionaries, whereas over 50% are listed in *GPC*.

These more detailed listedness scores are informative for Welsh-origin lexemes, too. Whereas most of this study's semantic categories have very few lexemes which are not in *GPC*, their listedness in other dictionaries varies. Most categories have few totally unlisted lexemes, but there is substantial variation in the distribution of the other listedness scores. Less than 15% of the domestic lexemes are listed in four or fewer dictionaries, compared to between 35% and 45% of colloquial, technical, and media lexemes. The media lexemes are the most likely to be unlisted, which mirrors their listedness status in *GPC*.

Turning to the synonym status of the English-origin lexemes, fig. F.4a shows how frequently lexemes are used in the two subcorpora, according to whether they have a strong Welsh-origin synonym, a weak Welsh-origin synonym, or no Welsh-origin synonym at all. The personal accounts are somewhat more likely to use English-origin words for which a Welsh-origin synonym is available, and, in particular, to choose an English-origin lexeme if it has only a weak Welsh synonym. Considering listedness in relation to synonym status, we see in fig. F.4b that words which have no clear Welsh-origin synonym are far more likely to be listed in dictionaries than those with either a weak or a strong Welsh-origin synonym. The latter two categories are roughly equally likely to be totally unlisted, but lexemes with a strong synonym are generally listed in fewer dictionaries than those with only a weak synonym.

6.iii.a.1 Discussion

The role of synonym status in register is less clear than that of semantic category. Further research would be required to make anything other than a very tentative claim that informal registers are less sensitive to synonym status in the use of ELUs than formal registers. Nonetheless, the data here suggests that this may be the case, and that informal registers may, in particular, be more willing to accept ELUs when there is no strong Welsh synonym available, whereas official registers seek to use ILUs even if this means using periphrastic forms or phrases instead of a single-word ELU.

For example, the Welsh-origin lexical unit {yn y bôn} (*basically, fundamentally*, lit. ‘at the root’) is used 17 times in the official dataset, while the English-origin {basically} appears only twice. In the personal data, in contrast, {yn y bôn} is used

104 times, and {basically} 148 times. A similar pattern is seen for the near-equivalents {yn ôl pob sôn} and {apparently}: they are used 7 times and twice, respectively, in the official data, and 74 and 142 times in the personal data. These examples suggest that further research on synonymy might illuminate the status of ELUs in different registers of Welsh.

Differences in the distribution of colloquial words are consistent with a characterization of the two subcorpora as separate registers. The personal accounts, which use more colloquial words, represent an informal, individual, and interpersonal register; the official accounts, with a greater proportion of words related to politics, media, science, and technology, represent a more informational, professional register. When official accounts do use colloquial terms, these are almost always ILUs, whereas the personal accounts use English-origin ELUs fairly frequently. The comparative prevalence of sociopolitical and technical ELUs in the official subcorpus, on the other hand, suggests that English-origin loanwords which have been embedded in Welsh technical and professional vocabulary are well accepted in formal registers, especially in comparison to colloquial terms. This is consistent with the general association of ELUs with informality which we have already seen is pervasive in Cymrophone discourse. Nonetheless, the Cymrophone authors in the personal subcorpus also prefer Welsh-origin colloquial terms overall.

The over-estimation of listedness that results from using *GPC* alone as a source is not evenly distributed, affecting some semantic fields more than others. Since colloquial language is more strongly associated with informal registers, the fact that *GPC*'s listing of colloquial terms is significantly higher than that of other dictionaries may make *GPC* a particularly inappropriate choice for analysing lexemes in informal language. Like English-origin colloquial lexemes, Welsh-origin colloquial lexemes are slightly more likely to be listed in *GPC* than in the other dictionaries, although the difference is much less extreme than it is for the English-origin data. In contrast, the media lexemes are slightly *less* likely to be listed in *GPC*. It is possible that the use of up-to-date modern dictionaries, instead of, or alongside, comprehensive historical dictionaries like *GPC*, has the advantage of being more likely to offer coverage of recent coinages. Smaller dictionary projects may be quicker to respond to linguistic developments, while the full scholarly process of an historical dictionary requires more time to absorb neologisms or newly-introduced loanwords.

6.iii.b Listedness and formality markers: treiglad and morphological inflection

Potential limitations of the criterion of listedness as operationalized in prior work on language-mixing in Welsh-English bilingualism have already been discussed in detail above. In accordance with the conclusions drawn in that discussion, this analysis provides each of the lexemes in the Twitter dataset with a listedness score between zero and six, and each variant with a score between zero and seven (after adding *GPC* to the list of dictionaries consulted). Information about lexemes' listedness in *GPC* alone will also be visualized, in order to establish whether a more granular listedness score is, in fact, more informative.

Before considering how listedness and frequency interact with graphemic sequences, we shall consider two measures used elsewhere to evaluate loanword assimilation in Welsh: treiglad and morphological inflection. Because of the scale of the data and the constraints of time, it was not possible to check every instance of the lexemes under consideration here to establish whether there was a context which should trigger treiglad. Counts of how often a lexeme undergoes treiglad should therefore be interpreted with caution, as they indicate only the raw counts of treiglad, and do not show where an expected treiglad is missing, or where treiglad has been overextended to contexts which would not trigger it in standard Welsh. Morphology counts were made by counting any inflectional morpheme on a noun or verb, and noting which language the morpheme comes from. Since there is some question about whether suffixes which allow an ELU to be treated as a verbnoun (usually “-io”) should be counted for these purposes (Deuchar & Stammers, 2016), they have been excluded for the morphological inflection data discussed in this chapter.

Rates of both treiglad and inflection are labelled here as “low”, “medium”, and “high”, rather than being more precisely quantified. This is because rates are, on the whole, quite low as a percentage of all uses of a lexeme; most lexemes, whether Welsh-origin or English-origin, appear most often in an uninflected form and without treiglad. The rates for both were therefore normalized using the maximum score recorded for Welsh-origin lexemes (which always had higher rates of Welsh morphology and of treiglad). Scores are measured on the same scale for both languages; however, the normalized numbers themselves do not communicate useful information about the actual instances of inflection or treiglad counted in the dataset. Since each of these measures can only be considered for a subset of the data (nouns and verbs, and lexemes with initial phonemes that can undergo treiglad), the total

number of lexemes involved in each case is not the same, and not as large as in the dataset as a whole. For each plot, the number of lexemes or variants involved is given underneath the relevant box, so that the robustness of the plots can be better evaluated.

Figure F.6a shows the rates at which English- and Welsh-origin inflections are applied to English- and Welsh-origin lexemes in the Twitter dataset⁴. English-origin morphemes are almost never applied to Welsh-origin ILUs. Welsh-origin morphology is comparatively more frequent on ELUs which are in *GPC* than those which are not. However, since only three of the Welsh-origin ILUs and twelve of the English-origin ELUs are unlisted in *GPC*, the difference is not particularly reliable. Figure F.6b shows the same data, this time split according to listedness in the six dictionaries. When listedness is judged according to whether the lexemes are listed in none of the dictionaries (“unlisted”), or in at least one dictionary (“listed”), the data looks fairly similar to the previous plot showing *GPC* listedness, although there are now 23 unlisted English-origin lexemes, making the counts somewhat more robust.

In contrast to the first two plots, figure F.7 gives a more useful view of the morphology rates for ELUs. As might be expected, most of the Welsh-origin ILUs are listed in five or six of the dictionaries; only three are unlisted, and fifteen are listed in between one and four dictionaries. The figures for the Welsh lexemes are again, therefore, not particularly useful. The spread for the English-origin ELUs, however, is twenty-three unlisted, forty-two in between one and four dictionaries, and sixty-four in five or six dictionaries. The rate of English morphological inflections falls markedly as listedness increases. Rates of Welsh morphology are generally higher for all listed lexemes, regardless of the number of dictionaries in which they are listed. This suggests that a drop in rates of English morphology may be more useful than a rise in rates of Welsh morphology as a sign of an ELUs assimilation into Welsh.

Figure F.8 shows how often inflections are applied to individual variant forms, rather than to lexemes, based on whether the variant is listed in *GPC*. Once more, English inflections are essentially never applied to Welsh-origin ILUs, regardless of listedness in *GPC*. The ILUs also have approximately equal rates of Welsh inflections regardless of listedness. This suggests that the standardness of variant spellings in ILUs is not correlated with other features of linguistic Welshness, such as Welsh

⁴ All boxplots in Appendix F have whiskers whose limits are set at 1.5 times the interquartile range.

morphology, which are used to evaluate the assimilation of ELUs. In the case of the ELUs, rates of English morphology are lower for listed variants than for unlisted, and the opposite is true for Welsh morphology. Since the numbers of variant forms are higher than the number of lexemes, and there is a better balance of listed and unlisted variants (121 listed and 93 unlisted), these numbers are rather more reliable than those for the lexemes given in fig. F.6a.

As already discussed, listedness in *GPC* is likely to be more useful as a measure for variants than for lexemes. Fig. F.9 shows rates of morphology according to listedness, adding *GPC* to the other six dictionaries. The middle listedness category is for variants listed in one to five of the dictionaries, and the top category for those listed in six or seven. The data here do not provide a particularly clear picture of the value of the expanded listedness criteria, although fig. F.9 suggests that rates of English morphology on ELUs, in particular, seem to show a correlation with listedness. Nonetheless, the picture of English morphology is clearer than that seen in fig. F.8, using only *GPC* listedness. In conjunction with the data for treiglad, which we will examine next, this indicative of the usefulness of a full listedness score including *GPC* measure for variant spellings specifically, rather than for lexemes.

Only lexemes whose initial phoneme can undergo treiglad were analysed to determine rates of treiglad. In figs. F.10a and F.10b, rates of treiglad on English-origin ELUs and Welsh-origin ILUs are shown. English-origin and Welsh-origin lexemes which are listed in *GPC* show higher median rates of treiglad than those which are unlisted. Once again, however, only a small number of lexemes of either origin are not in *GPC*, and counts are therefore not fully reliable. The listedness categories are much more informative than *GPC* listedness for English-origin lexemes, showing that rates of treiglad increase with listedness. The same appears to be true for Welsh-origin lexemes, although the less balanced distribution of lexemes across the three listedness categories makes this conclusion somewhat more tentative. In all cases, it is also clear that Welsh-origin lexemes undergo treiglad more consistently than English-origin lexemes in the same listedness category; nonetheless, the median rates of treiglad for the more-listed English-origin ELUs is consistently a little higher than the median treiglad of the Welsh-origin ILUs in the category immediately below.

For variant forms, measures of treiglad are shown in figs. F.11a–F.11b, and the patterns already noted are repeated here. The numbers of variant forms involved make it possible to rely rather more on the data for both languages. Lexemes of either origin which are listed in *GPC* are much more likely to undergo treiglad than those

which are not listed. When *GPC* and the other six dictionaries are combined, a clearer picture emerges. Rates of treiglad on unlisted English-origin variants are reduced to zero, and rates for both English- and Welsh-origin variants listed in six or seven dictionaries are considerably higher than rates for the middle category. Indeed, median rates of treiglad on variants English-origin lexemes are higher than those on variants of Welsh-origin lexemes in the listedness category directly below.

Finally, rates of inflection and of treiglad were compared for the two subcorpora separately, to see whether the register differences already identified extend to these features. In fig. F.12a, we see that English inflectional morphemes are used significantly more frequently in the personal subcorpus than in the official, and that the personal accounts use Welsh morphology much more frequently on Welsh-origin than on English-origin lexemes. Welsh inflectional morphemes are used much more frequently by the official accounts for both English-origin and Welsh-origin lexemes. This result implies that these accounts are rather more likely to use tensed verbs than the personal accounts, suggesting that the use of inflection rather than periphrastic constructions (which use a verbnoun plus a particle) is likely to be a marker of the official Twitter register. Further research is needed to confirm this hypothesis.

Patterns of treiglad on Welsh-origin lexemes in the two subcorpora are less distinctive, with, if anything, rather more use of treiglad by the personal accounts. Treiglad is used far less often on English-origin lexemes by the personal accounts than by the official, however. Further research in this area is needed, since the increased rate of treiglad for Welsh-origin lexemes in the personal data suggests that treiglad on English-origin ELUs specifically may be a more robust register marker than treiglad in general.

6.iii.b.1 Discussion

Overall, these results suggest that considering the exact variant form in use, rather than the lexeme to which it belongs, may be useful in analysing English-origin ELUs used by Cymrophones in writing. Labelling ELUs as code-switch or borrowing, or determining whether they are treated as “Welsh” or as “English”, at the lexeme level alone may obscure patterns of use which are revealed by looking at the variant form.

Looking at the behaviour of variants can reveal anomalies which may reward further analysis. The numerous outliers on the unlisted English- and Welsh-origin variants in fig F.11b show that, while most of the variants rarely undergo treiglad, there are still quite a few which have high rates. One such English-origin variant is

<cansyr> (*cancer*). The variants of the lexeme found in the dataset which are listed in one or more dictionaries include <canser> (6 dictionaries) and <cansar> (1 dictionary), as well as <caner> (3 dictionaries)⁵. The form <cansyr> represents phonological fidelity to a PSEE-type pronunciation of the word, with the grapheme <y> corresponding to a schwa vowel. The fact that this variant behaves like those with higher listedness may suggest that it is becoming a new norm for spelling this lexeme in Welsh.

Comparing patterns of orthography in ELUs and ILUs suggests that treigladd and spelling may be correlated, since the rate of treigladd of both Welsh- and English-origin variants falls as the listedness of the variant falls (fig. F.11b). This pattern is not repeated for morphology, however (fig. F.9), suggesting that rates of treigladd may be characteristic of linguistic standardness, whereas the use of Welsh inflections is not. This would make sense, since the use of an inflection generally has a semantic purpose; a plural is used because the referent is plural. Treigladd, on the other hand, is not as directly and unambiguously linked to semantic meaning.

6.iii.c Conclusions

This section has shown how aspects of a lexeme's place in relation to both of a bilingual's language systems, such as its semantic category or whether it has a strong synonym in the other language, may influence how it comes to be used as an ELU. In particular, we have seen that listedness as a measure is sensitive to these features of lexemes, and also provided support for the claim that basing listedness judgements on a wider range of sources is, in general, more informative than simply choosing the most comprehensive dictionary available. The effect of register on ELU use has also been examined here, showing that semantics and synonymy may affect ELU acceptability in formal registers. This suggests that attention to register, as well as to these characteristics of lexemes, is necessary in evaluating the use of ELUs by Cymrophones.

By analysing variant forms in conjunction both with measures such as listedness and with linguistic features traditionally associated with standardness and formality,

⁵ <caner> is a borrowing of the Middle English lexeme {canker} (*GPC Ar-lein*, 2024, s.v. *caner*, n.). I have, however, treated it as a variant of the lexeme {cancer} for the purposes of this study, since the development of the two words is closely related (*OED3*, n.d., s.v. *cancer*, n. and *canker*, n.), and Modern English {canker} is unlikely to be the intended meaning of {caner} for modern Cymrophones.

this section has shown that variant spellings can open up avenues for understanding language-mixing behaviour which are not available when treating lexemes as homogeneous units. This suggests that a model of language which can account for bilingual languaging must be able to understand the interaction between languages not only at the level of lexical or syntactic selection, as much traditional code-switching research does, but also at the more granular level of phonemic and graphemic patterns.

6.iv Qualitative analysis

As already mentioned in Chapter 5.vi.a, Principal Components Analysis was used as a technique to prepare for qualitative analysis. This study's data are varied and complex, derived from three separate sources which differ in a variety of ways. The survey covers a limited number of lexemes, but is accompanied by comparatively rich demographic metadata about the participants who produced the data. The Twitter corpus has a wider range of lexemes, but no true demographic data: the only metadata available about the Twitter participants is their assignment to either the official or personal subcorpus, and limited information about their linguistic behaviour on Twitter, such as how many Cymrophone tweets were recorded from their account, how often they use metalinguistic features such as emoji and hashtags, and how many words they use on average per tweet. The data from *GPC* contains the largest set of unique lexemes, but these are not used in context. They can therefore only provide background information about traditional and standard patterns of orthographic adaptation.

A number of separate PCAs were first performed on the individual datasets, providing a reduced set of features which, in total, captured over 85% of the variance in each dataset. Since the ultimate purpose of these analyses was not to provide interpretable results in themselves, but to assist in selecting objects for quantitative analysis, they will not be described in detail here. Performing PCAs on each dataset separately produced a small set of components which could be projected onto the Twitter data (which is the sole object of qualitative analysis because it is the only dataset to show ELUs used naturalistically). A final PCA was conducted on the Twitter data with all the components in place. For the top ten components identified by this final PCA (accounting for 77% of the variance in the data), the twenty-five highest and lowest scoring variants were selected, and the lexemes they represent

were added to a list. In total, the list contained 507 entries. Many of these were repeated, some numerous times; the list contained 150 unique lexemes. The repeated entries are those which either have extreme scores on multiple components, or multiple variants of interest (or both).

The list was manually examined for patterns of particularly frequent lexemes, and four coherent categories of lexeme emerged: discourse markers (e.g. {thank you}, {actually}, {whatever}, {please}); proper nouns (including brand names) (e.g. {google}, {ukraine}, {photoshop}, {eurovision}); lexemes relating to traditional and social media (e.g. {upload}, {video}, {mention}, {photobomb}); and abstract nouns relating either to internal states or to professional contexts (e.g. {session}, {focus}, {commission}, {perspective}). The full list of lexemes and variants identified is provided in Appendix F.vi.

Many of these lexemes could be used in a variety of contexts, and need not always belong to the categories I have suggested above. Not all lexemes in the list fit into these categories ({sausage}, for example), and alternative categories could undoubtedly be identified. Nonetheless, these categories form four distinct strands, offering a focus for the qualitative discussion. Of course, the form taken by the quantitative analysis is shaped by this project's research questions, and is by no means made into a neutral or unbiased endeavour simply by using quantitative methods to direct the focus. Nonetheless, this approach has the advantage of providing access to patterns of orthographic and lexical behaviour which may not be immediately visible when looking directly at the variant forms. Rather than pre-selecting variants of interest, this method allows those variants to emerge from the quantitative patterns. In the remainder of the chapter, examples of salient lexemes from these categories will be examined in the context of the Twitter messages to which they belong.

6.iv.a Overview

Lexemes may remain orthographically unadapted for a number of reasons. In some cases, adaptation is not necessary, because the standard English orthography is not scoligraphic within a Cymrophone frame of reference. In the orthographic adaptation survey, for example, the lexeme {cardigan} was left unadapted by 70 of the 83 respondents. Those who did choose to adapt it focused on one or more of the vowels. Two respondents chose to represent the first vowel, /ɑ:/, without using an <r>, producing the variants «câdigyn» and «cadigyn». Two chose to place a circumflex on the grapheme <i> representing the second vowel, suggesting that their

realization of this vowel is closer to /i/ than /ɪ/: «cardîgen» and «cardîgan». Eleven respondents altered the grapheme representing the unstressed final vowel, /ə/, either to <e> (three respondents) or to <y> (eight respondents). These changes are sensitive to details of the PSEE spoken form of the word, responding in particular to the non-rhotic nature of PSEE, and to the use of schwa in unstressed syllables. Nonetheless, the vast majority of respondents saw no need to adapt the word, suggesting that the phonological form /kərdɪɡən/, implied by a Cymrophone interpretation the orthography, is understood by most as a fair representation of the English word's phonology (particularly if Welsh-English accents are used as the basis for judgement).

The lexemes highlighted by quantitative analysis, however, are often at least somewhat scoligraphic in Cymrophone orthographic terms. The lexeme {project} is one notable exception, since the grapheme <j> is now generally accepted in Welsh orthography. The graphemic string therefore implies the pronunciation /'prɔdʒekt/ according to the Welsh orthographic system, which is essentially the same as the PSEE pronunciation. In a few cases, the only deviant feature in a lexeme's standard English orthography is the use of graphemes which are missing from the Welsh alphabet: apart from the grapheme <x>, {text} and {brexit} conform well to Welsh orthographic conventions. In most other cases, the standard English form of the lexeme is undoubtedly scoligraphic from a Cymrophone perspective: the graphemic sequence <ukraine>, for example, implies the phonological form /ɪk'reɪnɛ/, and {surely} implies /sɪr'elə/.

Most lexemes have both scoligraphic and non-scoligraphic segments: in {photobomb}, for example, only the final <mb> for /m/ is truly scoligraphic. The standard Welsh spelling for assimilated ELUs containing the initial morpheme {photo}, however, uses <ff> to represent the first phoneme. In modern ILU orthography, initial <ph> is reserved for forms which have undergone aspirate treigladd from a lexeme with initial /p/. The spelling <photo> therefore implies treigladd from a lexeme *{poto}. This kind of breach of Welsh orthographic convention does not render the standard English form scoligraphic, but it is nonetheless a source of pressure towards orthographic adaptation.

The discussion that follows explores some of the orthographic variants identified for closer analysis according to two major themes: level of adaptation, and the creativity and performativity of the orthography. The aim of this discussion is to situate different types of orthographic strategy in context, and to consider how orthography functions as a resource for making meaning rather than merely as the

necessary vehicle for the concrete representation of linguistic forms and structures. All examples in the following section are taken from the Twitter corpus.

In this section, a generic user identification is given for each example, to show whether tweets have come from official or personal accounts, and so that it is evident when tweets come from different accounts. All hyperlinks have been anonymized as *http://url.com*. Usernames of individuals have been anonymized as *@username_1*, with the final number incremented when necessary to distinguish different usernames in the same text. (The numbering is reset for each text; *@username_1* does not represent the same username in every tweet.) Usernames for organizations and public figures mentioned in tweets are not anonymized, since knowing the username can often aid interpretation of the text.

6.iv.b Unadapted variants

Unadapted variants appear far more frequently in the personal than in the official subcorpus: 44,216 unadapted variants feature in the personal subcorpus, as compared with 17,016 in the official. Unadapted variants are more likely to be read as code-switches rather than integrated loanwords and, as we have seen, code-switching is associated with informality and non-standard language in Welsh. The greater frequency of unadapted variants in the personal subcorpus therefore fits with other characteristics of the subcorpora discussed previously.

Examples of unadapted lexemes in the official subcorpus are most often found in translations or sequentially bilingual texts, as in ex. 27, rather than as intrasentential code-switches. Tweets with more than a certain proportion of wholly English sentences were excluded from the corpus, so the examples still present are those in which Welsh is more prominent.

Official accounts are far more likely to write messages in which a discrete part of the text is in English. This may be a translation of the Welsh text, a quotation, or reference to external content written in English. In ex. 26, the unadapted lexeme {process} is part of an English summation of the preceding Welsh text. In ex. 27, it appears in the English title of the linked website. Example 28 shows the unadapted lexeme {upload}, which appears to have been automatically generated by a service that allows users to inform their followers that new content has been added to their website. The latter example is, to some extent, “pseudo-English”, a counterpart to the automatically generated “pseudo-Welsh” content identified in Chapter 5.iv.b.2.

- (26) Torgalonnus a amhosib prosesu'r llanast a'r gwallgofrwydd – impossible to process, too much reality (*Official User 98*)
Heartbreaking and impossible to process the devastation and the madness – impossible to process, too much reality
- (27) Gallai'r broses fabwysiadu fod yn gyflymach nag y byddai rhywun yn ei feddwl 🕒 Darllenwch am y broses ar wefan @nw_adoption 🖱️: The adoption process – North Wales Adoption Service <https://url.com> (*Official User 64*)
 The adoption process could be quicker than one would think 🕒 Read about the process on @nw_adoption's website 🖱️: The adoption process – North Wales Adoption Service <https://url.com>
- (28) MonFM uploaded "Chris Roberts yn sgwrsio hefo Richard Parry Jones am Etholiadau Ynys Môn 2013 11.04.13" <http://url.com> (*Official User 124*)
 MonFM uploaded "Chris Roberts chatting with Richard Parry Jones about Anglesey Elections 2013 11.04.13" <http://url.com>

In the official subcorpus, unadapted lexemes are therefore often not part of the account's own text, but refer to external content or are created by external systems. When official users do include Anglophone text, these are sometimes direct translations of the Welsh content, but may also serve as briefer summations, which function to give non-Cymrophone readers a way to understand the main thrust of the tweet, while still implying that Cymrophones are the primary intended audience.

In each of these cases, the discourse is characterized by clear separation of languages. When a lexeme is used in both the Welsh text and the English, it is adapted, orthographically and in other ways. In exx. 27 and 26, for example, the lexeme {process} is used in the Welsh text in its established Welsh orthographic form <proses>. It undergoes treigladd in ex. 27, and gains the morpheme {-u}, turning it into a Welsh verbnoun, in ex. 26. When non-established English-origin ELUs are used in this way in English portions of tweets, they often correspond to an ILU in the Welsh section, rather than being repeated in adapted form, as in ex. 29.

- (29) @username_1 Rydym wedi cyfeirio at eich trydariad ar #CymruFyw // Your Tweet has been mentioned on #CymruFyw (*Official User 1*)
 @username_1 We have referred to your tweet on #CymruFyw // Your Tweet has been mentioned on #CymruFyw

The lexeme {mention} can be translated with the ILU {crybwyll} (*to mention; to hint*), but this Welsh-origin lexeme is much more popular with personal accounts than official in the Twitter corpus, being used around twice as often overall, and by more than twice as many users, in the personal subcorpus. Official accounts tend to prefer periphrases like “sôn am” (lit. *mention about*) or “cyfeirio at” (lit. “refer to”). It becomes difficult to find a Welsh counterpart for the ELU {mention} in the social-media-related sense of an online reference to a person by including their username in a tweet, and many personal accounts use the ELU in this sense even if they would use an ILU for more general senses.

6.iv.c Adaptation and code-switching

Throughout both subcorpora, but particularly in the personal subcorpus, examples of unadapted variants are found in otherwise Welsh-language discourse. Unadapted forms of this sort are, as we saw in chapter 1, traditionally more likely to be interpreted as code-switching. Nonetheless, forms which are orthographically unadapted may still show other kinds of adaptation, such as the presence of Welsh inflectional morphology or the application of treiglâd. It is important to consider the interaction of different types of adaptation, as well as the functions of orthography in the context of the text as a whole.

- (30) ma raid oeddon fucked felly! dim realy isho darllan huna pan dwyn buta spars pizza neithiwr wan! !haha (*Personal User 174*)

he must have been fucked then! don't realy wanna read that while im eatin last nights spar pizza now! !haha

Example 30 shows an unadapted form of the lexeme {fuck}, with English morphology, in a tweet which is very informal in style. The author uses non-standard capitalization and punctuation for example «oeddon» (“oedd o yn”, *he was*), whose representation with standard punctuation would be <oedd o'n>. Features of colloquial pronunciation are represented by the forms «wan» (from “rŵan”, *now*), «isho» (from “eisiau”, *want*) and «buta» (from “bwyta”, *eat*). Nonetheless, there is comparatively little code-switching in this tweet other than the discourse marker «realy» and the phrase «spars pizza», whose spelling and word order mark it as English. The colloquial lexemes, in particular, are mostly Cymrophone, representing northern Welsh regionalisms.

- (31) Cysidro unfriendio bob tool syn sgwennu status ffwcedig “new year new me”
ne “2013 has/hasn’t been my year..” (*Personal User 47*)

*Considering unfriending every tool who writes a fucking “new year new me” or
“2013 has/hasn’t been my year..” status*

In contrast to ex. 30, the tweet in ex. 31 adds a Welsh-origin derivational morpheme to <ffwc>, an orthographically adapted form of {fuck}, producing «ffwcedig». The suffix <-edig> is commonly used to produce verbal adjectives from verbs or verbnouns. Taken in isolation, this form is far more integrated into Cymrophone linguistic patterns than «fucked». The orthography of <fuck> requires almost total adaptation in order to meet Cymrophone standards. Without adaptation, the <k> immediately marks it as non-Cymrophone; even without the <k>, the remaining graphemes suggest the Welsh phonological form /vik/. The adapted variant <ffwc> is listed by GPC, and the addition of a (somewhat formal) Welsh morpheme to it, rather than adapting common English past participle form “fucked”, adds a humorous Welshness to the resulting wordform.

As a whole, nonetheless, this second tweet features far more ELUs than ex. 30, even if we discount the quoted English phrases. They range from the established, listed, phonologically and orthographically adapted «cysidro» ({consider}), to the wholly unadapted variant «status» (whose English-origin etymon is, however, listed as <statws>). The variant «tool» is also unadapted, as is «unfriendio»; the latter, however, has gained a Welsh verbnoun morpheme. As in the previous example, this author uses colloquial or dialectal forms of Welsh-origin lexemes, like «ne» (“neu”, *or*) and «sgwennu» (“ysgrifennu”, *write*). The overall picture given by this tweet is of a colloquial text, which is (perhaps humorously) offensive, and which uses code-switching as a deliberate part of its style.

The linguistically transgressive nature of code-switching in Cymrophone culture is one of many layers of informality in the tweet, along with the insult “tool”, the swearword “fuck”, the use of colloquial lexis, and the incongruous combination of “fuck” with the rather formal suffix <-edig>. However, the fact that two of the ELUs are orthographically adapted appears incidental to the meaning. Both <ffwc> and <cysidro> are listed variants of their respective lexemes; neither, therefore, shows any particular creativity on the author’s part. The adaptedness of <ffwc> is, arguably, an advantage in its morphological integration with the suffix <-edig>, but it is not otherwise obviously meaningful.

- (32) Fel dudodd ryw gynhyrchydd radio barfog nid anenwog o’r gogledd wrth

drafod Gwn Glân Beibl Budr – “mae hi jyst yn record ffocing wirion o dda”..... (*Official User 110*)

As some not-unfamous beardy radio producer from the north said when discussing Gwn Glân Beibl Budr – “it’s just a fucking crazily good record”.....

The tweet in ex. 32 is a rare example of the lexeme {fuck} in the official subcorpus. The account tweeting here is, evidently, one of the less formal official accounts, part of the music industry rather than representing a major broadcaster or institution. The tweet’s style is somewhat informal, especially in comparison with many of the other tweets in the official subcorpus, but far less so than the style of exx. 30 and 31. The author of ex. 32 uses one dialectal form, «dudodd» (“dywedodd”, *said*), but otherwise their choices are standard forms of Welsh-origin lexemes, with standard Welsh orthography. Treigladd is applied in a standard manner, and both punctuation and capitalization are used in a standard way (with the exception of the extended length of the final ellipsis).

In this case, the orthographic adaptation of the lexeme as <ffoc> seems relevant to the acceptability of the tweet. The variant <ffoc> is listed in *GPC*, although only in phrases such as “ffoc off”; the forms listed at the headword are <ffwc> and <ffyc>, with the text “&c.” suggesting other possible spellings. The examples in this study’s corpus suggest that <ffoc> is the only other major orthographic variant of {fuck} in common use. It is also the least phonologically faithful to the PSEE or Welsh-English pronunciation of the word. This, along with the fact that it is part of a quotation ascribed to an anonymous record producer (who is nonetheless slyly described so as to be recognizable to a certain audience), serve to distance the author from the direct impact of having used the lexeme {fuck}. Although the word’s morphological inflection is English, the orthographic adaptation of the root morpheme make «ffocing» distinctively Welsh, as well as reducing the taboo associated with {fuck}.

These three examples give a glimpse of the semiotic complexity of evaluating ELUs in Cymrophone texts. Morphological adaptation, phonological adaptation, and orthographic adaptation, along with the listedness of the lexeme and of the variant, provide potentially illuminating context for understanding an ELU’s linguistic status. In some cases, orthographic adaptation appears random, a matter of convention rather than a deliberate choice. In others, a lack of adaptation may indicate a code-switch, and yet the text as a whole may be more linguistically Welsh than one which uses orthographically adapted forms. And in some cases, some part of the linguistic function played by ELUs is possible only because of their level and type of adaptation. In the next section, we will complete our survey of the Twitter data by considering

deliberate adaptation strategies which use orthography in creative or performative ways as an active semiotic resource.

6.iv.d Performative adaptation

6.iv.d.1 Proper nouns

Proper nouns and brand names are very often left unadapted in the corpus, since there are pressures against orthographic adaptation which are not present for common nouns and other parts of speech. Proper nouns referring to companies, like personal names, have a stronger claim to invariance; the spelling and form of the name are integral to its identity. Nonetheless, there are situations in which proper nouns may be adapted. The names of countries, for example, have historically been adapted from their English forms (or directly from the name as it exists in the country's own language) to make their orthography non-scoligraphic in Welsh. The country whose official name is *Україна* is therefore Anglicized as “Ukraine”, but is Cymricized as “Wcráin” (*Geiriadur yr Academi* (1995), *Geiriadur Prifysgol Cymru* (2002)).

Country names are a rare example of proper nouns which are commonly seen in an orthographically adapted form in formal registers of Welsh. The unadapted variant <ukraine> is used 32 times in total in the personal subcorpus, by 21 users; in the official subcorpus, it is used three times in total, by three users. In contrast, the orthographic form of {ukraine} listed in *Geiriadur yr Academi* (1995) and *Geiriadur Prifysgol Cymru* (2002), <wcráin>, is used 39 times in the official subcorpus, by thirteen accounts, but only sixteen times, by six accounts, in the personal subcorpus. The form without the accent, <wcrain>, has 55 uses by eighteen users in the personal subcorpus, and 28 uses by thirteen users in the official subcorpus. This implies that the presence of the accent has been fairly strongly established as part of the standard spelling in Welsh. As we shall see, the personal accounts are less likely to use accents as they are prescribed by orthographic rules, and much more likely to use them indexically, as a way of enhancing the Welshness of a spelling, rather than simply to reflect pronunciation.

The variant <iwcrain> is less popular, but still has seventeen uses by eight users in the personal subcorpus (with only two uses by two users in the official). This form represents an example of the tension between tradition and faithfulness in loanword adaptation. The form with initial <w> represents a pronunciation closer to that of the the Ukrainian name's initial <y>, with the Welsh CWM or DRWG vowels; the form

with <iw> represents the Welsh **LLIW** vowel, implying a Welsh-English pronunciation of the English phoneme /jʊ/. Although the variant <iwcrain> is fully adapted, then, it uses English pronunciation conventions as its point of reference. This, along with the established nature of <wcráin>, may help to explain why the form <wcrain> is more popular in the personal subcorpus amongst those who opt for an adapted form of the word.

Despite the tradition of adapting at least some proper nouns to Welsh orthography, there is nonetheless often a sense that adapting a name, as opposed to a common noun or other part of speech, is strange. During the Eurovision Song Contest in 2017, an official BBC Twitter account posted a tweet mocking the Welsh name “Portiwgal” (*Portugal*) (Wright, n.d.), although, as already noted, the adaptation of country names is standard practice in English as it is in Welsh. As a result of the special status of proper nouns, they often appear unadapted, even in the official subcorpus, unless there is a very well established Welsh orthographic form. The name of the Eurovision Song Contest itself is an example of this tendency. The variant <eurovision> appears 55 times in the official subcorpus as a stand-alone word and 22 as a hashtag, with sixteen and twelve users, respectively. In the personal subcorpus, the unadapted variant, as a stand-alone word, sees 126 uses and 64 users (247 as a hashtag, with 65 users). Although the unadapted form is by far the most commonly used, however, some adapted forms are found. These are restricted to the personal accounts, and have only one or two users each:

- <eurofision> 2 uses / 1 user
- <ewrofishion> 1 use / 1 user
- <ewrofishon> 1 use / 1 user
- <ewrofision> 5 uses / 2 users (+ 9/2 as hashtag)
- <ewrofisiyn> 1 use / 1 user
- <ewrofishon> 1 use / 1 user
- <ewrovision> 2 uses / 2 users
- <iwrofishyn> 2 uses / 2 users (+ 1/1 as hashtag)
- <iwrofision> 2 uses / 1 user (+ 4/2 as hashtag)
- <iwrofisiyn> 1 use / 1 user

- <iwrofsiwn> 1 use / 1 user

These adaptations, as a whole, target the representation of four phonemes: /ju/, /v/, /ʒ/, and /ə/. Except in one case, where it is left unadapted, the original <v> is changed to <f>, the only acceptable Welsh representation of /v/. The /ju/ in initial position is salient for all but the first variant form. Since the graphemic sequence <eu> realizes the phoneme /əi/ in Welsh, its adaptation is necessary. The more popular solution is to use <ew> here, mirroring the standard Welsh orthography of integrated ELUs such as “Ewrop” (*Europe*) and “Ewro” (*Euro*). As with the adapted form <iwcraïn> already discussed, forms with <iw> suggest a distinctively Welsh-English pronunciation of the initial /ju/. The Welsh pronunciation implied by <ew>, /eu/, is less faithful to the PSEE pronunciation of the word, but more conventional. The adaptations representing /ə/ are also varied, reflecting the two languages’ different approaches to unstressed vowels. Some users choose to represent this phoneme with <y>, recognizing the phonological similarities of English **COMMA** and Welsh **YN** vowel. Only one adaptor chooses the representation <w>, reflecting a common conventional Welsh spelling for English-origin ILUs ending <-ion>. The others retain English <o>, suggesting that this unstressed vowel is not as salient as the other adapted phonemes to the Englishness of the word.

The phoneme /ʒ/ is not present in Welsh, and the adaptors can therefore not opt for a distinctively Welsh representation that is faithful to the phonology. For many, the original <si> is maintained; this conventionally represents /ʃ/ in English-origin ILUs in Welsh, and so its use for that phoneme’s voiced counterpart /ʒ/ is reasonable. Some users instead choose the adaptation <sh(i)>. This sequence is not graphemically coherent in Welsh ILU spelling, so the adaptation actually moves the forms closer to English orthographic norms. This suggests that, to some extent, the aim here is adaptation for the sake of defamiliarization, rather than purely for the avoidance of scoligraphy. These adaptations are not intended simply to be functional, but rather to perform the inscription of Welsh languageness on the lexeme. Using the forms <sh> or <shi> to replace <si> does not add anything to the Welshness of the word, but it aids in marking it as different from the English form. The preceding adaptations of /ju/ and /v/ are distinctively Welsh, and <sh> has no competing phonemic association in Welsh orthography, so the defamiliarizing effect of its use is treated by these users as an advantage to the adaptation of the word as a whole.

This kind of playful, performative adaptation is much more common in the personal subcorpus than in the official. In many cases, as with the {eurovision} examples, it is a reflection of a light-hearted, informal style of writing. This does not,

however, mean that these forms are necessarily associated with stigmatized linguistic styles, however. This kind of performative adaptation is often associated with a register which, while it is informal and humorous, avoids casual code-switching and is characterized by adherence to standards of spelling and punctuation, as in example 33.

- (33) Ife Mam a fi yw'r unig bobol sy' ddim yn gwyllo Iwrofishyn a BGT? Drama Ffrensh ar BBC4. (*Personal User 36*)

Are Mum and I the only people not watching Eurovision and B[ritain's] G[ot] T[alent]? French drama on BBC4.

Although the author uses a colloquial style, they are careful to use standard written conventions, using an apostrophe to mark the colloquial abbreviation «sy'» (from <sydd>), as well as using standard punctuation and capitalization throughout. The author deploys markedly adapted variants, not haphazardly, but as signals of a playful, and distinctively Welsh, linguistic register. The adapted ELU «Ffrensh» (*French*) stands out, since there is a commonly-used ILU, «Ffrangeg», which could be used here in its place. The ELU is also used in a Welsh syntactic structure, placed after the noun it modifies rather than before it, as would be natural in English. The style of the tweet is carefully standard, even if informal, and this standardness provides the necessary backdrop upon which «Ffrensh» can be displayed to full effect, leaving the reader in no doubt that this ELU is being used deliberately rather than because the author is unaware of the Welsh-origin alternative.

Coming after «Iwrofishyn», «Ffrensh» is easy to read as part of a somewhat snooty grumble highlighting the author's more refined cultural tastes, but the use of orthographic adaptation humorously undercuts the potential pretentiousness of the tweet. The stylistic association of these two words implies that «Drama Ffrensh» (*French drama*) is the same kind of proper noun as «Iwrofishyn», referring to a specific artistic genre rather than using «Ffrensh» as a more general descriptor. «Ffrensh» is invoked as a term of cultural evaluation; a French drama is not merely a drama made in France, or in the French language, but one associated with the cultural symbolism of French cinema as a particularly highbrow genre. The adapted spelling encourages the reader to think about pretentiousness and cultural snobbery by insisting on using a word from the culturally-dominant English language and adapting it, rather than using the commonplace Welsh lexeme instead.

Example 34 is rather more formal in style than ex. 33; the only other adapted ELU is «Iwnicorns» ({unicorns}), part of the final hashtag. This adaptation mimics

the initial <iw> of «Iwrofsiwn», and the use of the ELU itself may be explained by the fact that the standard ILU equivalent would be <uncyrn>, the plural of <uncorn>; a fairly uncommon form which might prove particularly hard to parse when combined with other words in a hashtag. As in the previous example, adding the orthographic adaptation allows the author to present the ELU as part of the overall humorous (or, in this case, sarcastic) tone of the tweet as a whole. The fact that the author leaves «Brexite» unadapted, despite its obviously non-Welsh orthography, might reflect an unwillingness to bring humour into what many Welsh Twitter users saw as a political catastrophe, but might also imply a desire to distance Welsh culture from Brexit; unlike «Brexite», «Iwrofsiwn» is a whimsical musical spectacle, and not inimical to the author's ideal of Welsh society and culture.

- (34) Tybed os ydi aros yng nghystadleuaeth Iwrofsiwn am fod yn un o gytundebau masnach ryngwladol mawr @LiamFox ar ôl Brexite? #IwunicornsAcEnfysau (Personal User 26)

I wonder whether staying in the Eurovision contest will be one of @LiamFox's big international trade agreements after Brexite? #UnicornsAndRainbows

Although the lexeme {brexite} appears most frequently as the variant <brexite> in both subcorpora (1,198 official; 1,262 personal), other variants have at least some currency. The most popular simply makes the orthography consistent with Cymrophone norms by replacing <x> with <cs> to create the variant <brecsite> (92 personal; 26 official). This is the only other variant found in the official subcorpus. A small minority of personal accounts (four in total) produce forms with <u> or <ù> instead of <i> for the final vowel. This is still phonologically coherent, but has the advantage of forming a pun: <brecsut> can now be interpreted as containing the Welsh interrogative pronoun “sut” (*how*), and serving as a political comment on what users see as the unanswered questions surrounding this particular political process. The forms with <ù> seem likely to be using the accents in the way we have already seen illustrated, drawing extra attention to the orthographic play inherent in the form rather than attempting to convey some specific phonological feature.

- (35) Galla i ddim cytuno mwy da'r brawddegau ola hyn... belifflop o gan llath mewn i bwl nofio heb ddŵr oedd, ac ydyw Brecsùt, o'r cychwyn cynta hyd heddi. 🙄🙄🙄 (Personal User 19)

I cannot agree more with these last sentences... Brecsùt was, and is, a bellyflop from a hundred yards into a swimming pool without water, from the very first up to today.

🙄🙄🙄

- (36) Yn y bôn dwi'n teimlo, rwy'n credu, y byddai'n well gen i weld Brecsut yn sleifio mas y drws cefn (*Personal User 49*)

Fundamentally I feel, I think, that I would prefer to see Brecsut slinking out the back door

These examples are both written in styles which alternate formal and informal linguistic features effectively. In both cases, a few added colloquial forms are used, such as the allegro forms «da'r» (<gyda'r>, *with-the*), «ola» (<olaf>, *last*), and «heddi» (<heddiw>, *today*), in the first text, and the dialect lexeme {mas} (*out*) in the second. The first text's orthographically adapted «belifflop» ({bellyflop}) is both semantically informal and, as an ELU, marks a colloquial register of Welsh. On the whole, however, the diction of both texts is very standard, and in some cases tends towards formality, as in the sequences “oedd, ac ydyw” and “byddai'n well gen i”, for both of which there are a variety of more informal and colloquial Welsh alternatives. Both texts set up a register of Welsh whose balance of informality and formality, colloquialism and standardness, help to make linguistic play more salient to the Cymrophone reader. When the pun comes in the orthographic adaptation of {brexit}, it need not be highlighted or explained. The style of the tweets reassures the reader that these are authors who are aware of language and confident in using its resources. The spelling of the variant forms is therefore unlikely to be an accident, and the pun can be appreciated as a deliberate act of creativity by authors who are making use of their full bilingual semiotic resources.

6.iv.d.2 Performing Welshness: Welsh and Welsh-English accents in orthographic adaptations

We have already seen how circumflexes may be used to add Welshness to an orthographic form, illustrated by examples from this study's corpora. This final section will consider how characteristic features of Welsh and Welsh English phonology are deployed by Cymrophones in orthographic adaptations, and how these may interact with traditional loanword adaptation practices.

The proper noun {google} features a word-final [V C_{sonorant}] sequence which, as we saw in Chapter 4.i.d, often triggers vowel epenthesis in WE, probably reflecting transfer of Welsh phonological constraints. Four adapted variants of {google} in the Twitter dataset show an orthographic representation of epenthesis, placing either <y>, <e>, or <a> before final <l>. The use of <y> implies the insertion of schwa, while <a> and <e> imply that the vowel has a more definite quality. None of these choices

reflect the distinctively Welsh tendency for an epenthetic vowel to belong to the same vowel set as the preceding vowel.

Each of the variants of {google} which represent epenthesis appears only once in the corpus. More common are forms with final <gl>, which do not attempt to represent a characteristically Welsh epenthetic vowel but rather the PSEE realization with a syllabic sonorant. All adapted variants use the graphemes <w> or <ŵ> to replace the <oo> of the unadapted orthography. As we have seen, the Welsh languageness of <w> as a representation of a vowel is strongly established; it is one of the graphemes most saliently Welsh to Anglophone viewers, as we saw in Chapter 4. It may therefore be that this single grapheme, especially when combined with a circumflex, is sufficient for a distinct performance of Welshness. It is also possible that the deliberate representation of PSEE phonology with Welsh orthography is the intention of the forms without epenthesis, with humour located in the tension between the Englishness of the phonology and the Welshness of the orthography.

Comparing {google} with listed loanwords such as {example} and {bottle}, it seems that established orthographies have some effect on users' adaptation strategies. Adapted forms of {bottle}, whose listed Welsh form is <potel>, are almost always written with an epenthetic vowel, usually <e>, but sometimes <a> (representing a northern Welsh pronunciation). Adapted forms of {example}, on the other hand, are almost always written with no epenthetic vowel. Only one instance with an epenthetic vowel, <egsampyl>, is found in the corpus. {example} is listed in *GPC* as <esiamp< (the most common modern form, borrowed from Middle English) or <ecsamp< (borrowed from Modern English, now not generally used in formal registers of Welsh). Both official and personal accounts almost always reflect features of the established loanword form, <esiamp<, including the sequence <si> (representing a pronunciation with /ʃ/), which is less faithful to the Modern English pronunciation with /ks/, instead opting for fidelity to the established Welsh form. In contrast to the salience of vowel epenthesis as an aspect of the performance of Welshness in the {google} instances, therefore, {bottle} and {example} show the influence of established loanword status as a standardizing force, particularly in lexemes without a particular emotive or pragmatic force (like discourse markers) or cultural salience (like proper nouns).

Durham (2022) found “g-dropping” and “h-dropping” to be strongly enregistered features of WE accents for users of other varieties, but not for users of WE themselves. In this study's Twitter dataset, “h-dropping” was unusual, even though the representation of /h/ with a zero-grapheme is not unusual in the

established loanwords analysed from *GPC* (see Appendix C.iii for graphs showing the distribution of graphemic representations of each phoneme in the *GPC*, Survey, and Twitter datasets). Forms with final /ɪŋ/ do not feature in the Twitter dataset, but in the Survey data, the lexemes {fetching}, {hatching}, {overtaking}, and {painting} together feature g-dropping on around 17% of forms. In contrast, h-dropping occurred in only 1.2% of relevant contexts in the Survey data. This suggests that Durham's (2022) finding that h-dropping is non-salient for WE users holds true for Cymrophones. On the other hand, although g-dropping is not a highly distinctive sign of Welshness for all respondents, many of them do consider it as distinctively Welsh. This is particularly interesting, since g-dropping is characteristically a phenomenon of English-language speech. The examples in the Survey dataset imply that, even in Welsh-language linguistic contexts, a significant proportion of Cymrophones may consider characteristics of Welsh English as contributing to overall Welshness in loanword adaptation.

WE features which are uncommon but nonetheless salient in the Twitter data are illustrated by the lexemes identified in the “discourse marker” category. The lexeme {thank you}, for example, shows examples of “th-fronting”, a feature of WE noted by e.g. Paulasto et al. (2021), where underlying /θ/ and /ð/ are realized as [f] and [v]. The variant <ffanciw> is produced 21 times by 16 different users, suggesting that it is a fairly well-recognized feature. The fact that this lexeme provides the only context in which th-fronting occurs, however, may suggest that it is considered marginally emblematic by Cymrophones, rather than a feature of their own speech which they regularly represent in orthographic adaptation. Since this feature is somewhat deprecated in British English, users who deploy it do so to create a humorous or colloquial tone, highlighting the informality and friendliness of the interaction with language that places the author in something of a linguistically subordinate position, using both a deprecated language and a deprecated pronunciation. This reflects an appropriate stance for the speech act of thanking, in which the focus is on complimenting the message recipient. More research in the use of WE features by Cymrophones might profitably consider the enregistered features of Welsh varieties of English, as distinct from PSEE, with a particular focus on the performance of social and interpersonal status using WE.

The discourse markers {whatever}, {whatevs}, and {beautiful} (used as an interjection rather than simply as an adjective) all show examples of a WE feature not noted by either Paulasto et al. (2021) or Durham (2022): the realization of intervocalic /t/ as an alveolar tap or flap, /ɾ/, with variants such as <worefyr>,

<warevs>, and <biwriffyl>. In total, this feature appears 22 times from 12 different users. Like th-fronting, this is not a frequent feature, but it is widespread enough to be interesting. The evidence from these examples suggests that further research studying the range and frequency of a flap/tap realization of /t/ in WE could add to our understanding of Welsh Englishes. Again, it seems no accident that this representation of a WE accent feature appears only in discourse markers. Like {thank you}, {beautiful} performs an interpersonal function, expressing the user's appreciation. Interpersonal functions are also to the fore in use of {whatever} and {whatevs}, but here the meaning is dismissive and potentially rude. The use of an informal WE pronunciation may work to mitigate any potential rudeness, suggesting a satirical persona using a voice that is not, or not entirely, the user's own.

Chapter 7

“TIREZ LE FFYCIN RIDEAU”: Conclusions

The delicate and beautifully composed red flowers alongside the coffin read
TIREZ LE FFYCIN RIDEAU
By spelling ‘fucking’ in Welsh, I suppose, Wil had got it past a monoglot
florist.

(C. Meredith, 2012, p. 229)

“The problem with the French is that they don’t have a word for entrepreneur”, words attributed (but likely never spoken by) U.S. president George W. Bush (Mikkelsen, 2002). The amusement value of this aphorism hinges on the languageness of *entrepreneur*, a French-origin word so wholly assimilated into the English lexicon that the then “village idiot” (Mikkelsen, 2002) of U.S. politics could allegedly not realise the incongruity of his statement.

In literal terms, of course, the claim is ridiculous. Within a connotative semiotic framework, however, the claim is at least potentially coherent. Borrowed words are denotational signs, unions of content and expression. Like that of any other denotational sign, it is possible for the content plane of a loanword to shift over time. The meaning of the word changes, in the original language, in the borrowing language, or in both, so that the expression becomes an expression for two apparently different contents in the two languages.

Even at the point of initial entry into the IL, the ELU will not always bring its entire semiotic content with it. When an ELU is borrowed, it is usually a combination of form (a phonetic and/or orthographic string) and meaning (an abstract concept) that enters the IL. F. W. Field (2002). Information about the word’s morphosyntactic categorization in the EL is not a necessary part of the borrowing process, especially if it is incompatible with the IL. This means that an ELU’s assignment to a syntactic category in the IL may occur analogically, according to

similarity between the ELU's form-meaning set and similar form-meaning sets in the IL, rather than based exclusively on its EL category F. W. Field (2002). A borrowed ELU therefore need not have the same meaning or contexts of use in the IL as it does in its original language. It is theoretically possible, therefore, for French to have no word with the same denotational or connotational content as the English lexeme {entrepreneur}, despite having been the source of that lexeme.

Agha's (2005) enregisterment framework, along with his insight about the semiotic potential of multiple type-faithfulness evaluations in bilingual language-mixing Agha (2009), offers a rich resource for understanding how Cymrophones use language. Orthographic adaptation of ELUs appears, on the face of things, to be an unmarked choice. This study's data show that the use of standard and conventional adaptation is a marker of tweets from official accounts on Twitter, as is the use of unadapted English words as short translations or pointers to the meaning of a primary, Welsh text. Even in a context where the normal rules of formality and standardness might be expected to relax, standard Welsh orthographic practices are strongly enregistered as official, institutional practices.

In tweets from personal accounts, on the other hand, non-standard orthography serves a variety of functions. At a macro level of analysis, it is clear that a lack of orthographic adaptation of ELUs which do not conform to the Welsh system characterizes informal, colloquial communication. On the other hand, a different kind of informality by extreme orthographic adaptation practices, which add "Welshness" to variants by the strategic use of particular graphemes, grapheme sequences, and even simply by adding diacritics. For the Anglophone gaze, Welsh orthography has long been enregistered as foreign, troublesome, and transgressive. For Cymrophones on Twitter, Welsh identity is enregistered in the deliberate use of orthographic forms which are scoligraphic to an Anglophone audience.

The model of language introduced in chapter 2 described the relationship between phonology and orthography as a micro-semiotic one, but also permitted orthography a direct relationship with cultural context within the a connotative semiotic framework. The Twitter corpus and the elicited orthographic adaptation data show the utility of this model for situating ELU orthography in Cymrophone language. The SFL models which, as we saw, allow context no direct relationship with the expression-substance of language are insufficient to explain how different kinds of context are enregistered in different kinds of orthographic and phonological practice.

The act of orthographic adaptation itself carries meanings relating to

languageness, but also potentially to formality. Individual graphemes carry meanings which directly relate them to phonemes, or even to phones, as language users seek to represent particular speech sounds and ways of speaking (Durham, 2022). Individual graphemes also carry meanings of languageness, humour, and resistance, however, which may work with their phonological meanings, or overcome them altogether, as we saw in the non-standard uses of diacritics as marks of orthographic Welshness even though this implied a phonological form which was not faithful to any common PSEE or WE pronunciation of the lexeme.

The evaluation of faithfulness to type (Agha, 2009) appears, therefore, to recur at multiple levels of analysis. One and the same linguistic object is evaluated, not only against two linguistic systems, but also against different kinds of potential meaning. A graph is evaluated against its respective types: the graphemes of the two languages' orthographies. It also, however, participates in a semiotic connection with a phone as a result of the second articulation, forming a micro-sign which is evaluated for its faithfulness to the paired grapheme-phoneme types which make up each language's writing system in its role as an expression of phonology. The graphemic string which makes up a morph is another token, this time of the morphemes; and, on the expression plane of the denotative sign, the graphemic string is a component of the lexeme which is evaluated against the connotational type-systems of the two languages.

The written form, like the phonological form, is embedded at every stage of semiosis. The units of the second articulation, the *figuræ* Hjelmslev (1961), are by no means incidental to the macro-level, contextual, connotational meanings of language as a social semiotic. As a concrete embodiment both of language and of *languages*, the *figuræ* are the interface between the meaner and her meaning on the one hand, and the interpreter and her interpretation on the other.

7.i Regina progenitrix: the contextual power of languageness

The poster in fig. 7.1, designed by Welsh graphic artist Carwyn Hedd (Hedd, 2023), is an ideal example for testing the framework developed during this study, illustrating how complex layers of Welshness and Englishness can interact to create what is, ultimately (and paradoxically), a distinctively Cymrophone text. The poster is a complex intertextual, multimodal, and multilingual pun whose ludic effect relies upon



Figure 7.1 “God Sêf Ddy Chwyn” (Hedd, 2023).

the recipient's knowledge of both English and Welsh. The title of the English national anthem, "God Save the Queen", is altered in two ways: the final word, "Queen", is replaced by the Welsh "chwyn" ("weeds"); and the preceding words are adapted to Welsh orthographic rules. <God sêf ddy> implies a phonetic form /gɒd sɛv ðə/, representing a Welsh English pronunciation of the phrase via the monophthongal realization of the vowel of {save}. The final word, <chwyn> is pronounced /χwɪn/, and its similarity to /kwi:n/ licenses a pun referring to cannabis, colloquially called "weed". The pun is visually supported by the image of Queen Elizabeth II smoking a joint, and the use of a colour palette associated with the Rastafari religion, one of whose sacraments involves smoking cannabis.

The linguistic status of the ELUs and ILU in the poster's slogan is difficult to tease apart. One reading might consider the whole phrase to be in English, with a code-switch into Welsh for the pun on "chwyn". The matrix language here is undeniably English: English word order is exclusively used, and none of the English-origin ELUs can reasonably be understood as a borrowing in this context. However, the orthographic adaptation of "save" and "the" renders these words scoligraphic from an Anglophone perspective. Of course, "God Save the Queen" is a fixed phrase, rather than a spontaneously constructed sequence of independent lexemes. It is therefore possible to think of the slogan as a multiword code-switch into English, if the semiotic context establishes Welsh as a kind of matrix language at the connotational level. By this, I mean that Welsh is understood to be the language of the communicative act, even if the dominant language of the phrase's syntax and lexis is not Welsh.

Continuing the metaphor established by the use of the Latin "matrix", this overarching contextual language might be called a "progenitrix" language. The progenitrix language is determined at the level of the communicative context, rather than by the syntactic patterns of the words being used. It is negotiated through connotation, by participants in an act of communication signalling their linguistic capacity and preferences, and by features of the context in which the act of communication takes place. Unlike the matrix language, therefore, the progenitrix language is not wholly controllable by the participants. In a classroom in a Welsh-medium school, for example, there will often be a strong expectation that Welsh is the progenitrix language; producing a single utterance in English in such a context would generally be insufficient to change this expectation and establish English as the new medium of communication. Indeed, as we saw in Chapter 1.vi.e, a pupil who produced an English utterance might well be censured by the teacher.

Only an extraordinary change in the context, such as if a non-Cymrophone guest were invited to speak to the class, would be likely to establish English as the progenitrix language; and then only for the duration of the guest's visit. In the words of one of Selleck's (2013, p. 31) respondents from the Welsh-English bilingual school, "even though it's a bilingual school they want you to always talk Welsh".

In the case of the "God Sêf ddy Chwyn" poster, the progenitrix language is, to some extent, dependent on the context in which the poster is seen. On Hedd's website, amongst other designs whose textual content is predominantly in Welsh and whose references are to Welsh popular culture and politics, the poster's progenitrix language is implicitly Welsh. In this context, the orthography is easily decoded by a Cymrophone, but is not essential to establish the progenitrix language. Outside this context, however, there is no guarantee that anything *but* the orthography will establish that Welsh is the poster's progenitrix language. The unchanged nature of the orthography of <God>, as well as the poster's strong stylistic echoes of the iconic sleeve design for the Sex Pistols' single "God Save the Queen" (fig. 7.2) (Reid, 1977)), lead the viewer to expect the English phrase.

The orthography of <sêf> is the first indication that these linguistic expectations will be subverted. The presence of the circumflex on the middle grapheme of <sêf> is crucial to signal the switch from English to Welsh. The subsequent correspondences of <f>, <dd>, and <y> to [v], [ð], and [ə], respectively, are also strongly imbued with Welsh languageness, and prepare the Cymrophone reader for the final pun. A non-Cymrophone, in contrast, is unlikely to be able to decode the orthography, and will not understand the pun.

This analysis suggests that code-switching may profitably be examined at a level above that of denotational meaning. Below the level of the lexeme, the grapheme-phoneme correspondences beginning with the <ê> of <sêf> signal a connotational switch to Welsh, even as the denotational meaning of the lexical items remains English. Orthographic code-switching, like lexical code-switching, relies on pairs of signifier and signified: the grapheme is the signifier, and the phoneme is the signified. Welsh and English share the majority of their graphemic inventories. The major exceptions are <k>, <q>, <v>, <x>, and <z> in English, and the accented characters (particularly those with a circumflex) in Welsh. Using any of the exclusively English characters when the denotational language is Welsh, or vice versa, is a sign of orthographic code-switching. Since Welsh and English share so many characters, however, orthographic code-switching is also marked by the use of language-specific grapheme-phoneme correspondences, or by orthographic



Figure 7.2 "God Save the Queen" (Reid, 1977)

sequences which are legal in one language but not in the other.

It is no accident that the first sign of orthographic Welshness in “God sêf ddy Chwyn” is the unambiguously non-English grapheme <ê>. The Welsh word “sef” (a particle meaning roughly “that is”) already has the pronunciation /se:v/; the addition of the circumflex is therefore not strictly necessary to establish a Welsh-English pronunciation of “save”. However, without the circumflex, the semiotic Welshness of the phrase is harder to recognise; this is the only character in the phrase which, taken alone, is obviously Welsh rather than English. Its use therefore cues the Cymrophone reader to a switch of the progenitrix language, and prompts the interpretation of what follows as a sequence of Welsh grapheme-phoneme correspondences.

The most important aspect of this interpretation of “God sêf ddy Chwyn” is that the matrix language of the phrase does not change. Indeed, the maintenance of English as the matrix language is essential for the correct interpretation of the phrase as a whole. The reader must recognise the need to switch to Welsh as the progenitrix language in order to decode the graphemes correctly and reconstruct the phonetic sequence /gɔd se:v ðə/, corresponding to the English words “God Save the”. The visual cues of the poster itself prompt the recognition of the English phrase at the denotational level, but the orthography prompts a switch to Welsh as the language of contextual meaning. Without this switch, the final pun would be much harder to interpret.

Having used the combination of visual symbolism and orthography to prepare the Cymrophone reader to expect the English phrase but to decode it through the medium of Welsh, interpreting the final word <chwyn> correctly requires recognizing a code-switch at the denotational level. The faithfulness to Welsh-English pronunciation displayed by the preceding words is inconsistent with the use of <chwyn> to represent [kwi:n]; a more likely sequence would be <cwîn>. Although the reader will recognize the similarity of [χwin] to [kwi:n], she is therefore prompted by the *dissimilarity* of the two phonetic sequences to seek an alternative interpretation of “chwyn”. The appropriateness of the Welsh denotational meaning “weed” is confirmed by the spliff in the Queen’s mouth and the poster’s background colours.

The languageness of “God sêf ddy Chwyn” is simultaneously English and Welsh; it is inextricably bilingual. The reader must understand both the Englishness and the Welshness of the slogan’s cultural, syntactic, lexical, and orthographic content in order to appreciate its humour. The poster is distinctively Cymrophone, not despite the English-origin elements, but because of them: it addresses Cymrophones as

bilinguals, rewarding them for their natural language-mixing abilities at the moment they deploy them to decode the pun.

7.ii Primary research question: orthographic variation and semiotic meaning

The remainder of the chapter will be spent offering answers to the research questions posed at the beginning of this study. Particular attention will be given to pedagogical and policy implications, including desiderata for future research.

RQ1: Does variation in the orthographic adaptation of English-origin lexical items in modern colloquial Welsh function as a vehicle for semiotic meaning?

Naturalistic data from Twitter and elicited data from the survey both support the claim that orthographic practices in the representation of ELUs by Cymrophones are vehicles of semiotic meaning, in multiple ways. It is no accident, for example, that the quantitative analysis described in the previous chapter indicated the particular salience of discourse markers amongst the ELUs examined. These are often informal, colloquial lexemes which are used to manage discourse structure within and between utterances, and range from the standard and commonplace to the taboo and the transgressive. Their use aligns the language user with a variety of possible connotational meanings, and the various types of adaptation strategy are used to enhance, dampen, or shift the emotive force of the ELUs in use.

Orthography, whether in itself, as a representation of phonology, or as the expression-form of a lexeme, is able to bear multiple meanings and to encode multiple linguistic, personal, and cultural strategies. The epigraph to this chapter, from C. Meredith's (2012) novel *The Book of Idiots*, illustrates how orthography can be used, not only for constructing and performing identity, but also as a strategy of cohesion and exclusion, as a taboo English-origin lexeme is placed within a French phrase, and hidden from the Anglophone gaze through using Welsh orthography. The word made Welsh need not be a Welsh word, or even a word used within a Cymrophone text. What is important is that the context of use offers the audience a moment of linguistic solidarity. Welsh orthography becomes not merely a vehicle for phonology, but a means of imagining a truly bilingual community in which the

skillful use of both languages, acting in concert, is the crux of the text. “Yt’s basicyli a sicret langwej!”.

7.iii Secondary research questions: orthography, enregisterment and languageness

Q2: To what extent can the observed variation in orthographic adaptation be ascribed to enregisterment?

Q3: To what extent can the observed variation in orthographic adaptation be characterized as variation in languageness?

These questions appear, by this point, to be two different perspectives on the same question. This conclusion comes via a semiotic model which, in the tradition of Hjelmslev (1961) and Barthes (1965), imagines a language as a co-occurring set of linguistic objects which share the same connotator of “languageness”. In answering the main research question, moreover, we have also answered these questions, showing that orthography, whether as an expression of phonology or as a visible representation of some larger syntagm such as a lexeme, can be imbued with “Welshness” and “Englishness”, “not-Welshness” and “not-Englishness”, as well as a variety of other types of languageness indexing, amongst other things, geography, standardness, and formality. Orthography allows Cymrophones to perform a wide variety of personae. Moreover, it has affordances which are unique to its own system, such as diacritics, which gain connotations of Welsh languageness through their consistent use in the standard language, and which can therefore be used in non-standard ways to perform Welshness.

7.iv Tertiary research questions: scholarly and pedagogical implications

7.iv.a Research Question 4

Q4: Do the observed patterns of variation have implications for scholarly approaches to code-switching and borrowing?

This study has raised a number of issues of relevance to scholarly research on ELUs. The construct of listedness which, as we have seen, features prominently in many previous studies of Welsh-English bilingualism, was critiqued and a refinement proposed. By representing listedness as a cline, rather than as a binary category, I was able to achieve more useful and informative results than by using a large and comprehensive dictionary on its own. Furthermore, I showed that different approaches may work for different objects of focus: abstract lexemes benefit from a restricted definition of listedness, whereas forms with concrete orthography benefit from expansive and exhaustive listedness measures. This study therefore supports paying particular attention to this way of evaluating ELUs, and making a critical examination of the lexicographical resources to be used rather than assuming that bigger is better for all purposes.

The different approaches needed for best results with lexemes and variants also implies that ELU research might benefit from varying its objects of study. Critiques of previous research have questioned an all-or-nothing approach to ELU integration, in which a word is either a fully integrated borrowing, or an unintegrated code-switch (Deuchar et al., 2018). This study suggests that caution must be taken in any categorization of ELUs which applies at abstract levels, such as lexemes. Indeed, the wide variety of orthographic adaptation strategies employed in the dataset examined here, and the corresponding richness of semiotic functions of those strategies, suggests that paying attention to the adaptation strategies themselves could be a fruitful approach. Welsh orthography is a characteristic and highly enregistered aspect of the language. A fruitful approach to language mixing might, therefore, be to ask whether a second code-switching has happened in the orthography after a first code-switching of the lexeme. Rather than seeing an adapted ELU, we might see a deliberately and obviously English lexical unit, with a deliberately and obviously code-switched Welsh orthography. This study's data provides strong evidence that, in some cases, the English languageness of the lexemes is the point of using them: they exist within a Cymrophone's discourse in order to be English, and in order to have their orthography code-switched into Welsh. Although traditional research assumes that code-switching below the level of the morpheme is impossible, I suggest that a model of language which gives the micro-semiosis of the second articulation its proper place provides us with a framework in which to situate sub-morphemic code-switching as a valid theoretical construct.

7.iv.b Research Question 5

Q5: Do the observed patterns of variation have implications for Welsh-language pedagogy?

Question 5 is the one with the most significant real-world impact, and therefore requires a more detailed response than the previous questions.

Welsh-language pedagogy does not exist in a vacuum. It is inextricably linked to the political, cultural, historical, and geographical landscapes of Wales. Understanding the various ideals of Welshness which are embedded in these landscapes, and how they affect language ideologies, is essential. Chapter 3 therefore gave a detailed account of attitudes towards English and its influence on Welsh, showing that a long history of prescriptivist condemnation of language-mixing lies behind contemporary attitudes. The ideological construction of Welsh-languageness as a determinant of Welshness is bound up with romanticized ideals of Welsh rural life, imagining a pure, monolingual Cymrophony. The impossibility of this ideal in modern Wales does not make it less potent, particularly given the nationalist monolingualism which dominates contemporary political (Blackledge, 2002; Fozdar & Low, 2015; Van Avermaet, 2009) and pedagogic (Iannacci, 2008; I. U. Khan et al., 2023; Willans, 2010) ideologies across Europe and beyond.

Having looked in detail at the treatment of English-origin ELUs in two pedagogical texts in Chapter 2.iii.b.1, and at attitudes in culture and in educational contexts in Chapter 1.vi, it is evident that the theoretical construct of languageness is often treated inconsistently and incoherently by modern Cymrophones. The ideological weight of monolingualism privileges the use of Welsh-origin lexis, and takes a punitive approach to language-mixing. English-origin lexis, on the other hand, has a liminal position. Although it is clear from the texts analysed in Chapter 2.iii.b.1 that some English-origin lexis is considered to have acquired Welsh-languageness, and therefore to be acceptable for use by Cymrophones, there is no clear and coherent basis for deciding which words have Welsh-languageness, and which retain English-languageness. As a result, different sources of information about language have different opinions about what, exactly, constitutes an “English” word, for the purposes of Cymrophone discourse.

In a review of prior work on code-switching and borrowing in Chapter 1, we saw that most recent work on Welsh-English bilingualism has drawn on the approaches of Poplack and Meechan (1998) and Myers-Scotton (1993), seeking to distinguish code-switching from borrowing using factors such as frequency of use

and level of integration. I proposed, instead, to reconceptualize the distinction as one of languageness: English-origin ELUs are borrowings if they have sufficient Welsh-languageness, and are code-switches otherwise. As a result, different variants of the same lexeme may be treated differently, depending on a variety of features which affect their overall languageness. These include features inherent to the variant and its immediate context, such as orthographic adaptation and morphosyntactic integration, as well as more broadly contextual features, such as frequency of use and dictionary listedness. The data analysis in Chapter 6 showed some of the ways in which these various features interact, but further research is needed to tease out the ways in which an ELU's languageness is established in context. As we saw, orthographic resources are used by Cymrophones in complex ways, to perform different types of Welshness and Englishness (including a distinctively Welsh Englishness).

The data collected for this study provides evidence that Cymrophones can and do combine detailed knowledge of the phonologies and orthographies of PSEE, WE, and Welsh in their orthographic code-switching practices, creating meanings which could not be conveyed by any of these codes in isolation. It also shows the importance of ELUs to establishing registers of Welsh: both lexemes which retain greater Englishness and those which have been adapted to display a performative Welshness characterize informal communication. Adaptedness does not necessarily correlate monotonically with formality, therefore; both very low and very high levels of orthographic adaptation signal different types of informality.

Clearly, orthographic adaptation, like other kinds of language mixing, is a practice which takes skill and detailed knowledge of the codes in question. Nonetheless, as this study has shown, attitudes towards language mixing of all kinds, including in some cases the use of well established loanwords, is strongly deprecated in Cymrophone culture. The bias against language mixing is particularly strong in Welsh-medium education. This not only risks disadvantaging pupils unfairly, by punishing them for using forms which are elsewhere treated as legitimate loanwords; it also enregisters an impossible variety of Welsh as the only "good" Welsh. This can only exacerbate an already-difficult linguistic situation in which, as we have seen, young Cymrophones fail to acquire enough distinctive registers of Welsh to be able to achieve their educational potential, and are prevented from enregistering Welsh as a language of casual, social relationships.

7.iv.b.1 Language policy: from monolingualism to plurilingualism

Chapter 3 presented evidence suggesting that an insistence on Welsh, and Welsh alone, as the language of communication in Welsh-medium and bilingual schools engenders both a lack of linguistic confidence and a sense of grievance and rebellion in children. Stigmatizing the use of English, while it may be intended to protect Welsh as a minoritized language, has the practical effect of removing pupils' linguistic agency. As Selleck (2013, p. 35) notes, pedagogical practice is therefore at odds with the liberal, equitable language of Welsh policy makers, described in Chapter 3.ii.b, which is "premised on choice and on the inclusive principle that the option to learn and use Welsh (and English) is open to all, regardless of linguistic background".

The treatment of English-origin lexemes by Cymrophones, both in educational settings and elsewhere, further reflects tensions between what Selleck (2013) terms "flexible" and "separate" bilingualism. The former allows users to select their language of choice, as well as permitting (and even encouraging) the use of varieties which mix languages in various ways. The latter insists on keeping the languages as discrete, unmixed codes, and often enforces the use of a specific language in a given context. Selleck (2013, p. 31) describes the use of Wenglish (by which she means a mixed variety combining English and Welsh in various ways) as a strategy of flexible bilingualism, which allows "greater social inclusion between students with vastly differing linguistic experiences and competencies". One of her respondents, a student at a predominantly English-medium school, admits that a Welsh teacher might not like to hear Wenglish, but asserts that "it could get you bilingual though if you don't know that much Welsh" (Selleck, 2013, p. 30). Although the likelihood that informal mixing of languages amongst friends would truly be sufficient to "get you bilingual" may seem fairly low, it is nonetheless interesting that the English-medium schoolchildren in Selleck's (2013) study volunteered fairly positive attitudes towards Welsh, seeing bilingualism as a valued skill.

Young people in mixed-language social groups outside school report sensitivity to perceptions of the Welsh language by non-Cymrophones, too (L. Evans et al., n.d.). As E. M. Thomas and Roberts (2011, p. 105) state, "English is the language of inclusion". For some young people, this dynamic may translate into less use of Welsh in social contexts; for others, however, it spurs the adoption of social strategies to include their non-Cymrophone peers by using a hybrid, bilingual language whose use of both inter- and intra-sentential code-switching makes it more accessible for those without advanced competence in Welsh (L. Evans et al., n.d.). Selleck (2013, pp. 30–31) notes that the use of Wenglish is not only a way for groups of pupils to

accommodate each other's varied linguistic repertoires, but also a means of "identity performance" by which pupils who class themselves as non-Cymrophone can "negotiate a position within the national category of being 'Welsh'". For some of Selleck's respondents, therefore, displaying bilingualism through language mixing is a way to bolster one's Welsh identity. For the pupils at the bilingual school, however, this is not the case. In predominantly Welsh-medium education, displaying bilingualism through language mixing is instead interpreted as a threat to the integrity of the Welsh language, and as an expression of alignment with English (linguistic, cultural, or national) values.

It is useful to bring in the notion of a progenitrix language here to explain the asymmetry. Where English is the implied or prescribed progenitrix language, there is little scope for other languages to be perceived as a genuine threat at a macrolinguistic level. English is in a dominant position, both within Wales and globally. When Welsh is the progenitrix, however, English is perceived as intrusive and threatening because, at a macrolinguistic level, Cymrophony is genuinely threatened by the spread of Anglophony in Wales. Opportunities to use Welsh are limited, and teachers are aware that many pupils have little exposure to the language outside the classroom. Every English word or phrase, in this context, symbolized one fewer opportunity to expose pupils to Welsh.

We have seen in previous chapters that the stigmatization of English where Welsh is the implied progenitrix language extends outside the classroom. Cymrophones who choose to speak English are seen by some as violating an unspoken social rule: an introduction to Cymrophone television presenter Elinor Jones on the BBC RadioCymru website states "Does dim yn ei gwylltio mwy na Chymry Cymraeg sy'n dewis siarad Saesneg a'i gilydd"¹ (Llwyd, 2010). In 2021, former Member of the Senedd Cynog Dafis wrote an open letter expressing his disappointment that, in the official business of the Senedd, "fe glywn ni siaradwyr Cymraeg rhugl yn optio am y Saesneg, y *lingua franca* ryngwladol honno y mae ei lledaeniad yn erydu'r defnydd cymdeithasol o'r Gymraeg o ddydd i ddydd ac o flwyddyn i flwyddyn"² (Dafis, 2021). For some, therefore, there is an implicit sense that Welsh is the default progenitrix language for Cymrophones, unless its use is

¹ *Nothing enrages her more than Cymrophone Welsh people who choose to speak English with one another*

² *we hear fluent Welsh speakers opting for English, that global lingua franca whose spread is eroding social use of Welsh day by day and year by year*

barred, for example by the presence of a non-Cymrophone interlocutor or audience. To use English instead is to violate an unspoken linguistic contract, and there is no contextual or personal circumstance which can make it acceptable for a Cymrophone to select English where Welsh is possible.

The patterns of orthographic practice portrayed in this study are illustrative of a kind of cross-linguistic creativity and play which is either stigmatized or ignored in Cymrophone educational contexts, yet is extensively used to imbue English words with Cymrophony. This maintains Welsh at the level of the progenitrix language even as the matrix language is switched to English, since the words can only be interpreted easily by those who are familiar with Cymrophone orthographic conventions. Orthographic code-switching therefore becomes an important means of satisfying the unspoken linguistic contract, maintaining the Cymrophony of the discourse even if the lexical and syntactic content switches to English. The examples shown in the previous chapter include orthographic switching for proper nouns such as company and product names, which are more often treated as having a fixed orthographic form (for reasons such as brand recognition and trademark registration). Respelling even these names is therefore a strongly marked choice, drawing attention to the Cymrophony of the discourse as well as potentially fulfilling other functions such as humour.

7.iv.b.2 Plurilingual strategies for educational policy and practice

Despite their importance in contemporary colloquial discourse, the creative uses of orthography seen on Twitter are non-standard and, therefore, essentially invisible to many traditional pedagogic approaches. Transgressive language of many kinds can have functions in minoritized-language contexts which make them culturally important. Code-switched taboo words may be used to mock the common conception that swearwords do not exist in minoritized languages (Muehlmann, 2008), to signal emotional engagement (Dewaele, 2010), or to create humour (Fägersten, 2017), and their use should not be dismissed out of hand merely as deliberate defiance of social norms (Martínez & Morales, 2014). Informal, colloquial, playful language is a crucial part of young bilinguals' enregisterment of a language as *theirs*, a language of play and fun and creativity (Auckle, 2017; de León, 2019; D. C. S. Li, 2000; Paugh, 2019) rather than a language imposed upon them by some authority. Indeed, within reason, taboo language should be included in plurilingual schooling, helping pupils to grasp the emotional and social impact of swearwords (Adaros & Tironi, 2017; Dewaele, 2016a, 2018; Finn, 2017; Horan, 2013).

Simon (2001) suggests that learners acquiring an additional language develop a double identity, tied to language: the learner-identity is associated with the target language being learned, and the person-identity is associated with the learner's primary language. The ability to express feelings and personal identity tends to remain tied to the primary language, at least while the individual lacks the resources to express themselves fully in the target language. Code-switching in the classroom may therefore, according to Simon, function to assert the individual's wish to be understood as a person rather than being bounded by her institutional role as a learner. For the teacher to reject code-switching, by insisting on speaking the target language exclusively, is to reject this renegotiation of the relationship, and to maintain the formal boundaries of the pedagogical context. Alternatively, however, teachers may engage in facilitative code-switching, reframing the context as one where the goal is to maximize learning, rather than to follow unquestioningly a target-language-only pedagogy (Simon, 2001). In order for code-switching to become a useful pedagogical tool, Simon (2001) recommends taking a bilingualist approach to learning, seeing the development of bilingual competence through code-switching and other mixed-language strategies as an end in itself, rather than merely a stage on the journey to strict monolingual use of the target language.

Although many pupils in Welsh-medium education come from Cymrophone families, or have community exposure to Welsh outside school, many do not. Students who begin Welsh immersion at a young age are not in exactly the same position as the additional-language learners described by Simon (2001), yet they share some similarities. In particular, if they only encounter Welsh at school, learners may enregister Welsh solely as a language of education. E. M. Thomas and Roberts (2011), and never develop a strong sense of Cymrophone person-identity. When their use of English, whether as the language of discourse or in code-switching, is denigrated and repressed in school, it may be internalized as an attempt to repress their individuality, and to force them into a solely institutional identity. The supposition that immersion-educated Cymrophones fail to enregister Welsh as belonging to them on an individual level is supported by the fact that many abandon the language when they leave school *Cymraeg 2050* (2017), Price (2020), and E. M. Thomas et al. (2014b).

Enregistered as an institutional language, Welsh becomes irrelevant once the institutional context is left behind. This phenomenon has also been observed, for example, in Gaelic Medium Education in Scotland: Smith-Christmas (2017) found that immersion education in Scottish Gaelic does not necessarily translate into use of

the language outside the classroom, and that the strong association of Gaelic with education means that the language “is parceled up according to domain instead of existing as an integrated part of community life” (Smith-Christmas, 2017, p. 43). Immersion education provides pupils with resources that open linguistic options to them, but they often choose not to use the minority language outside school (Smith-Christmas, 2017).

The plurilingual proposals offered here advise promoting connections between languages, including mixed-language varieties which are currently more often stigmatized. The aim of these proposals is not only to offer pupils the resources that expand their linguistic options, but also to give them the best chance possible of enregistering English, Welsh, as well as mixed varieties of the two languages, expansively and inclusively, as languages for living rather than merely for schooling.

Despite research showing the benefits of home-language maintenance and linguistic diversity in education, “additive orientations toward linguistic incorporation remain largely unsupported and underdeveloped in schools” Iannacci (2008, pp. 115–116). The following discussion proposes a variety of ways in which plurilingualism might be incorporated into different aspects of the Welsh educational landscape. The proposals here are by no means exhaustive. The aim is to open a discussion about how plurilingualism could be used to inform Welsh education both in policy and in practice.

7.iv.b.3 Plurilingual parents, teachers, and institutions

Pedagogy benefits from involving parents, giving them the resources to make informed choices about their children’s linguistic development (Luckett, 1993). In settings where one language (often a former colonial language) dominates over others, parental freedom of choice may simply become a route to the hegemony of the dominant language as the language of schooling, if parents are not informed of the alternatives and their benefits, since the benefits of the majority language will otherwise be the most visible (Luckett, 1993). Parents should be advised not to hide their own language(s) and dialect(s) from their children; instead, they should be helped to support their children’s linguistic development at home, and (where appropriate) be invited to participate as experts in language awareness activities in the classroom (Beacco et al., 2016; Castellotti et al., 2008).

Successful plurilingual pedagogy also requires the incorporation of plurilingualism into teacher training programmes (Castellotti et al., 2008). The topic of linguistic anxieties should be addressed explicitly with trainees, including exploring

the anxieties of both pupils and teachers (Gabryś-Barker, 2020). Teachers' guilt about their own code-switching should be mitigated by explaining the advantages of using different languages for various aspects of the lesson (Gabryś-Barker, 2020). Training should include discussion of attitudes towards language mixing and social variation, addressing trainees' concerns about plurilingual pedagogy, and introducing current knowledge about language acquisition and the cultural dimension of language (Beacco et al., 2016). Trainees should be given opportunities to experience plurilingual methods, for example by attending training courses run in multiple languages, by learning an unknown language, or by developing plurilingual curricula (Beacco et al., 2016). It is also important that teachers have training in different varieties of languages (*CoE: LPD*, 2007).

The academic literature describes a variety of ways in which teachers use code-switching strategically, for particular pedagogic purposes. This may involve switches of single words within an utterance, or switching languages for larger stretches of discourse. Teachers have been found to use switching for purposes including to check comprehension (Jiang et al., 2014; Reyes, 2008); to explain difficult concepts (I. U. Khan et al., 2023; Ngcobo, 2013), for metalinguistic discourse (Gabryś-Barker, 2020; Üstünel & Seedhouse, 2005), for classroom management (Gabryś-Barker, 2020; Üstünel & Seedhouse, 2005), and to make teacher talk more comprehensible (Jiang et al., 2014; Reyes, 2008). This evidence should be incorporated throughout Welsh teacher training. Trainees in all subject areas should, in particular, be introduced to C. Williams's (1994) *trawsieithu*, and have opportunities to practise implementing it in lessons. *Trawsieithu* not only offers a plurilingual framework, but has the specific advantage of having arisen in the context of Welsh-medium education. This may reassure trainees that the framework is relevant to them and their practice.

Plurilingual teacher training should include up-to-date information about code-switching, addressing myths about code-switching as a sign of deficit, and showing how it can be used as a strategy for teaching, learning, and creativity. The role of language in identity formation should be discussed in detail. Teachers must enter the profession with a firm awareness that "early exposure to both languages has a good chance of leading to fluent bilingualism, which will include code switching" (Deuchar et al., 2016, p. 235). W. M. Collins's (2003) work with Guatemalan users of the minoritized language Mam highlight the disconnection between community attitudes to code-switching and those of trained teachers. Teacher training in Wales should encourage teachers to respect the attitudes and linguistic practices of the wider

community, and discourage any tendency to assume that education and training make teachers unassailable as authorities on linguistic matters.

Trainees should learn how to develop students' acquisition of academic Welsh without reprimanding them for code-switching or for using English. Teachers must be trained to praise pupils for achievements, regardless of which language they are using, and to use pupils' knowledge of one language as a scaffold for developing equivalent knowledge in the other (Iannacci, 2008).

School language policy in Wales cannot ignore concerns about ensuring sufficient exposure to Welsh. However, the main focus should be on developing students' competences, rather than preserving the language. Schools and teachers should not use the language's minoritized status as a stick with which to beat students who code-switch or use English. Schools must be encouraged and supported by policy makers to replace monolingualist approaches to pupils' language with plurilingual ones. Strategies to encourage the use of Welsh alongside English, and to raise awareness of contexts in which code-switching is inappropriate, must be trialled (ideally within a formal research framework, so that their effectiveness can be properly measured).

Schools should take particular care to avoid promulgating ethnonationalist understandings of the Welsh and English languages, for example by telling pupils who use English to move to England. Schools should also ensure that pupils have a thorough understanding of terms such as "Cymro"/"Cymraes" and "Cymraeg"/"Cymreig", emphasizing the fact that the "Cymry" are not all Cymrophone, and raising awareness of the social history of languages in Wales. English should be represented in schools as an authentically Welsh language, which is as much part of pupils' heritage as is the Welsh language itself.

On the national level, bilingual institutions such as the Senedd provide an excellent and public model of bilingualism, with both Welsh and English used regularly. The Senedd should consider establishing a Welsh programme along the lines of the European Youth Parliament (EYP) (*European Youth Parliament*, n.d.), but with a bilingual ethos where the EYP uses British English as its working language. As part of such a programme, children from across Wales would take part in activities to develop their political and civic awareness, using either Welsh or English, with professional translation and interpreting services provided to support all participants. This would encourage closer links between pupils in Welsh-medium and English-medium education without requiring the Cymrophone pupils to use English. It would also allow all pupils to experience a professional multilingual context,

showing them that Welsh and English can co-exist on an equal level in such a setting. Other opportunities for children from English- and Welsh-medium schools in Wales to meet equally and plurilingually should also be explored, to build bridges between language communities which could persist beyond the school years.

7.iv.b.4 Plurilingualism in academic language

Academic language should be taught explicitly as a part of plurilingual education, including both written and spoken forms of the language of schooling, since it may not be easily acquired outside school (Castellotti et al., 2008). In Welsh-medium schools, this should include attention to academic English, since some students will go on to attend English-medium universities, and may find the transition difficult if they do not have an equal command of academic English as of academic Welsh (R. Jones & Desforjes, 2003).

The language of schooling needs targeted pedagogy for different academic disciplines, and particular attention must be paid to children whose primary language is not the language of schooling (Castellotti et al., 2008). In Welsh-medium schooling, this means both particular attention to academic Welsh for pupils from non-Cymrophone homes, and particular attention to academic English for pupils from Cymrophone homes, ensuring that pupils gain as symmetrical a competence as possible.

A particular need in Wales is for policy makers, curriculum experts, and examination boards to come together and agree on an approach to code-switching and the use of English-origin ELUs in academic language. The approach must be realistic and take into account the natural linguistic behaviour of Cymrophones. As Chapter 2.iii.b.1 showed, current advice about ELUs is inconsistent, lacks a solid theoretical foundation, and is sometimes unreasonably punitive. Teachers, pupils, and parents need clear and rational guidance on the use of ELUs, which does not, for example, insist that extremely common ELUs such as {ffrind} (*friend*), which are listed in practically every modern Welsh dictionary, are unacceptable and must be replaced in all contexts with the more formal {cyfaill}.

Welsh pupils should be presented with information about the normality of code-switching, its greater frequency amongst more balanced bilinguals, and its appropriateness in many registers of informal Welsh. They should also learn that all languages develop distinct registers, and that code-switching is merely one feature amongst many which marks less formal Welsh registers. Having learned that code-switching is neither lazy nor shameful, however, pupils can then be taught

specific strategies to avoid code-switching in formal oral or written contexts, where it is less acceptable.

Drawing on Poplack's (1980) observation that code-switching is characterized by fluency and a lack of hesitation, a prominent strategy to be taught to pupils should be the use of flagging formulae. These indicate the pupil's awareness that she is using an unassimilated ELU or EL phrase, rather than incorporating it seamlessly into the discourse and giving the impression that she is unaware of its potentially transgressive presence. The somewhat archaic set phrase "ys dywed y Sais" ("*as the English say*") might be appropriate in less formal contexts, particularly if a humorous tone is intended. More formal set phrases such as "yr hyn a elwir 'X' yn Saesneg" ("that which is called 'X' in English") should also be taught. The pupil should ultimately develop a range of flagging strategies for different contexts of use, which she is able to use confidently and fluently. Flagging allows the student to remain in a monolingual mode, "quarantining" the English word by advertising its presence, leaving the audience in no doubt that the student is aware of the non-Cymrophony of the word. Pupils should also be trained to employ circumlocution, finding ways to "talk around" the lexical lacuna in Welsh rather than filling it with an English word.

In addition, tasks which raise awareness of borrowing patterns in Welsh should be designed, to help pupils predict which English-origin lexical items are more or less likely to be acceptable in formal registers. Awareness of borrowing patterns should also include studying how English-origin lexical items are morphologically, orthographically, and phonologically adapted to Welsh, allowing pupils to produce properly adapted "nonce-loans" from an English lexeme. As this study has shown, Cymrophones are often instinctively aware of the orthographic resources available to them, and deploy them for a variety of functions in informal contexts. However, colloquial and creative orthographic patterns differ from the traditional patterns found in established loanwords in GPC, and from the standard rules of Welsh orthography established in (Orgaff, 1987). Teaching standard adaptation practices would therefore also help to raise awareness of spelling patterns in Welsh ILUs, including the appropriate use of diacritics, and how these reflect elements of etymology and word history. Instruction in the standard patterns of adaptation would enable pupils to use their instinctive knowledge strategically, teaching them how to use standard adaptation techniques in formal contexts, as well as heightening their awareness of the creative possibilities of non-standard adaptations.

An advantage of teaching strategies such as flagging, circumlocution, and awareness of borrowing patterns is that they are not confined to use in a bilingual's

primary languages. They can be transferred as tools in learning additional languages, and may well improve confidence and fluency in general. Research into the effects of learning such strategies in Welsh on overall fluency in Welsh, English, and other languages would be enlightening.

Alongside teaching accepted strategies for avoiding code-switching into English, official policy also needs to take a more open and realistic approach to the presence of English-origin ELUs in academic language. No student should be marked down for using an English-origin ELU if that ELU can be found in *GPC* or in one of the major modern Welsh dictionaries, for example. In order to encourage greater tolerance, examiners (and teachers) should be made aware of factors affecting pupils' confident acquisition of Welsh technical vocabulary, particularly the potential for Welsh technical terminology to be less confidently acquired when there are multiple alternative terms, when those terms are long or periphrastic, and when the terms are rarely encountered by the student. Terminology planning for Welsh should be implemented with the goal of developing Welsh equivalents with strong perceived validity; that is, a Welsh term of art should not seem less formal or less precise than the counterpart English term. Further research into factors affecting Cymrophones' willingness to use technical terms, and confidence in the correctness of those terms, is needed.

Corpora of contemporary Welsh should be mobilized to show how often children are likely to be exposed to equivalent English and Welsh technical terms. Corpus data should also be examined to discover how often a given English-origin ELU is used instead of an ILU by contemporary Cymrophones, since this is likely to have an influence on pupils' enregisterment of the ELU as a "correct" Welsh term for the referent. Corpora will give examiners and educators a realistic sense of the pressures affecting the acquisition of Welsh terminology, as well as giving authors of learning materials the ability to focus on raising awareness of otherwise underrepresented terminology.

An effort to expand the Welsh Wikipedia might profitably focus on adding articles for technical vocabulary, or expanding those that exist so that they match the equivalent English articles in length and quality. The use of Wikipedia is now widespread, and finding a Welsh technical term treated thoroughly on Wikipedia might encourage younger Cymrophones to enregister that term as official and correct. As ever, further research is needed to establish how the presence or absence of technical terms on the web, in dictionaries, in textbooks, and in general usage affects their acquisition and subsequent use.

7.iv.b.5 Plurilingualism in standardized assessment

The translation of formal examinations and tests is a complex matter, because fairness demands that different versions should not disadvantage one language group more than another. This is an ongoing problem in the Welsh context (D. V. Jones, 1998). Translators note that achieving equivalence of language versions is often very difficult, for a variety of reasons including inherent differences between the languages (Arffman, 2012). The factors affecting difficulty of test items are not limited to the item's abstract content, but also include the language in which the item is presented. Lexis, syntactic complexity, register differences, and cultural factors all affect the perceived and actual difficulty of an item (Asil & Brown, 2016; Grisay & Monseur, 2007; D. V. Jones, 1998).

Particular problems in ensuring the equivalence of English and Welsh tests arise in the differing familiarity of the lexis used and the differing transparency of some technical terminology (D. V. Jones, 1998). The use of treigladau must also be treated with care; a Cymrophone student could potentially be presented with three or four different written variants of a lexeme, which may affect word recognition (D. V. Jones, 1998). Dialectal differences in Welsh (which has no single standard variety comparable to Standard British English) add further complexity (D. V. Jones, 1998). In 1991, Lyon (1991) found that many Cymrophones preferred to hear radio and television programmes in English, rather than listen to an unfamiliar dialect of Welsh. It is not clear whether this preference persists to any degree thirty years later, amongst Cymrophones who have increasing exposure to media in a variety of accents and dialects. If it does persist, however, it only adds to the pressure on standardized-test authors to select a variety of Welsh which strikes readers as both authentic and readable. Research into the factors affecting readability in Welsh is needed, and a readability measure which is valid for both English and Welsh would be an extremely valuable tool for examination writers, as well as for textbook authors and teachers.

Any translation of a standardized assessment, however good, has an inherent equivalence cost (Grisay & Monseur, 2007). Differences in performance between different language versions of tests have been found not only in tests of language skills (such as reading and writing) but also in tests of mathematical and scientific skill (Gökçe et al., 2021). G. Johnes (2020) analysed scores on the international PISA assessments in Wales, concluding that the performance of Welsh-medium and English-medium pupils showed that Welsh-medium schooling was less efficient overall. Nonetheless, Johnes also notes “that pupils attending schools that deliver their

teaching through the medium of Welsh are bilingual, and so have skills in both languages that many of those attending English-medium schools do not; yet these skills are not tested by PISA” (G. Johnes, 2020, p. 62). A plurilingual approach to assessment of Welsh pupils requires further research to establish the extent to which the comparatively poorer results of Welsh-medium pupils on the PISA assessments is an artefact of the monolingual testing instrument or, indeed, of lack of equivalence between English and Welsh versions of the test.

Bilingual children, particularly children whose language of instruction differs from the language spoken in the community or at home, have often been found to underperform on standardized testing, compared with their monolingual peers (De Angelis, 2014; DeDonno et al., 2014; S. S. Khan, 2019; Pearson, 1993). Bilinguals’ performance in standardized tests may also be more strongly affected by time pressures (DeDonno et al., 2014). Standardized tests reproduce inequalities (Urrieta, 2004), as they are based on twin assumptions of linguistic and cultural uniformity which often disadvantage both bilingual and minority students, (S. S. Khan, 2019). As a result, students with less exposure to the dominant culture, as well as students who code-switch or exhibit other bilingual behaviours, are likely to underperform.

Research is urgently needed in Wales to establish whether and how the varying bilingualism and biculturalism of Welsh pupils affects their performance on standardized tests. C. Ellis et al. (2015) showed that Welsh-English bilinguals require significantly more processing effort when reading English texts containing factual information about Wales than when reading the same information in Welsh. This example, specific as it is, indicates that bilinguals’ cognitive processes may be affected by a multitude of (perhaps unexpected) factors, and that it is always risky to compare them directly to monolinguals without first seeking evidence that the assessment is fair to both groups. Regardless of the language(s) used at school, at home, or in the wider community, assessment of all pupils in Wales should take their language profile into account, and evidence should be sought to guide the development of appropriate instruments of assessment.

The particular difficulties facing Welsh-medium education must also be taken into account in assessment, notably the dearth of professionally-produced Welsh-language textbooks and learning materials (C. Baker, 1995; BBC News, 2016; C. H. Thomas, 1982; C. Williams, 1994), which is compounded by the limited Welsh-language content available online. Whether or not schools successfully employ *trawsieithu* or other mixed-language learning strategies in order to

incorporate Anglophone materials, it is still undeniable that the imbalance in availability of materials is a significant difference between Welsh-medium and English-medium schooling. Research is needed to uncover any effects on academic performance and development of the materials deficit in Welsh-medium education.

Attention should also be paid to the potential effects of home and community language(s) on performance in standardized assessments. It seems highly possible, for example, that Welsh-medium pupils from non-Cymrophone homes and/or living in largely non-Cymrophone communities may well experience some of the same educational disadvantages well-documented in children with different home and school languages in other contexts (Mancilla-Martinez & Kieffer, 2010; Reyes, 2008; Urrieta, 2004). If this does prove to be the case, the wealth of international literature describing methods proven to support such pupils must be drawn upon in Welsh-medium education and assessment.

It is also possible that students from non-Cymrophone families are at a disadvantage as regards the cultural content of the Welsh-medium curriculum, since they may have little exposure outside school to Cymrophone literary, musical, artistic, and cultural traditions. As discussed in Chapter 3.iv.b, high culture and popular culture are more closely related in Cymrophone society than they are, for the most part, in Anglophone society. Children with Cymrophone parents, or living in predominantly Cymrophone areas, may be advantaged not only linguistically, but also in exposure to high literary and cultural traditions.

Again, research is needed to discover whether pupils from non-Cymrophone families are, paradoxically, disadvantaged by being raised in the majority culture, which is nonetheless minoritized in school. Particular attention should be paid to the experiences and academic progress of pupils whose parents come from outside Wales, and who may therefore bring a complex constellation of cultural and linguistic identities to the Welsh-medium classroom, whilst also having less instinctive familiarity with traditional Welsh culture. Finally, further research should consider the academic experiences of Welsh-medium students from Cymrophone families, particularly as regards their acquisition of academic genres of English, since this may have an impact on their confidence in accessing Anglophone higher education.

Crucially, a plurilingual approach to assessment should not take a deficit view of pupils' linguistic and cultural backgrounds. If research finds that bilinguals are systematically disadvantaged in standardized testing, or that home/community language systematically affects test performance, the appropriate response is to explore how the test can be made fair, rather than to ask how the bilingual's competence can

be better made to conform to the test.

7.iv.b.6 Plurilingual learning materials

As has already been mentioned, Welsh-medium education suffers from a lack of professionally-produced learning materials such as textbooks, as well as from a corresponding lack of informal learning sources, particularly online: the Welsh Wikipedia had over 436,000 pages in September 2024; the English Wikipedia had over 47.5 million (*Wikimedia Statistics*, [n.d.](#)). The Welsh Government should allocate significant funding for the production of Welsh-language textbooks and other learning materials, as well as supporting (financially and otherwise) efforts to expand the Welsh Wikipedia and similar public-domain resources. Welsh materials should not solely be translations of existing English-language texts, but should also include wholly original content, geared to the specific context of Welsh-medium education, and helping pupils from all backgrounds to engage with Welsh history and culture alongside, and in connection with, global histories and cultures.

Welsh-language materials must pay particular attention to the problems outlined in Chapter 3.ii.a, and ensure that the image of Welsh people and Welsh culture they represent does not unintentionally reproduce a homogenous stereotype of traditional Welshness. A full range of modern Welsh identities, including Welsh and Cymrophone people of colour and disabled people, must be represented to pupils. Not only do white Cymrophone pupils need to see an expansive version of the personae which can be expressed through Welsh, but pupils of colour, as well as non-Cymrophone parents of colour who may be considering a Welsh-medium education, need to see themselves represented throughout Welsh pedagogic materials. Issues such as Welsh involvement in the British Empire and colonialist projects should be taught alongside information about the repression of the Welsh language and exploitation of Wales' natural resources, ensuring that the complexities of Wales' relationship to empire are not elided.

Welsh-medium teaching resources, particularly materials for language teaching, should take care not to include value judgements about the use of language(s), and in particular must avoid labelling code-switching as "lazy". Where possible, a variety of authentic linguistic styles should be included in textbooks, to allow pupils to enregister their linguistic features and contexts of use. When materials present texts for linguistic discussion, these should not simply be translated from English examples, since this risks flattening out the distinctive features of Welsh registers, and assumes that a Welsh register will use the same linguistic resources as the corresponding

English one. In particular, this is likely to mean that bilingual practices such as code-switching and innovative orthographic adaptation are never represented, despite (as this study has shown) being an important variable distinguishing informal Welsh registers from formal ones. Materials for language study should therefore, wherever possible, be based on authentic Welsh texts.

Materials for studying English in Welsh schools should pay specific attention to Welsh varieties of English, encouraging pupils to draw out the influence of the Welsh language on Welsh English. Pupils, whatever their medium of education, must be encouraged to value English as a Welsh language, and not only in the form of Standard British English. Pupils should be encouraged to reflect on their own use of English, as well as that of friends and family, and to collect examples of distinctively Welsh language for classroom discussion.

The language experience of immigrants to Wales who have learnt Welsh should particularly be sought, with materials designers creating videos, audio interviews, and/or written texts which describe these people's experiences and their motivations for learning Welsh. (This should include interviewees whose primary language is English, whether they come originally from the United Kingdom or from another Anglophone country). The interviewees' reflection on similarities and differences between Welsh and their primary language(s) should be highlighted, and used as the basis for activities designed to encourage pupils to reflect on their own experience of cross-linguistic transfer.

Humorous and ludic approaches to language-mixing should play a prominent role in the design of teaching materials. The kinds of orthographic adaptation and respelling exemplified in this study's data provide an accessible way to encourage pupils to use their linguistic knowledge playfully, with plentiful exemplars showing how spelling can be a tool of humour. Extracts from works of literary humour which play with spelling and homophony should also be included. *N'Heures Souris Rames* ("Nursery Rhymes") (de Kay, 1980), for example, presents texts which are readable (if obscure, and somewhat semantically nonsensical) in French, but which sound like English nursery rhymes when read aloud. *Anguish Languish* ("English language") (Chace, 1956), meanwhile, constructs new versions of well-known texts using English near-homophones of the original words. It includes texts originally in English, such as *Sinker Sucker Socks Pants* ("Sing a Song of Sixpence"), and in French, such as *Alley Wetter* ("Alouette"). Activities based on these examples could ask students to produce similar versions of texts which are familiar to them (using Welsh, English, and other languages), and then encourage their fellow pupils to decode

them.

A plurilingual approach to materials design might also include bilingual learning materials which deploy both languages (and, potentially, others) strategically. Such materials could draw on the trawsieithu framework to encourage pupils to read primary sources in one language, and discuss or summarize in another. It would also be valuable to include primary sources (for example in History textbooks) in their original language, accompanied by a translation into English, Welsh, or both. Pupils with knowledge of the original language would be enabled to engage directly with the text, and the teacher could encourage them to share their knowledge with the class. Members of the school community (teachers, school support staff, pupils from other classes, and/or parents) with knowledge of the languages in question could be invited to the classroom as expert informants, helping pupils to interpret the texts. Seeing their other language(s) represented in the linguistic landscape of the school can have a positive effect on pupils, encouraging them to develop confidence and to use their full linguistic repertoire (Iannacci, 2008).

Castellotti et al. (2008) suggest that presenting documents in other languages can “servir [...] de support dans la construction de connaissances”³ (Castellotti et al., 2008, p. 20), a description which is reminiscent of the model of trawsieithu proposed by C. Williams (1994). Beacco et al.’s (2016, p. 162) summary of what they call “plurilingual activities”, too, sounds very like Williams’ trawsieithu, including activities such as watching a film in one language and summarizing it in another, or reading a story in one language, hearing a recording of it in a second language, and rewriting it in the language of schooling. Materials design for the Welsh context should ideally be informed by plurilingualism and trawsieithu, as ways to support and encourage comfortable, strategic bilingualism.

7.iv.b.7 Plurilingual curricula and classrooms

Plurilingual curricula, like plurilingual materials design, should be geared towards involving pupils’ full linguistic repertoires, as well as introducing them to previously unknown languages and varieties (Castellotti et al., 2008). The language(s) of schooling must be represented in full, including different varieties, registers, and genres, helping pupils to appreciate “que la variation est constitutive des langues”⁴

³ *serve [...] as a support in the construction of knowledge*

⁴ *that variation is constitutive of languages*

(Castellotti et al., 2008, p. 20). Classroom activities in all subject areas should incorporate linguistic skills and knowledge, highlighting to pupils the importance of subject-specific genres and norms, and (where necessary) making them aware of differences between English and Welsh.

Metalinguistic and sociolinguistic reflection should be included throughout the plurilingual curriculum, and incorporated as early as possible Castellotti et al. (2008). Language awareness activities are therefore central to a plurilingual approach. These raise awareness of how language functions in general, rather than teaching a specific language, and should include critical language awareness, which introduces discussion of sociolinguistic issues such as the relative power and status of languages (Ballinger et al., 2017).

Language awareness tasks might include comparing proverbs with the same meaning from different languages, or considering the syntax of negation in a known and an unknown language, or considering cognates and word families in different languages (Ballinger et al., 2017).

The awareness approach allows all the pupils' (and the wider community's) languages, whether primary or additional, at whatever level of competence they are known, to make a contribution to the discussion. In an immersion context, this has the advantage of providing a space for plurilingualism and respect for students' diverse identities, while allowing the language of schooling to remain the primary language of the classroom discussion (Ballinger et al., 2017). Deliberate exposure to language variation and metalinguistic discussion should benefit pupils' awareness of the language of education (Beacco et al., 2016; Castellotti et al., 2008). In Welsh-medium schools, pupils awareness of, and confidence in using, different registers of Welsh should be enhanced by activities which raise explicit awareness of register variation, including comparative consideration of the conventions for a given register in English and in Welsh.

An ideal plurilingual curriculum in Wales would see Welsh and English replaced with integrated subjects (perhaps called "Language" and "Literature"), which would encompass the study of both languages, including their literary traditions, as well as specific classes teaching academic varieties of the language(s) of schooling. This would have the added advantage of making a substantially unified curriculum possible for all Welsh schools, regardless of their medium of instruction. Welsh schoolchildren would have a common frame of educational reference, rather than seeing major curricular differences depending on the medium of their schooling.

Importantly, Welsh children in English-medium education would develop the

same familiarity with Cymrophone literature as those in Welsh-medium education, albeit through the medium of English where necessary. They would be offered access to the cultural traditions of Wales in the same depth as their Cymrophone counterparts, ultimately contributing to a reduction of the cultural distance between Cymrophony and Anglophony in Wales. This would also contribute to achieving one of the recommendations of a 2020 independent review of Welsh education, “that consideration should be given to how the provision for latecomers [to Welsh-medium education], or parents looking for a second opportunity to choose Welsh education, is a part of the education regime” (LinguaNi, 2020). A largely identical curriculum for English- and Welsh-medium schooling, with differences only in how much content is taught in Welsh and how much in English, would facilitate movement of pupils from one medium to the other.

During integrated Literature instruction, students might read Cymrophone and Anglophone novels from the same period, comparing themes, characterization, and style. Alternatively, they might study two texts, one of which inspired the other, such as Orwell’s *Nineteen Eighty-Four* and Owain’s *Y Dydd Olaf* (mentioned in Chapter 1.vi.d). In English-medium schools, the Cymrophone texts would be read in translation where necessary, with Welsh-language instruction based on short passages or a simplified version of the text. Students would be encouraged to read more Welsh texts in the original language as their competence increases.

Students might also study the use of cynghanedd, including attempts to use it in Anglophone poems, and its influence on the style of poets such as Gerard Manley Hopkins, as well as considering the linguistic differences which make producing correct cynghanedd in English a challenge. Mererid Hopwood’s excellent *Singing in Chains* demonstrates that it is entirely possible to teach the basic principles of cynghanedd without using Welsh as the medium of instruction. There is no reason why the unique and remarkable traditions of Welsh literature should not be studied by all pupils in Wales in as much depth as those of English literature.

In additional-language classrooms, where languages like French and German are taught, pupils might encounter (or, at higher levels, produce) translations of texts they have already studied in Welsh and/or English. Ideally, the Welsh additional-language curriculum should be expanded, supplementing the traditional European languages (including, in some schools, classical European languages like Latin and Greek) with languages commonly used by people who have immigrated to Britain, such as Polish, Romanian, Panjabi, and Urdu (*Language, England and Wales: Census 2021*, 2022).

Similarities and differences between Welsh/English and other languages should

be discussed. Knowledge of cognates, in particular, is a cross-linguistic strategy which can help with vocabulary acquisition. Additional-language cognates of a word in one's primary language are generally more easily learned than noncognates (De Groot & Van Hell, 2005). Some researchers argue that morphologically related words, including cross-linguistic cognates, are stored together in memory, meaning that learning a cognate in a new language involves adapting an entry already present in the mental lexicon rather than creating an entirely new one (De Groot & Van Hell, 2005). Cognate awareness in Welsh education could include raising awareness of lexical cognates (Ballinger et al., 2017), such as French {eglise} and Welsh {eglwys}. These might then be compared with the cognates German {Kirche}, and English {church}, prompting a discussion of the historical reasons for these relationships. In the Welsh context, teaching cognates also provides a foundation for critical language awareness tasks, considering the differences in social roles and status between languages which accounts for the kinds of Germanic-French-Latin triplets in English described in Chapter 1.iii.c, and the reduced acceptability of many English borrowings in formal Welsh.

This kind of metalinguistic discussion can be invaluable for students (Ballinger et al., 2017; Beacco et al., 2016; Castellotti et al., 2008). Helping them to understand why certain types of language are considered acceptable in some situations and not in others will benefit their register awareness and their confidence that they know how to use Welsh appropriately in different contexts. Amongst other things, discussing cognates would increase pupils' awareness that lexical borrowing is very frequent amongst all of these languages, and that Welsh contains borrowings from Latin and French as well as from English, helping to balance perceptions that Welsh solely borrows (or is forced into using) English-origin ELUs. In turn, this will help establish the fact that lexical borrowing in Welsh (as in English) is a complex process, and is not merely a sign of the language's political and social subordination. Drawing parallels with the status of loans from different languages in English may further help to normalize the Welsh situation, reassuring students that the unacceptability of English-origin ELUs in certain registers is not a matter of denigrating English, but rather is an example of a process of lexical differentiation which also takes place in other languages.

Alongside awareness of cognates, awareness of patterns of derivational morphology in both primary and additional languages has been shown to aid pupils' vocabulary learning and reading comprehension (Baumann et al., 2003; Choi, 2015; Morin, 2003; Templeton, 2012). Indeed, readers with limited vocabulary may rely

heavily on morphological information for comprehension, since they are less able to understand and make use of contextual information from the surrounding text (Raudszus et al., 2021). Primary-language morphological awareness may also facilitate morphological awareness in additional languages (Choi, 2015; Ramírez et al., 2013). Teaching morphology in both Welsh and English, as well as in any additional languages, would therefore be a valuable addition to Language classes, as well as providing an opportunity to discuss the morphological assimilation of loanwords in both languages.

Integrated Language classes might include an element of translation, not for its own sake, but to facilitate cross-linguistic awareness (Ballinger et al., 2017; Christofferson, 2014). Translation exercises could help draw attention to asymmetries in bilingual pupils' competences, allowing teachers to target teaching to those areas which need work, and to supply pupils with linguistic resources which their translations show them to possess in one language but not in the other. Learning how to translate effectively includes learning when not to translate literally from one language to the other. Awareness of this issue should help pupils avoid one of the most common bugbears of Cymrophone language experts: the use of calques and the direct transference of English syntax into Welsh (C. Williams, 1999; R. H. Williams, 2018).

Pupils' on-line linguistic processing and production skills might also be developed with interpreting tasks, where one pupil speaks in one language and another provides a simultaneous interpretation. As far as I am aware, there is very little literature on the use of interpretation as a classroom task, perhaps because the skill is considered too difficult for young learners. Nonetheless, short interpretation tasks with older pupils, after the skill of asynchronous translation has been well established, might well help their oral skills in both languages. It would be valuable to see research trialling such tasks.

Of course, translation and interpretation are also valuable skills in themselves, ones which can be transferred to other languages, and which might prove useful in future employment. A related skill, recommended for use in plurilingual education, is intercomprehension, where pupils are encouraged to try and interpret texts in unknown languages which are closely related to their primary language(s) (Beacco et al., 2016; Lenz & Berthele, 2010). In Welsh-medium education, for example, pupils might attempt to make sense of texts in Cornish, or in Breton (particularly if they also have some knowledge of French). Being able to read authentic documents in other languages is potentially a very valuable skill, which need not be paralleled by

the ability to speak or write in that language (Lenz & Berthele, 2010). Pupils can also be encouraged to attempt intercomprehension together, especially when some of them have knowledge of languages other than English and Welsh.

Intercomprehension increases linguistic sensitivity and the development of communicative strategies Beacco et al. (2016). Enhancing pupils' ability to communicate effectively and to manage misunderstanding and communication barriers should be of benefit throughout their educational career, and beyond.

Returning to translation, this need not only be a task in itself but could also provide interesting objects of study. Students might compare and contrast different translations into one language of a text in another, and then be asked to make their own translation, perhaps into a specific style or variety. They might compare and contrast two translations of the same text, for example a translation of a poem which retains poetic features such as rhyme and meter, and a second translation into prose. The Welsh poet Gwyneth Lewis's experiences of translating her own Welsh poems into English ((e.g. G. Lewis, 1996)) would add a valuable metalinguistic perspective, and encourage students to think about their own use of language and how translation between languages or varieties may be an expression of different aspects of their personal experience.

Students might also study dialect literature, perhaps studying the same story told in different dialects of English, and then writing the story in their own variety. Orthographic switching of the sort seen in the data collected for this study could be used as resources for inspiring dialect-writing tasks, and would also open up discussions about the relationship between speech and writing, showing how different languages deploy orthography (or orthographies) in different ways (Castellotti et al., 2008). Students might be encouraged to consider the widespread use of the Latin alphabet for European languages, and to ask whether the current form of the alphabet used for Welsh could be improved. The orthographies of other languages should be investigated to provide inspiration. Representation of Welsh phonemes might, for example, be improved by following Icelandic and using the grapheme <ð> to represent /ð/, rather than the current Welsh convention of using <dd>. The experimental Welsh alphabet created by design student Carl Edwards in 2019 (*Road signs: Man designs new Welsh alphabet*, 2019), which introduces new characters to replace digraphs such as <dd> and <ll>, would also be an excellent prompt, and a project to design new alternatives might be co-ordinated across the Language syllabus and the Design and Technology syllabus. Other alphabets and writing systems should also be explored, particularly any which are known to pupils

in the class or to the teacher. The use of language on social media could also be studied, including practices such as the romanization of non-Latin-alphabet languages, with exploration of how new technologies facilitate or hinder different orthographic representations of languages.

Texts from social media would, in general, provide relatable material for language-awareness activities. The use of Welsh-English code-switching on social media could be studied, contextualized by examples from other languages. Pupils should engage in activities which lead them to recognize the different functions performed by different types and styles of code-switching, such as those explored in the previous chapter. Pupils who have other languages in their repertoires might also be encouraged to give examples of code-switching in those languages, and to help their classmates interpret them. Discussing social media language would be an appropriate context to introduce issues of dialect, particularly the well-documented Wenglish (R. E. Lewis, 2008). Students might compare and contrast tweets about a particular event in both Welsh and English, or in Welsh English and other varieties. Comparison of standard and non-standard varieties of both languages would lead naturally to discussions about language purism. The row about football pundit John Hartson, mentioned in Chapter 3.iv.a, might be considered alongside texts mocking or complaining about the language of Anglophone footballers and pundits. This would be a constructive opportunity for students to express their own opinions about language, to consider how context of use may change our beliefs about acceptability, and to become aware of the classism, racism, sexism, and ableism which often inform linguistic prescriptivism and purism. It would also open up a discussion about the social status of English and Welsh, and how this affects purist discourses in the two languages.

Curriculum planning must account for everything from national and international policy and practice, to the individual linguistic practices and competences of teachers and students (Beacco et al., 2016). Curricular coherence and learning efficiency can be maximized by taking a broad, integrational view of all aspects of schooling, and by co-ordinating school subjects so that competences can be transferred between them wherever possible (Beacco et al., 2016). Translation activities, for example, need not be confined to the Language classroom. The Welsh curriculum should include lessons in Welsh and British history. A module about Welsh political history might, for example, include reactions to the investiture of Charles Windsor in 1969, and could include reading tweets about the role of Prince of Wales which take different perspectives on the issue. The texts could be

introduced in the Language classroom and analysed there, ensuring that all pupils understand the texts in both languages, before being used as primary sources in the History classroom. With appropriate support, such activities could use original texts in both languages, regardless of the school's medium of education. Opportunities to use different varieties of English and Welsh should also be sought, allowing pupils to experience both languages in a variety of styles and contexts. Practical subjects such as physical education and art, for example, are particularly compatible “avec des formes légères et ludiques d'enseignement bilingue”⁵ (Castellotti et al., 2008, p. 20).

Throughout their education, Welsh pupils should be presented with examples of skilful, artistic code-switching Christofferson (2014). This need not be restricted to Welsh-English code-switching, but should instead aim to show how code-switching is employed in a variety of contexts worldwide, to normalize its use. Visual artworks, films and television, music, and plays which contain code-switching or which are in more than one language should be used extensively in subjects such as Music, Drama, Art, and Media Studies, as well as in the Language and Literature classrooms. Songs which contain lyrics in both English and Welsh might be discussed in Language and/or Music classes, and compared with mixed-language songs from other cultures.

Schools must accommodate the realities of bilingualism in their institutional attitudes towards language, celebrating code-switching as a natural and creative way of being bilingual, while encouraging awareness of the preference against switching in formal contexts, rather than setting up an artificial monolingualism as the only way to use Welsh (Musk, 2010). The emphasis in Welsh education should be on developing skills and strategies for successful communication (Musk, 2010) within and between languages and cultures, while negotiating ambiguity and uncertainty Castellotti et al. (2008), so that the pupil gains confidence in using her full repertoire appropriately in a variety of contexts.

Of course, the ideal of an integrated plurilingual curriculum could not be achieved in a single stroke. This does not, however, entail that plurilingualism cannot be implemented at all in Welsh education in its current form. Many of the strategies, activities, and approaches described here, as well as many others which implement plurilingual ideals, could be incorporated into the current curriculum. Others would require full-scale systemic changes. The development of a plurilingual educational policy should proceed realistically and incrementally, taking small steps, and setting

⁵ *with light and playful forms of bilingual teaching*

goals which are achievable given the resources available, whilst maintaining creativity and ambition (Castellotti et al., 2008).

7.iv.b.8 Plurilingualism for grown-ups: adult education

As I have argued throughout this study, replacing purist language attitudes with inclusive, creative, and (above all) realistic ones in Wales is a crucial component of encouraging the confident use of Welsh throughout the country. The plurilingual strategies described above for formal schooling should also inform adult Welsh-language education, drawing on learners' knowledge of their primary languages, and raising their awareness of the Welshness, and the Welsh languageness, of Welsh varieties of English, to help them to incorporate Welsh as a valuable addition to an integrated linguistic repertoire.

Adults in Wales who are learning Welsh should be taught about the value and naturalness of code-switching and language mixing behaviours, and encouraged to observe how primary-language Cymrophones engage in these behaviours. Adults acquiring Welsh as an additional language should not be taught artificial varieties which avoid English-origin ELUs. Indeed, they should be encouraged to use established ELUs, both because (as discussed above) cognates, including loanwords, facilitate vocabulary acquisition, and because the use of ELUs is an integral feature of a target-like use of the language. Welsh teaching for adults could usefully include raising awareness of the complex sociolinguistic factors at play in Wales, including the harmful effect of stigmatizing language-mixing. This would be of particular benefit to adults who are learning because they have children or grandchildren in Welsh-medium education, who make up a significant proportion of learners of Welsh (Newcombe, 2007), enabling them to pass on healthy attitudes towards language-mixing to their children.

Like resources for schoolchildren, Welsh-language resources for adults must also avoid enregistering Welsh as exclusively the language of the *gwerin*, of white people, or of people born in Wales. Although corpus data showing how Welsh is used in contemporary contexts is extremely important for the development of learner resources, it is therefore also important to ensure that drawing on what is already said in Welsh does not prevent Welsh-language education from providing learners with the resources to express novel ideas and enregister Cymrophone identities beyond the traditional ones. In the production of the *Geirfan* corpus-based wordlists for learners of Welsh, for example, two lexemes referring to Christian places of worship – {*eglwys*} (“church”) and {*capel*} (“chapel”) – were found to be particularly frequent

in contemporary Welsh. The lexemes {synagog} (“synagogue”), {teml} (“temple”), and {mosg} (“mosque”) were therefore also added, despite not being of high frequency in the corpus (Knight et al., 2022), redressing the overrepresentation of Christian religious lexis in the corpus and expanding learners’ linguistic resources beyond those expressing a restricted idea of traditional Cymrophony.

7.v Conclusions

This study argues for a plurilingual approach to language in Welsh education. The hardest part of such an approach might be persuading Cymrophone pupils, parents, and educators of the benefits of integrating English with Welsh in Welsh-medium schools. Indeed, there are valid arguments in favour of maintaining monolingualism in minoritized-language immersion education, where the dominant language is well-supported in the community and there is a severe imbalance in status between the languages (Ballinger et al., 2017; Juan-Garau & Lyster, 2019). Nonetheless, the primary aim of Welsh-medium education should not be to revitalize the Welsh language, or to produce ideal, monolingual Cymrophones, but to support pupils in their personal, social, and academic development. Doing so requires educating the student as a whole person, and celebrating their entire linguistic identity, not forcing them into the system’s preferred mould.

Paradoxically, a plurilingual approach might help students develop a more expansive and secure Cymrophony than is possible in a context of prescribed monolingualism. The data analysis presented in Chapter 6 shows that orthographic adaptation in Cymrophone discourse is a way of adding social, cultural, political, and interpersonal meaning to the mixing of languages. It is a way for Cymrophones to display Welshness, even when the language of the syntax and lexis is predominantly English. It also permits them to deploy both their languages skilfully to make meanings that are greater than anything that could be achieved by either language alone. Making room for colloquial and transgressive languaging in educational spaces offers pupils the chance to enregister the difference between informal, formal, colloquial, academic, and other registers, without which their ability to use the language as effectively as they can use English may suffer (Price, 2020). In the absence of such opportunities, many will acquire a semiotically indistinct, educational

variety of Welsh (Durham & Morris, 2016; Hornsby, 2014), which is “rhy llafar”⁶ (CBAC, 2017) for written use, yet “rhy ffurfiol”⁷ (CBAC, 2019) for oral use.

It is essential that pupils are given the opportunity to acquire the Welsh language in a way that makes it relevant to their own lives E. M. Thomas et al. (2014b), and that integrates their full linguistic and cultural repertoire. As this study and that of Durham (2022) show, Welsh people’s language on social media makes use of a variety of mixed-language and non-standard practices, with even individual graphemes having a role to play in the semiotic performance of Welshness. “Defnydd o eiriau Saesneg”⁸ (CBAC, 2019) in informal registers, far from being a sign of linguistic insufficiency, must be understood as a valid and widespread mode of performing modern Cymrophony.

I have argued here for educational policy which places the development of the pupil’s confidence and competence as a greater priority than language maintenance, but support for the former may ultimately translate into promotion of the latter. Smith-Christmas (2017), for example, recommends paying explicit attention to the ideological and affective associations of the minoritized language in order to increase the effectiveness of language maintenance efforts. Plurilingualism, by promoting the value of all languages and varieties in the individual’s repertoire, offers some potential for helping pupils throughout Wales to view Welsh and English on equal terms, and to understand the value of both languages, used separately or mixed together, in the formation of a distinctive Welsh identity. Its potential therefore extends beyond the classroom into the community, encouraging pupils to develop confident linguistic identities and to achieve academically, and strengthening their affective bonds with Welsh, which make its continued use beyond school more likely.

7.v.a Education and the right to a Welsh identity

Although this study has encompassed language use and attitudes in multiple contexts, from official political statements to anonymous social-media posts, the educational context has emerged as the primary focus in these conclusions. Welsh-medium education, as we have seen, plays a central role in the development and perpetuation of language ideologies, and in the formation of Cymrophones’ understanding of their

⁶ *too oral*

⁷ *An overly formal register.*

⁸ *Use of English words.*

place in the linguistic landscape. In its position as the common social and developmental context of young Welsh people, “il reste à l’école cette fonction et ce pouvoir de donner droit de cité à cette pluralité, de la réfléchir, d’en faire prendre pleine conscience, notamment par le biais, au contact et aussi pour le bénéfice de la langue de scolarisation”⁹ (Castellotti et al., 2008, p. 17). Welsh-medium schools are key milieux of language socialization. Changes in educational policy and practice, away from punitive monolingualism and deficit models of language mixing, and towards plurilingualism and the celebration of diverse linguistic competence, are needed if all children in Welsh-medium education are to thrive.

This is not to say that Welsh-medium schooling should be the sole locus of change, however. The discussion above has also suggested ways in which greater knowledge of and awareness of Welsh could be introduced into the curriculum for English-medium schooling. The policy shift against the “second-language” label for Welsh as a subject in English-medium schooling, aiming to provide a “continuum of learning” (*One language*, 2013, p. 2) for all Welsh pupils, represents a step towards a plurilingual approach. I have recommended above the extension of this approach into a curriculum which is the same for all pupils in Wales, with Language and Literature taught from the same syllabus, and Welsh texts provided either in English translation or in simplified versions where necessary. English-medium and Welsh-medium pupils in Wales should not feel that they exist in different ideological worlds (Lee, 2015), separated by a chasm of cultural knowledge and experience which is caused by linguistic difference. Rather, they should be educated in substantially the same content, with adjustments made to the language in which the content is presented. English-medium pupils in Wales have as much right, and as much need, to acquire a deep and thorough knowledge of Welsh histories, Welsh languages, and Welsh cultures. Welshness, whose highly contested and troubled nature was explored in Chapter 3, can only be made truly expansive and inclusive when every pupil – indeed, every person – in Wales is treated as the inheritor of the full breadth of Welsh culture. Positioning non-Cymrophony as a barrier to accessing this heritage guarantees the perpetuation of the Welsh linguistic divide. Bilingualism and “the focus on Welsh identity” are a particular attraction of the education system in Wales, as reported by Donaldson (2015, p. 15), yet the current system offers these benefits

⁹ *the school still retains the function and the power to legitimize plurality, to reflect it, to raise awareness of it, particularly by means of, in contact with, and also for the benefit of, the language of schooling*

much less consistently and thoroughly to pupils in English-medium schools.

A constructive, liberal attitude in policy and education will, I have argued, reduce language anxiety, increase confidence in the use of Welsh, and encourage its use outside educational contexts. Ultimately, Welsh-English bilinguals must be permitted personal ownership of both their languages, enregistering them as expansively as possible, in order to feel that their individual identity can be successfully expressed in Welsh, in English, or in a mixture of the two. This can best be achieved through educational programmes which forge connections between the languages, enregistering both equally as “school language” and as “social language”, rather than limiting Welsh to the academic domain and making English (including Wenglish) into a forbidden code used secretly and rebelliously with friends. Rather than seeing language as a zero-sum game, where every English element is a lost opportunity for Welsh, we must instead understand all of a Cymrophone’s linguistic behaviours as inherently Cymrophone, just as they are also inherently Anglophone. Bilinguals cannot be expected to become monolinguals, although they can be taught strategies for producing unmixed varieties of their languages in appropriate contexts. The production of unmixed varieties cannot be the only aim of Welsh instruction, however, and such varieties should not be reified as the only correct, acceptable, or moral way to use a language.

7.v.b Iaith Gymreig, iaith Gymraeg: the Welsh-linguageness of the English language

Code-switching “symbolizes identities beyond the linguistic fact” (Auer, 2005, p. 404). This study has argued that a connotative semiotic framework of language explains how Welshness permeates the Welsh language down to its sub-lexical, sub-meaningful orthographic units. Departing from de Saussure and Hjelmslev, I argue that graphemes have their own, unique semiotic content, and that writing can therefore convey meaning in ways that cannot be replicated in speech. This framework leads to the conclusion that codeswitching, traditionally thought of as operating on units no smaller than the lexeme or the morpheme, can in fact operate at the orthographic level. The diffusion of contextual meaning through the entire linguistic system is what allows orthographic choices to express Cymrophony even when the lexis is English. Understanding how the language practices of Cymrophones are enregistered with Welshness encourages an inclusive attitude towards the full linguistic identities of Welsh-English bilinguals, offering the insight

that Cymrophones do not cease to be Cymrophones when they use English.

Indeed, the Welshness of English-origin words, phrases, and sentences permeates Welsh people's linguistic behaviour, whether the users are primary-language Cymrophones, additional-language Cymrophones, or non-Cymrophones. As Mererid Hopwood reminds us, “nid Saesneg Lloegr yw ein Saesneg ni”¹⁰ (Hopwood, 2003, p. 59). The intimate relationship between Welsh and the English of Wales means that there is really no such thing as a Welshless Welsh person Hopwood (2019). Welsh English is a Welsh language, its syntax, vocabulary, phonology, and idioms imbued with the history of contact with Welsh. When an English word is made Welsh, it represents the equality of the two languages which contributed to its construction. Welsh and English, in collaboration, can construct meanings imbued with various types of both Welshness and Englishness; meanings that neither language can express alone.

Giving all Welsh pupils “a fair and equal opportunity to achieve in Cymraeg, regardless of which type of school they attend” (*Qualifications Wales*, 2022), and achieving the aims of the *Cymraeg 2050* strategy (*Cymraeg 2050*, 2017), require systematic changes, particularly in Welsh education. Despite vast differences of social reach and status, English and Welsh may co-exist as equals in the bilingual's linguistic competence, but only if they are permitted to do so freely. Paradoxically, the flourishing of Welsh in Wales may depend on recognizing and honouring its relationship with English, as well as on celebrating acts of linguistic cross-fertilization which make English not only a “iaith Gymreig”, but also a “iaith Gymraeg”¹¹.

¹⁰ “our English is not the English of England”

¹¹ That is, not only a language of Wales, but also a language inseparable from, and imbued with, the Welsh language.

Bibliography

Primary sources

Fiction and poetry

Amis, K. (1986). *The old devils*. Hutchinson.

Popular non-fiction

Davies, J. (2014). *The Welsh language: A history*. University of Wales Press.

James, E., & Robins, J. (2018). *The Holy Vible*. Orion.

Meredith, J. (2010). *Yr hwn ydwyf*. Y Lolfa.

Owen, H. D. (2015). *Sesiwn yng Nghymru*. Y Lolfa.

Rhys, D. (2009). *Hiwmor y Cofi*. Y Lolfa.

Williams, C. (2002). *Sugar and slate*. Parthian.

Journalism, essays, and opinion pieces

“@BeiroCoch”. (2015). *Cymraeg cywir?* BBC Cymru Fyw. Retrieved October 4, 2015, from <https://www.bbc.co.uk/cymrufyw/34269623>

‘Caervallwch’. (1834). Geiriau Cymraeg. *Greal y Bedyddwyr*, 8(89), 141.

‘Gweinidog Heb Ofal’. (1903). Meddyliau ar bregethu. *Y Diwygiwr*, 60(809), 117–121.

Anffyddiaeth. (2012). *Hiliaeth amrwd ar S4C*. Retrieved November 19, 2012, from <http://anffyddiaeth.blogspot.com/2012/11/hiliaeth-amrwd-ar-s4c.html>

Anon. (1851). To our correspondents. *Yr Addysgydd*, 1(1).

ap Iwan, E. [A. J. (1889). Plicio gwallt yr hanner Cymry. *Y Geninen*, 7, 116–119, 169–176.

BBC Cymru Fyw. (2019). *Meic Stevens yn gwadu sylwadau hiliol*. Retrieved July 16, 2019, from <https://www.bbc.co.uk/cymrufyw/49008186>

- BBC Cymru Fyw. (2022). *Meinir Pierce Jones yn ennill Gwobr Goffa Daniel Owen*. Retrieved August 2, 2022, from <https://www.bbc.co.uk/cymrufyw/62392799>
- BBC News. (2016). *Lack of Welsh language school textbooks 'unfair'*. Retrieved October 3, 2016, from <https://www.bbc.co.uk/news/uk-wales-politics-37545993>
- BBC One. (2022). *Our Lives series 6: Searching for my other mam*. Retrieved September 2, 2022, from <https://www.bbc.co.uk/iplayer/episode/m001bn33/our-lives-series-6-searching-for-my-other-mam>
- Brooks, S. (2010, August 3). *Welsh-medium education is not all-white*. Retrieved February 1, 2019, from <https://www.iwa.wales/click/2010/08/ethnic-cleansing-in-civic-wales/>
- Ceiro, G. (2021). *'Dewch 'mlaen canol y cae??'* Retrieved October 14, 2021, from <https://golwg.360.cymru/cylchgrawn/2071699-dewch-mlaen-canol>
- Coslett, R. L. (2019). *Cymru am byth! – How speaking Welsh became cool*. The Guardian. Retrieved June 13, 2019, from <https://www.theguardian.com/uk-news/2019/jun/13/welsh-speaking-bilingual-tv-thrillers-cool>
- Critically Speaking. (2020). *Mali Ann Rees*. Retrieved February 23, 2020, from <https://criticallyspeaking.co.uk/2020/02/23/mali-ann-rees-2/>
- Davies, D. (2020). *Hiliaeth 'systemig' yn rhan annatod o Gymru*. BBC Cymru Fyw. Retrieved June 18, 2020, from <https://www.bbc.co.uk/cymrufyw/53089768>
- Dwyfor, G. (2021). *John Hartson a'r plismyn iaith*. Retrieved October 13, 2021, from <https://golwg.360.cymru/cylchgrawn/2071454-john-hartson-plismyn-iaith>
- Ellis, L. (2024). *'Cymry di-Gymraeg yn teimlo fel estroniaid'*. Retrieved January 23, 2015, from <https://golwg.360.cymru/newyddion/cymru/175380-cymry-di-gymraeg-yn-teimlo-fel-estroniaid>
- Evans, D. E. (1897). *Y Gymraeg â geiriau tramor*. *Y Cerddor*, 9(103), 69–70.
- Gallagher, A. (2021). *Datblygu frontman David R. Edwards has died, aged 56*. Retrieved June 23, 2021, from <https://www.nme.com/news/music/datblygu-frontman-david-r-edwards-dies-aged-56-2975868>
- Golwg360. (2012). *Eisteddfod y Ffermwyr Ifainc: Ymateb chwyrn*. Retrieved November 19, 2012, from <https://golwg.360.cymru/newyddion/cymru/92169-eisteddfod-y-ffermwyr-ifainc-ymateb-chwyrn>
- Golwg360. (2018). *Cyfle i 'agor y drysau' i Gymry di-Gymraeg*. Retrieved July 14, 2018, from <https://golwg.360.cymru/newyddion/cymru/524178-cyfle-agor-drysau-gymry-gymraeg>

- Golwg360. (2019). *Meic Stevens a sylw am Fwslimiaid: 'dw i ddim yn hiliol'*. Retrieved July 16, 2019, from <https://golwg.360.cymru/newyddion/cymru/549871-meic-stevens-sylw-fwslimiaid-ddim-hiliol>
- Hastings, C. (2000). *BBC staff 'suffered for not being welsh'*. The Telegraph. Retrieved July 6, 2000, from <https://www.telegraph.co.uk/news/uknews/1346799/BBC-staff-suffered-for-not-being-Welsh.html>
- Hopwood, M. (2003). Llenwi'r bylchau. *Transactions of the Honourable Society of Cymmrodorion*, 2002, 9, 57–73.
- Hopwood, M. (2019). *Doethineb iaith*. Retrieved August 2, 2019, from <https://pedwargwynt.cymru/dadansoddi/doethineb-iaith>
- Hughes, V. (2024). *Cwrs y byd*—RS. Retrieved May 2013, from <https://barn.cymru/node/1425>
- Huws, S. (2021). *Procio'r plismyn iaith*. Retrieved November 3, 2021, from <https://golwg.360.cymru/cylchgrawn/2067186-procio-plismyn-iaith>
- Jones, D. (2018). Hiliaeth a Chymreictod. *O'r Pedwar Gwynt*. <https://pedwargwynt.cymru/adolygu/hiliaeth-a-chymreictod>
- Jones, E. D. (1933). Iaith ac ymadrodd. *Y Llenor*, 12(1), 45–53.
- Jones, E. L. (2007). [review of the book *Amen, Dyn Pren*, by Gwiliym Tudur and Mair E. Jones]. *Y Traethodydd*, CLXIII [CLXII](681 [680]), 59–60.
- Jones, G. H. (1926). Cadwraeth yr iaith. *The transactions of the Honourable Society of Cymmrodorion*, 1924–1925, 2–29.
- Jones, G. (1926). Cadwraeth yr iaith. *The transactions of the Honourable Society of Cymmrodorion*, 1924–1925, 79–87.
- Jones, M. D. (1888). Y Gymraeg fel iaith masnach. *Y Celt*, 7.
- Jones, R. W. *Methiant prifysgolion Cymru: Darlith yr Eisteddfod Genedlaethol*. Casnewydd, 2004.
- Jones, S. (2020). *Araith llywydd y dydd: Seren Jones*. BBC Cymru Fyw. Retrieved August 1, 2020, from <https://www.bbc.co.uk/cymrufyw/53569056>
- Jones, S. (1938). Speech sounds and their bearing on the linguistic problem in Wales. *Transactions of the Honourable Society of Cymmrodorion*, 1937, 227–239.
- Lewis, G. (1996). Whose coat is that jacket? Whose hat is that cap? *Columbia: A Journal of Literature and Art*, 27, 58–68.
- Lewis, G. (2006). *Caught between two cultures*. Retrieved November 18, 2006, from <https://www.theguardian.com/lifeandstyle/2006/nov/18/familyandrelationships.family1>

- Lewis, R. (2011). *How i tried to be one of the boyos*. Retrieved August 12, 2011, from <https://www.dailymail.co.uk/home/books/article-2024591/Exploited-raped-controlled-punished-English-mans-ambition-Welsh.html>
- Llywelyn-Williams, A. (1940). Y Gymraeg ar y radio. *Y Llenor*, 19, 143–153.
- McCarthy, J. (n.d.). *Someone told Leanne Wood to 'moderate' her Welsh accent so people 'need not squirm'*. Retrieved February 23, 2017, from <https://www.walesonline.co.uk/news/politics/leanne-wood-advised-moderate-accent-12649075>
- Mears, T. (n.d.). *Leanne Wood fires back at critics after 'getting stick' for her Valleys accent*. Retrieved November 7, 2016, from <https://www.walesonline.co.uk/news/local-news/leanne-wood-fires-back-critics-12138782>
- Mikkelsen, B. (2002). *Snopes fact check: Bush and French word for entrepreneur*. Retrieved July 29, 2002, from <https://www.snopes.com/fact-check/french-lesson/>
- Misstear, R. (2012). *'Children admonished for using English in Welsh schools'—parents' claim*. Wales Online. Retrieved November 15, 2012, from <https://www.walesonline.co.uk/news/wales-news/children-admonished-using-english-welsh-2017412>
- Môn, D. (1889). Yr Eisteddfod Genhedlaethol: A ydyw yn werth y draul o'i chynnal? *Y Geninen*, 7(1), 7–12.
- Morris-Jones, J. (1926). Cadwraeth yr iaith. *The transactions of the Honourable Society of Cymmrodorion*, 1924–1925, 29–33.
- Mwyn, R. (2023). *'Plismyn Iaith' fel Brexiteers fydd byth yn fodlon*. Retrieved September 6, 2023, from <https://golwg.360.cymru/cylchgrawn/2128777-plismyn-iaith-brexiteers-fydd-byth-fodlon>
- Onllwyn, H. (2022). *Herio'r heddlu iaith*. Retrieved August 24, 2022, from <https://golwg.360.cymru/cylchgrawn/2103329-herio-heddlu-iaith>
- Orcutt, M. (2013). *The many tongues of Twitter*. Retrieved 2013, from <https://www.technologyreview.com/s/522376/the-many-tongues-of-twitter/>
- Oremus, W. (n.d.). *Twitter's new order*. Retrieved March 5, 2017, from https://www.slate.com/articles/technology/cover_story/2017/03/twitter_s_timeline_algorithm_and_its_effect_on_us_explained.html
- Owain, O. (1964a). Y Fro Gymraeg. *Tafod y Ddraig*, 4, 2.
- Owain, O. (1964b). Oni enillir y Fro Gymraeg... *Y Cymro*.

- Owain, O. (1977). Debygwn i: Safbwynt personol ar wyddoniaeth heddiw. *Y Gwyddonydd*, 15(1), 127–128.
- Owen, G., & Lewis, J. (1799). A history of Pembrokeshire. *The Cambrian Register for 1796*, 2, 53–230.
- Pick, R. (n.d.). *Twitter is testing timelines that aren't in chronological order*. Retrieved December 8, 2015, from <https://www.vice.com/en/article/vv7g9d/twitter-is-testing-timelines-that-arent-in-chronological-order>
- Rees, I. W. (2015, October 4). *Ffordd iawn o siarad Cymraeg?* <https://www.bbc.co.uk/cymrufyw/34325522>
- Road signs: Man designs new Welsh alphabet*. (2019). Retrieved August 30, 2019, from <https://www.bbc.com/news/uk-wales-49521620>
- Stapley, N. (2008, March 9). *Broadcasting the bards*. https://www.transdiffusion.org/2008/03/09/broadcasting_th
- Traynor, L. (2012). 'Speak English and you'll be punished': Strict rules for pupils at Welsh secondary school. *The Mirror*. Retrieved November 24, 2012, from <https://www.mirror.co.uk/news/uk-news/pupils-caught-speaking-english-will-be-punished-1453709>
- Wightwick, A. (n.d.). *Major changes to Welsh language and literature GCSEs*. Retrieved February 16, 2022, from <https://www.walesonline.co.uk/news/education/major-changes-welsh-language-literature-23104143>
- Williams, R. (1985). Community. *London review of books*, 7(1).
- Wood, L. (n.d.). *Leanne Wood: Challenge me, but don't try to use my working class accent to shut me down*. Retrieved April 9, 2018, from <https://inews.co.uk/opinion/leanne-wood-challenge-me-but-dont-try-to-use-my-working-class-accent-to-shut-me-down-142536>
- Wright, M. (n.d.). *Eurovision 2017: BBC forced to defend 'absolutely SHAMEFUL' tweet about Welsh language*. Retrieved May 15, 2017, from <https://www.express.co.uk/showbiz/tv-radio/804489/Eurovision-2017-BBC-forced-to-defend-shameful-tweet-about-Welsh-language>
- Ymateb cryf i honiad fod 'hanner y geiriau a ddefnyddia John Hartson ar S4C yn Saesneg'*. (2021). Retrieved October 11, 2021, from <https://golwg.360.cymru/newyddion/cymru/2071326-ymateb-cryf-honiad-hanner-geiriau-ddefnyddia-john>
- Youde, M. (2018). *No, anti-Welsh bigotry is not the last acceptable form of racism*. *Nation.Cymru*. Retrieved April 17, 2018, from

<https://nation.cymru/opinion/no-anti-welsh-bigotry-is-not-the-last-acceptable-form-of-racism/>

Websites

- ‘@kurrik’. (2013). *Twitter Developer Platform Blog: Introducing new metadata for Tweets*. Retrieved February 13, 2013, from https://blog.twitter.com/developer/en_us/a/2013/introducing-new-metadata-for-tweets#:~:text=The%20new%20lang%20attribute%20specifies,no%20language%20could%20be%20detected.
- Animal Farm*. (2018). Retrieved August 5, 2018, from <https://www.blaketheater.com/production/animal-farm/>
- BBC Cymru Fyw. (2020). *Ymddeol o'r Talwrn ar ôl 40 mlynedd*. Retrieved March 4, 2020, from <https://www.bbc.com/cymrufyw/51662071>
- Bethel English Baptist Chapel*. (2010). Retrieved 2010, from <https://coflein.gov.uk/en/site/9826/>
- Comisiynydd y Gymraeg. (n.d.). *Amdanom ni*. <https://www.comisiynyddygyymraeg.cymru/>
- Dafis, C. (2021). *Siaradwyr Cymraeg y Senedd yn optio am y Saesneg*. Retrieved May 20, 2021, from <https://golwg.360.cymru/cylchgrawn/2051154-siaradwyr-cymraeg-senedd-optio-saesneg>
- ESRC. (2019). *Internet-mediated research*. Retrieved September 27, 2019, from <https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/frequently-raised-topics/internet-mediated-research/>
- European Youth Parliament*. (n.d.). <https://eyp.org/>
- Ffrancis, F. (2015). *Welsh-medium education for all, or just for the lucky ones?* Retrieved January 20, 2015, from <https://www.iwa.wales/agenda/2015/01/welsh-medium-education-for-all-or-just-for-the-lucky-ones/>
- Jones, E. L. (2018). *Datganiad pellach gan Eifion Lloyd Jones*. Eisteddfod Genedlaethol Cymru. Retrieved August 12, 2018, from <https://eisteddfod.cymru/datganiad-pellach-gan-eifion-lloyd-jones>
- Little, D., Goullier, F., & Hughes, G. (2011). *The European Language Portfolio: The story so far (1991-2011)*. Retrieved February 2011, from <https://www.coe.int/en/web/portfolio/the-european-language-portfolio-the-story-so-far-1991-2011-executive-summary>

- Llwyd, D. (2010). *Dewi Llwyd ag Elinor Jones*. Retrieved November 21, 2010, from https://www.bbc.co.uk/webarchive/https%3A%2F%2Fwww.bbc.co.uk%2Fblogs%2Fradiocymru%2F2010%2F11%2Fdewi_llwyd_ag_elinor_jones.shtml
- McGuire, A. (2014). *Tell Tale are back with Animal Farm*. Retrieved October 20, 2014, from <https://liverpoolculturalchampions.wordpress.com/2014/10/20/tell-tale-are-back-with-animal-farm/>
- Morus, I. (2016). *Y Dydd Olaf (The Last Day)*. Retrieved 2016, from <https://web.archive.org/web/20160504125257/http://unsettlingscientificstories.co.uk/y-dydd-olaf>
- National Records of Scotland. (2013). *Scotland's Census*. <https://www.scotlandscensus.gov.uk>
- Penalltau, Y. G. (n.d.). *Cymro - cymraes yr wythnos*. Retrieved September 15, 2022, from <https://www.ysgolpenalltau.com/news/detail/cymrocymraes-yr-wythnos-1/>
- Pontybrenin, Y. G. G. (n.d.). *Cymro - cymraes yr wythnos / welsh speaker of the week*. Retrieved September 15, 2022, from <https://www.yggpontybrenin.com/cymro-cymraes-yr-wythnos-welsh-speaker-of-the-we-8/>
- Safle Owain Owain. (n.d.). Retrieved 2008, from www.owainowain.com
- Senedd. (2022). *Official languages*. Retrieved January 18, 2022, from <https://senedd.wales/commission/senedd-commission-policy/official-languages/>
- Tornes, A., & Trujillo, L. (2021). *Twitter Developer Platform Blog: Enabling the future of academic research with the Twitter API*. Retrieved January 26, 2021, from https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api
- Twitter. (2019). *Developer agreement and policy*. Retrieved 2019, from <https://developer.twitter.com/en/developer-terms/agreement-and-policy>
- Wikimedia Statistics. (n.d.). <https://stats.wikimedia.org/>
- Y Talwrn. (n.d.). <https://www.bbc.co.uk/programmes/b007sczc>

Official reports and records

- Arffman, I. (2012). *Translating international achievement tests: Translators' view* (No. 44). Finnish Institute for Educational Research.

- Beacco, J.-C., Byram, M., Cavalli, M., Coste, D., Cuenat, M. E., Goullier, F., & Panthier, J. (2016, August). *Guide for the development and implementation of curricula for plurilingual and intercultural education*. Council of Europe.
- Castellotti, V., Coste, D., & Duverger, J. (2008, March). *Propositions pour une éducation au plurilinguisme en contexte scolaire*. Association pour le Développement de l'Enseignement Bi-/plurilingue.
- CBAC. (2017). *TGAU adroddiadau arholwyr: TGAU (newydd) Cymraeg Iaith Haf 2017*. <https://www.wjec.co.uk/media/y0tdb53w/tgau-cymraeg-iaith-adroddiad-haf-2017.pdf>
- CBAC. (2019). *TGAU adroddiadau arholwyr: Cymraeg Iaith Haf 2019*. <https://www.cbac.co.uk/umbraco/surface/blobstorage/download?nodeId=25610>
- CBAC. (2022). *TGAU adroddiadau arholwyr: Cymraeg Iaith Haf 2022*. <https://www.cbac.co.uk/umbraco/surface/blobstorage/download?nodeId=42236>
- Cymraeg 2050: A million welsh speakers* (No. WG31851). (2017). Welsh Government. <https://gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf>
- Cymraeg 2050: Changes to Welsh Language qualifications*. (2022, February). Qualifications Wales. Retrieved February 16, 2022, from <https://qualifications.wales/news-views/cymraeg-2050-changes-to-welsh-language-qualifications-eng/>
- Donaldson, G. (2015). *Successful futures* (No. WG23258). Welsh Government. <https://gov.wales/sites/default/files/publications/2018-03/successful-futures.pdf>
- Evas, J., Morris, J., & Whitmarsh, L. (2017, June). *Welsh language transmission and use in families* (No. 29/2017). Welsh Government.
- From linguistic diversity to plurilingual education: Guide for the development of language education policies in Europe (main version)*. (2007). Council of Europe: Language Policy Division.
- Hafferty, C. (2020). *What factors are linked to people speaking the Welsh language?* (No. 27/2020). Welsh Government. https://gov.wales/sites/default/files/statistics-and-research/2020-03/what-factors-are-linked-to-people-speaking-the-welsh-language_0.pdf
- Knowledge of language and life in the UK* (tech. rep. No. 29.0). (2024, July). United Kingdom Home Office.
- Language in England and Wales: 2011*. (n.d.). Retrieved March 4, 2013, from <https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/>

- [language/articles/languageinenglandandwales/2013-03-04#main-language-in-england-and-wales](#)
- Language, England and Wales: Census 2021 (Statistical bulletin). (2022, November). Office for National Statistics. Retrieved November 29, 2022, from <https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/bulletins/languageenglandandwales/census2021>
- Lenz, P., & Berthele, R. (2010, September). *Assessment in plurilingual and intercultural education*. Council of Europe: Language Policy Division.
- LinguaNi. (2020, December). *The Independent Review Report to re-define schools according to their Welsh-medium education provision* (No. WG41371).
- Morgan, N., Bajjada, T., Wise, C., & Bebb, H. (2022). *Welsh Government consultation – Welsh in English Medium schools and settings* (Analysis of Feedback / Consultation Responses). <https://gov.wales/sites/default/files/consultations/2022-10/draft-framework-for-welsh-in-english-medium-education-responses.pdf>
- Northern Ireland Statistics and Research Agency. (2012, December). *Census 2011: Key statistics for Northern Ireland*.
- Office for National Statistics. (2015). *2011 census: UK comparability* (Information Paper).
- One language for all: Review of Welsh second language at Key Stages 3 and 4*. (2013, September). Welsh Government. <https://dera.ioe.ac.uk/id/eprint/20135>
- Qualified for the future: The right choice for Wales*. (2022, March). Qualifications Wales.
- Ray, C. (2019). *Acceptable qualifications for knowledge of Welsh or Scottish Gaelic for naturalization*. Retrieved December 4, 2019, from https://www.whatdotheyknow.com/request/acceptable_qualifications_for_kn
- Record of Proceedings* (No. 2020-01-28). (2020). Senedd Cymru. <https://record.senedd.wales/Plenary/6077#C261315>
- Record of Proceedings* (No. 2021-07-13). (2021a). Senedd Cymru. <https://record.assembly.wales/Plenary/13951#C601555>
- Record of Proceedings* (No. 2021-12-08). (2021b). Senedd Cymru. <https://record.senedd.wales/Committee/12510#C394042>
- Record of Proceedings* (No. 2022-10-13). (2022a). Senedd Cymru. <https://record.senedd.wales/Committee/12989#C452819>
- Record of Proceedings* (No. 2022-09-28). (2022b). Senedd Cymru. <https://record.senedd.wales/Plenary/12977#C447642>

- Record of Proceedings* (No. 2023-07-05). (2023a). Senedd Cymru.
<https://record.senedd.wales/Plenary/13405#C517324>
- Record of Proceedings* (No. 2023-01-24). (2023b). Senedd Cymru.
<https://record.senedd.wales/Plenary/13189#C478803>
- Record of Proceedings* (No. 2024-06-25). (2024). Senedd Cymru.
<https://record.assembly.wales/Plenary/13951#C601555>
- Striking the right balance: Proposals for a Welsh Language Bill* (White Paper Consultation Document No. WG32353). (2017, August). Welsh Government. <https://gov.wales/sites/default/files/consultations/2018-02/170807-consultation-document-en.pdf>
- Welsh Arts Anti Racist Union. (2021). *Welsh arts anti racist union report*. Arts Council of Wales. <https://arts.wales/resources/welsh-arts-anti-racist-union-report>
- Welsh in education: Action plan 2017-21* (No. WG33270). (2017). Welsh Government.
- Welsh Language (Wales) measure* (No. nawm 1). (2011). National Assembly for Wales. <https://www.legislation.gov.uk/mwa/2011/1/contents/enacted>
- Welsh Language Act* (No. c. 66). (1967). U.K. Parliament.
<https://www.legislation.gov.uk/ukpga/1967/66/enacted>
- Welsh Language Act* (No. c. 38). (1993). U.K. Parliament.
<https://www.legislation.gov.uk/ukpga/1993/38/contents/enacted>
- Welsh Language Commissioner. (2020). *Recruitment: Good practice advice document*. Welsh Government.
<https://www.welshlanguagecommissioner.wales/media/0trjreze/20210923-dogfen-gyngor-arferion-da-recrwtio-saesneg.pdf>

Source code and electronic corpora

- Bangor Siarad corpus*. (n.d.). Retrieved June 29, 2020, from <http://bangortalk.org.uk/>
- Canolfan Bedwyr (Ed.). (n.d.). *Cysgeir*. <https://www.cysgliad.com/cy/cysgeir/>
- Carnegie Mellon Speech Group. (2014). *CMUdict (the Carnegie Mellon Pronouncing Dictionary)*. <https://github.com/cmuspinx/cmudict>
- Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (n.d.). *Cronfa Electroneg o Gymraeg (ceg): A 1 million word lexical database and frequency count for Welsh*. Retrieved 2001, from <http://corpws.cymru/ceg>
- Ethnologue : Languages of the world*. (2022). SIL International.
<https://www.ethnologue.com/>

- Halford, M. (n.d.). *Prince*. <https://github.com/MaxHalford/prince>
- Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E.-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., ... Scannell, K. (2020). *CorCenCC: Corpwys Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh*. <http://doi.org/10.17035/d.2020.0119878310>
- Mottl, D., & Henrique, J. (2018). *GetOldTweets3: GitHub repository*. Retrieved December 9, 2019, from <https://github.com/Mottl/GetOldTweets3>
- SurveyMonkey. (n.d.). <https://www.surveymonkey.co.uk/>

Language and reference

Reference works

- Convery, A., & Thorne, D. (Eds.). (1991). *Collins–Spurrell pocket Welsh dictionary* (rev. ed.). Harper–Collins.
- Duden online*. (n.d.). Retrieved January 12, 2019, from www.duden.de
- English short title catalogue*. (n.d.). Retrieved January 12, 2019, from estc.bl.uk
- Evans, H. M., & Thomas, W. O. (Eds.). (1953). *The new Welsh dictionary—y geiriadur newydd*. Christopher Davies.
- Evans, H. M., & Thomas, W. O. (Eds.). (1983). *Y Geiriadur Mawr* (11th ed.). Christopher Davies a Gwasg Gomer.
- Geiriadur Prifysgol Cymru ar lein*. (2024). Retrieved 2024, from geiriadur.ac.uk/gpc/gpc.html
- Griffiths, B., & Jones, D. G. (Eds.). (1995a). *Geiriadur yr Academi: The Welsh Academy English–Welsh dictionary*. Gwasg Prifysgol Cymru.
- Griffiths, B., & Jones, D. G. (Eds.). (1995b). *Geiriadur yr Academi: The Welsh Academy English–Welsh dictionary*. Gwasg Prifysgol Cymru.
- Gruffudd, H. (Ed.). (2007). *The Welsh learner's dictionary* (Pocket ed.). Y Lolfa.
- Johnston, D. (Ed.). (n.d.). *Dictionary of Welsh biography*. <https://biography.wales/>
- King, G. (Ed.). (2007). *Modern Welsh dictionary*. Oxford University Press.
- Lewis, D. G. (2015). *Geiriadur gwybod y geiriau Gomer*. Gwasg Gomer.
- Lewis, D. G., & Lewis, N. (Eds.). (2016). *Geiriadur Cymraeg Gomer*. Gwasg Gomer.
- Oxford English Dictionary online* (ed. 2) (2nd ed.). (n.d.). www.oed.com
- Oxford English Dictionary online* (ed. 3) (3rd ed.). (n.d.). www.oed.com

- Rees, E. (1987). *Libri Walliae* [2 vols.] National Library of Wales.
- Spurrell, W. (1903). *Geiriadur cynaniaethol Seisoneg a Chymraeg* (4th). W. Spurrell a'i Fab.
- Thomas, R. J., Bevan, G. A., & Donovan, P. J. (Eds.). (2002). *Geiriadur Prifysgol Cymru: A dictionary of the Welsh language*. Gwasg Prifysgol Cymru. (Original work published 1967)
- Williams, J. L. (Ed.). (1973). *Geiriadur termau—Dictionary of terms*. Gwasg Prifysgol Cymru.

Grammar and language-teaching materials

- Breese, N., & Clement, B. (2017). *Seren iaith! 2*. Atebol.
- King, G. (2003). *Modern Welsh: A comprehensive grammar* (2nd ed.). Routledge.
- Williams, C. (1999). *Cymraeg clir*. Gwasg Dwyfor.
- Williams, R. H. (2018). *Cymraeg yn y gweithle*. Gwasg Prifysgol Cymru.

Secondary sources

- Aaron, J., & Williams, C. (Eds.). (2005). *Postcolonial Wales*. University of Wales Press.
- Adamou, E. (2016). *A corpus-driven approach to language contact*. De Gruyter Mouton.
- Adaros, J. S., & Tironi, A. C. (2017). “What the f*ck was that?!” A preliminary study on the use of swear words by university English as a foreign language learners in Chile. *Lenguas Modernas*, (49), 9–25.
- Agarwal, P., Sharma, A., Grover, J., Sikka, M., Rudra, K., & Choudhury, M. (2017). I may talk in English but gaali toh Hindi mein hi denge: A study of English-Hindi code-switching and swearing pattern on social networks. *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, 554–557.
- Agha, A. (2005). Voice, footing, enregisterment. *Journal of linguistic anthropology*, 15, 38–59.
- Agha, A. (2009). What do bilinguals do? A commentary. In A. Reyes & A. Lo (Eds.), *Beyond Yellow English: Toward a linguistic anthropology of Asian Pacific America* (pp. 253–260). Oxford University Press.
- Agha, A. (2015). Enregisterment and communication in social history. In Frog & A. Agha (Eds.), *Registers of communication* (pp. 27–53). Finnish Literature Society.

- Ahmad, F., & Barner-Rasmussen, W. (2019). False foe? When and how code switching practices can support knowledge sharing in multinational corporations. *Journal of International Management*, 25.
- Aitchison, J., & Carter, H. (1994). *A geography of the Welsh language 1961–1991*. University of Wales Press.
- Akbari, F. (2020). *Iran's language planning confronting English abbreviations*. Springer.
- Alexander, D. M. (1929). Why not “U” for “You”? *American speech*, 5, 24–26.
- Alim, H. S., Rickford, J. R., & Ball, A. F. (Eds.). (2016). *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press.
- Amuzu, E. K. (2013). Producing composite codeswitching: The role of the modularity of language production. *International Journal of Bilingualism*, 18, 384–407.
- Amuzu, E. K. (2005). Revisiting the classic codeswitching–composite codeswitching distinction: A case study of nonverbal predication in Ewe–English codeswitching. *Australian Journal of Linguistics*, 25(1), 127–151.
- Amuzu, E. K. (2012). Socio-pragmatics of conversational codeswitching in Ghana. *Ghana Journal of Linguistics*, 1, 1–22.
- Andrews, T., & Prys, G. (2016). Terminology standardization in education and the construction of resources: The Welsh experience. *Education Sciences*, 6.
- Antia, B. E. (2000). *Terminology and language planning : An alternative framework of practice and discourse*. John Benjamins.
- ap Gareth, O. L. (2009). *Welshing on postcolonialism: Complicity and resistance in the construction of Welsh identities* [PhD Thesis]. University of Wales, Aberystwyth.
- ap Morgan, S. (1642). *The Welch-Man's Warning-Piece*. Robert Wood; Henry Marsh.
- Apel, K., Henbest, V. S., & Masterson, J. (2019). Orthographic knowledge: Clarifications, challenges, and future directions. *Reading and Writing*, 32, 873–889.
- Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, 16, 71–93.
- Assenza, E. (2016). The asymmetric multi-language model: A cognitive-pragmatic pattern to explain code-switching by unbalanced multilinguals. In K. Allan, A. Capone, & I. Kecskes (Eds.), *Pragmemes and theories of language use* (pp. 861–899). Springer.

- Auckle, T. (2017). Youth discourse in multilingual Mauritius: The pragmatic significance of swearing in multiple languages. *Stellenbosch Papers in Linguistics Plus*, 51, 61–86.
- Auer, P. (2022). ‘Translanguaging’ or ‘doing languages’? Multilingual practices and the notion of ‘codes’. In J. MacSwan (Ed.), *Multilingual perspectives on translanguaging* (pp. 126–153). Multilingual Matters.
- Auer, P. (2005). A postscript: Code-switching and social identity. *Journal of Pragmatics*, 37, 403–410.
- Awbery, G. M. (1996). *Tafodiaith yn y Gymraeg*. Uned Iaith Genedlaethol Cymru CBAC.
- Babino, A., & Stewart, M. A. (2020). *Radicalizing literacies and languaging: A framework toward dismantling the mono-mainstream assumption*. Palgrave Macmillan.
- Bache, C. (2010). Hjelmslev’s Glossematics: A source of inspiration to Systemic Functional Linguistics? *Journal of Pragmatics*, 42, 2562–2578.
- Badiola, L., Delgado, R., Sande, A., & Stefanich, S. (2018). Code-switching attitudes and their effects on acceptability judgment tasks. *Linguistic Approaches to Bilingualism*, 8, 5–24.
- Baker, C. (1995). Bilingual education in Wales. In O. García & C. Baker (Eds.), *Policy and practice in bilingual education: A reader*. Multilingual Matters.
- Baker, C. (2019). A Tribute to Ofelia García. *Journal of Multilingual Education Research*, 9, 175–182.
- Baker, S., Brown, B. J., & Williams, E. (2014). Illuminating a resilient rural culture in twentieth century Y Fro Gymraeg using Bourdieu’s *Bearn*. *Sociologia Ruralis*, 54, 40–56.
- Ball, M. J., & Williams, B. J. (2001). *Welsh phonetics*. Edwin Mellen Press.
- Ballinger, S., Lyster, R., Sterzuk, A., & Genesee, F. (2017). Context-appropriate crosslinguistic pedagogy: Considering the role of language status in immersion education. *Journal of Immersion and Content-Based Language Education*, 5, 30–57.
- Balsom, D. (1985). The three-Wales model. In J. Osmond (Ed.), *National question again: Welsh political identity in the 1980s* (pp. 1–17). Gwasg Gomer.
- Barakos, E. (2016). Language policy and governmentality in businesses in Wales: A continuum of empowerment and regulation. *Multilingua*, 35, 361–391.
- Barthes, R. (1965). *Éléments de sémiologie* (1964). In *Le degré zéro de l’écriture* (pp. 77–172). Editions Gonthier.

- Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford University Press.
- Baumann, J. F., Edwards, E. C., Boland, E. M., Olejnik, S., & Kame'enui, E. J. (2003). Vocabulary tricks: Effects of instruction in morphology and context on fifth-grade students' ability to derive and infer word meanings. *American Educational Research Journal*, 40, 447–494.
- Bautista, M. L. S. (2004). Tagalog-English code switching as a mode of discourse. *Asia Pacific Education Review*, 5, 226–233.
- Baxter, L. (2010). Lexical diffusion in the early stages of the merry-marry merger. *U. Penn Working Papers in Linguistics*, 16, 11–19.
- Beatty-Martínez, A. L., & Dussias, P. E. (2017). Bilingual experience shapes language processing: Evidence from codeswitching. *Journal of Memory and Language*, 95, 173–189.
- Bell, E., Archangeli, D. B., Anderson, S. J., Hammond, M., Webb-Davies, P., & Brooks, H. (2023). Northern Welsh. *Journal of the International Phonetic Association*, 53, 487–510.
- Benedict, B. M. (2004). Readers, writers, reviewers, and the professionalization of literature. In T. Keymer & J. Mee (Eds.), *The Cambridge companion to English literature 1740–1830* (pp. 3–23). Cambridge University Press.
- Ben-Rafael, M. (2008). French: Tradition versus innovation as reflected in English borrowings. In J. Rosenhouse & R. Kowner (Eds.), *Globally speaking: Motives for adopting English vocabulary in other languages* (pp. 44–67). Multilingual Matters.
- Bentahila, A., & Davies, E. E. (2001). Codeswitching: An unequal partnership? In R. Jacobson (Ed.), *Codeswitching worldwide* (pp. 25–50, Vol. 2). Mouton de Gruyter.
- Bhatt, R. M., & Bolonyai, A. (2019). Code-switching and translanguaging. In *Handbook of pragmatics* (pp. 61–78, Vol. 22). John Benjamins.
- Bhatt, R. M., & Bolonyai, A. (2022). Codeswitching and its terminological other-translanguaging. In J. MacSwan (Ed.), *Multilingual perspectives on translanguaging* (pp. 154–180). Multilingual Matters.
- Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 31–56). Oxford University Press.
- Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge University Press.

- Blackledge, A. (2002). 'What sort of people can look at a chicken and think dofednod?': Language, ideology and nationalism in public discourse. *Multilingua*, 197–226.
- Blackledge, A. (2006). The racialization of language in British political discourse. *Critical Discourse Studies*, 3, 61–79.
- Blackledge, A. (2009a). 'As a country we do expect': The further extension of language testing regimes in the United Kingdom. *Language Assessment Quarterly*, 6, 6–16.
- Blackledge, A. (2009b). Being English, speaking English. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration* (pp. 83–108). John Benjamins.
- Blackledge, A. (2009c). Inventing English as convenient fiction: Language testing regimes in the United Kingdom. In G. Extra, M. Spotti, & P. Van Avermaet (Eds.), *Language testing, migration and citizenship* (pp. 66–86). Continuum.
- Bodley, J. H. (2015). *Victims of progress* (5th ed.). Rowman; Littlefield.
- Boeschoten, H. (2001). Codeswitching, codemixing, and code alternation: What a difference. In R. Jacobson (Ed.), *Codeswitching worldwide* (pp. 15–24, Vol. 2). Mouton de Gruyter.
- Bohata, K. (2004). *Postcolonialism revisited*. University of Wales Press.
- Bokamba, E. G. (1988). Code-mixing, language variation, and linguistic theory: Evidence from Bantu languages. *Lingua*, 76, 21–62.
- Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. (2005). Onset entropy matters—letter-to-phoneme mappings in seven languages. *Reading and Writing*, 18, 211–229.
- Borsley, R. D., Tallerman, M., & Willis, D. (2007). *The syntax of Welsh*. Cambridge University Press.
- Bourhis, R. Y., Giles, H., & Tajfel, H. (1973). Language as a determinant of Welsh identity. *European Journal of Social Psychology*, 3, 447–460.
- Bowen, E. (1941). Ar ddifancoll. *Y Llenor*, 20, 189–196.
- Brake, P. (1998). *Cymraeg graenus*. Gwasg Gomer.
- Broersma, P., & Hamann, S. (2009). Loanword adaptation as first-language phonological perception. In A. Calabrese & W. L. Wetzels (Eds.), *Loan phonology* (pp. 11–58). John Benjamins.
- Brooks, S. (2021). *Hanes Cymry: Lleiafrifoedd ethnig a'r gwareiddiad Cymraeg*. Gwasg Prifysgol Cymru.

- Bugge, P. (2022). The history of the dichotomy of civic Western and ethnic Eastern nationalism. *National Identities*, 24, 505–522.
- Burchfield, R. W. (1973). The treatment of controversial vocabulary in the Oxford English dictionary. *Transactions of the Philological Society*, 72(1), 1–28.
- Burchfield, R. W. (1975). The art of the lexicographer. *Journal of the Royal Society of Arts*, 123, 349–361.
- Burton, S., & Soboleva, A. (2011). Interactive or reactive? Marketing with Twitter. *Journal of Consumer Marketing*, 28, 491–499.
- Camangian, P. (2008). Untempered tongues: Teaching performance poetry for social justice. *English Teaching: Practice and Critique*, 7, 35–55.
- Canagarajah, S. (2011). Codemeshing in academic writing: Identifying teachable strategies of translanguaging. *The modern language journal*, 95(3), 401–417.
- Carter, D., Davies, P., Deuchar, M., & Parafta Couto, M. d. C. (2011). A systematic comparison of factors affecting the choice of matrix language in three bilingual communities. *Journal of Language Contact*, 4(2), 153–183.
- Casaponsa, A., Carreiras, M., & Duñabeitia, J. A. (2015). How do bilinguals identify the language of the words they read? *Brain research*, 1624, 153–166.
- Casaponsa, A., & Duñabeitia, J. A. (2016). Lexical organization of language-ambiguous and language-specific words in bilinguals. *The Quarterly Journal of Experimental Psychology*, 69, 589–604.
- Choi, Y. H. (2015). Roles of L1 and L2 derivational morphological awareness in L2 reading through the mediation of L2 vocabulary knowledge. *The Journal of Asia TEFL*, 12, 81–114.
- Chomsky, N. (1993). A minimalist program for linguistic theory. *MIT Occasional Papers in Linguistics*, 1.
- Christofferson, K. O. (2014). Does child code-switching demonstrate communicative competence?: A comparison of simultaneous and sequential bilinguals. *Arizona Working Papers in SLA and Teaching*, 20–40.
- Cisło, A. (2012). Between tradition and innovation: The alternative Gaelic script. In Z. Wąsik, P. Czajka, & M. Szawerna (Eds.), *Alternate construals in language and linguistics* (pp. 131–142). Philological School of Higher Education in Wrocław.
- Clair, R. N. S., & Guadalupe, V. (1980). The sociology of code-switching. *Language Sciences*, 2(2), 205–221.
- Clapham, J. (2010). Dadansoddiad o ddefnydd athrawon dan hyfforddiant o gyfnwidiad cod mewn dosbarth uwchradd dwyieithog: Achos o gymru. *Gwerddon*.

- Clyne, M. (1987). Constraints on code switching: How universal are they? *Linguistics*, 25, 739–764.
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th). Routledge.
- Collins, B., & Mees, I. M. (1990). The phonetics of Cardiff English. In N. Coupland (Ed.), *English in Wales* (pp. 87–103). Multilingual Matters.
- Collins, W. M. (2003). Code-switching behavior as a strategy for Maya-Mam linguistic revitalization. *OSUWPL*, 57, 1–39.
- Connolly, J. H. (1981). On the segmental phonology of a South Welsh accent of English. *Journal of the International Phonetic Association*, 11, 51–61.
- Connolly, J. H. (1990). Port Talbot English. In N. Coupland (Ed.), *English in Wales* (pp. 121–129). Multilingual Matters.
- Constantine, M.-A. (2007). ‘Viewing most things thro’ false mediums’: Iolo Morganwg (1747–1826) and English perceptions of Wales. In C. Lamont & M. Rossington (Eds.), *Romanticism’s debatable lands* (pp. 27–38). Palgrave Macmillan.
- Constantine, M.-A. (2017). ‘British bards’: The concept of laboring class poetry in eighteenth-century Wales. In J. Goodridge & B. Keegan (Eds.), *A history of British working class literature* (pp. 101–115). Cambridge University Press.
- Costa, J. (2015). New speakers, new language: On being a legitimate speaker of a minority language in Provence. *International Journal of the Sociology of Language*, 231, 127–145.
- Coste, D., & Simon, D.-L. (2009). The plurilingual social actor. language, citizenship and education. *International Journal of Multilingualism*, 6, 168–185.
- Coupland, N. (1988). *Dialect in use*. University of Wales Press.
- Coupland, N. (1990). ‘Standard Welsh English’: A variable semiotic. In N. Coupland (Ed.), *English in Wales* (pp. 232–257). Multilingual Matters.
- Coupland, N., Bishop, H., & Garrett, P. (2003). Home truths: Globalisation and the iconising of Welsh in a Welsh-American newspaper. *Journal of Multilingual and Multicultural Development*, 24, 153–177.
- Couto, P., del Carmen, M., Deuchar, M., & Fusser, M. (2015). How do Welsh-English bilinguals deal with conflict? Adjective-noun order resolution. *Code-switching between structural and sociolinguistic perspectives*, 65–84.
- Crawford, R. (2001). *The modern poet: Poetry, academia, and knowledge since the 1750s*. Oxford University Press.

- Dafydd Jones, R. (2010). Islam and the rural landscape: Discourses of absence in west Wales. *Social & Cultural Geography*, 11(8), 751–768.
- Daland, R., Oh, M., & Kim, S. (2015). When in doubt, read the instructions: Orthographic effects in loanword adaptation. *Lingua*, 159, 70–92.
- Dampier, G. (2012). The (f)utility of additive bilingualism in South African education. *Education as Change*, 16, 69–81.
- Davies, P., & Deuchar, M. (2014). Auxiliary deletion in the informal speech of Welsh-English bilinguals: A change in progress. *Lingua*, 143, 224–241.
- Davies, P. G. C. (2010). *Identifying word-order convergence in the speech of Welsh-English bilinguals*. Bangor University.
- Davies, R. R. (1997). Presidential address: The peoples of Britain and Ireland, 1100–1400: IV language and historical mythology. *Transactions of the Royal Historical Society*, 7, 1–24.
- Day, G. (2002). *Making sense of Wales: A sociological perspective*. University of Wales Press.
- De Angelis, G. (2014). A multilingual approach to analysing standardized test results: Immigrant primary school children and the role of languages spoken in a bi-/multilingual community. *Intercultural Education*, 25, 14–28.
- De Groot, A. M. B., & Van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 9–29). Oxford University Press.
- de León, L. (2019). Playing at being bilingual: Bilingual performances, stance, and language scaling in Mayan Tzotzil siblings' play. *Journal of Pragmatics*, 144, 92–108.
- de Saussure, F. (1907). [1996] *premier cours de linguistique générale* (E. Komatsu, Ed.). Pergamon Press.
- de Saussure, F. (1909). *Deuxième cours de linguistique générale* (E. Komatsu, Ed.). Pergamon Press.
- de Saussure, F. (1911). [1993] *troisième cours de linguistique générale* (E. Komatsu, Ed.). Pergamon Press.
- De Vries Jr., H. J. (2008). Dutch: Is it threatened by English? In J. Rosenhouse & R. Kowner (Eds.), *Globally speaking: Motives for adopting English vocabulary in other languages* (pp. 68–81). Multilingual Matters.
- Declerck, M., & Grainger, J. (2017). Inducing asymmetrical switch costs in bilingual language comprehension by language practice. *Acta Psychologica*, 178, 100–106.

- DeDonno, M. A., Rivera-Torres, K., Monis, A., & Fagan, J. F. (2014). The influence of a time limit and bilingualism on Scholastic Assessment Test performance. *North American Journal of Psychology*, 16, 211–224.
- Detey, S., & Nespoulous, J.-L. (2008). Can orthography influence second language syllabic segmentation?: Japanese epenthetic vowels and French consonantal clusters. *Lingua*, 118, 66–81.
- Deuchar, M., Davies, P., Parafta Couto, M., & Carter, D. (2014). Building bilingual corpora. In E. M. Thomas & I. Mennen (Eds.), *Advances in the study of bilingualism* (pp. 93–111). Multilingual Matters.
- Deuchar, M. (2005a). Congruence and Welsh–English code-switching. *Bilingualism: Language and Cognition*, 8(3), 255–269.
- Deuchar, M. (2005b). Congruence and Welsh–English code-switching. *Bilingualism: Language and cognition*, 8, 255–269.
- Deuchar, M. (2005c). Minority language survival: Code-mixing in Welsh. *Proceedings of the 4th international symposium on bilingualism*, 608–620.
- Deuchar, M. (2006). Welsh–English code-switching and the Matrix Language Frame model. *Lingua*, 116(11), 1986–2011.
- Deuchar, M. (2020). Code-switching in linguistics: A position paper. *Languages*, 5(2), 22.
- Deuchar, M., & Davies, P. (2009). Code switching and the future of the welsh language.
- Deuchar, M., Donnelly, K., & Piercy, C. (2016). ‘Mae pobl monolingual yn minority’: Factors favouring the production of code switching by Welsh–English bilingual speakers. In M. Durham & J. Morris (Eds.), *Sociolinguistics in Wales* (pp. 209–239). Palgrave Macmillan.
- Deuchar, M., Muysken, P., & Wang, S.-L. (2007). Structured variation in codeswitching: Towards an empirically based typology of bilingual speech patterns. *International Journal of Bilingual Education and Bilingualism*, 10, 298–340.
- Deuchar, M., & Stammers, J. R. (2012). What is the “Nonce Borrowing Hypothesis” anyway? *Bilingualism: Language and Cognition*, 649–650.
- Deuchar, M., & Stammers, J. R. (2016). English-origin verbs in Welsh: Adjudicating between two theoretical approaches. *Languages*, 1.
- Deuchar, M., Webb-Davies, P., & Donnelly, K. (2018). *Building and using the Siarad corpus: Bilingual conversations in Welsh and English*. John Benjamins.

- Dewaele, J.-M. (2004a). Blistering barnacles! What language do multilinguals swear in?! *Estudios de Sociolingüística*, 5, 83–105.
- Dewaele, J.-M. (2004b). The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of multilingual and multicultural development*, 25, 204–222.
- Dewaele, J.-M. (2010). Christ fucking shit merde! Language preferences for swearing among maximally proficient multilinguals. *Sociolinguistic Studies*, 4, 595–614.
- Dewaele, J.-M. (2016a). Thirty shades of offensiveness: L1 and LX English users' understanding, perception and self-reported use of negative emotion-laden words. *Journal of Pragmatics*, 94, 112–127.
- Dewaele, J.-M. (2016b). Why do so many bi-and multilinguals feel different when switching languages? *International Journal of multilingualism*, 13, 92–105.
- Dewaele, J.-M. (2018). “Cunt”: On the perception and handling of verbal dynamite by L1 and LX users of English. *Multilingua*, 37, 53–81.
- Di Sciullo, A.-M., Muysken, P., & Singh, R. (1986). Government and code-mixing. *Journal of linguistics*, 22, 1–24.
- Dinkin, A. J. (2005). Mary, darling, make me merry; say you'll marry me: Tense-lax neutralization in the Linguistic Atlas of New England. *University of Pennsylvania Working Papers in Linguistics*, 11(2), 5.
- Dissanayake, W. (2008). Language in cinema. In B. B. Kachru, K. Yamuna, & S. N. Sridhar (Eds.), *Language in South Asia* (pp. 395–406). Cambridge University Press.
- Dobrin, L. M., & Good, J. (2009). Practical language development: Whose mission? *Language*, 85(3), 619–629.
- Dodd, W. S. (1998). Twentieth-century Welsh lexicography as a context for the *Geiriadur yr Academi*. *International Journal of Lexicography*, 11, 55–64.
- Drinkwater, S. J., & O'Leary, N. C. (1997). Unemployment in Wales: Does language matter? *Regional Studies*, 31, 583–591.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19, 61–74.
- Durham, M. (2022). How many <a>s in Wales? performing a Welsh accent on Twitter. In C. Cutler, M. Ahmar, & S. Bahri (Eds.), *Digital orality: Vernacular writing in online spaces* (pp. 189–217). Springer.
- Durham, M., & Morris, J. (2016). An overview of sociolinguistics in Wales. In M. Durham & J. Morris (Eds.), *Sociolinguistics in Wales* (pp. 3–28). Palgrave Macmillan.

- Durkin, P. (2014). *Borrowed words: A history of loanwords in English*. Oxford University Press.
- Durvasula, K., Huang, H.-H., Uehara, S., Luo, Q., & Lin, Y.-H. (2018). Phonology modulates the illusory vowels in perceptual illusions: Evidence from Mandarin and English. *Laboratory Phonology*, 9, 1–27.
- Edwards, E. (2019). Romantic Wales and the eisteddfod. In G. Evans & H. Fulton (Eds.), *The Cambridge history of Welsh literature* (pp. 285–305). Cambridge University Press.
- Edwards, J. (2009). *Language and identity: An introduction*. Cambridge University Press.
- Edwards, J. (2010). *Language diversity in the classroom*. Multilingual Matters.
- Edwards, T. (1847). *A brief analysis of Welsh orthography*. Thomas Gee.
- Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19, 161–188.
- Ellis, C., Kuipers, J. R., Thierry, G., Lovett, V., Turnbull, O., & Jones, M. W. (2015). Language and culture modulate online semantic processing. *SCAN*, 10, 1392–1396.
- Ellis, N. C., & Hooper, A. M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics*, 22, 571–599.
- Engesser, S., & Humprecht, E. (2015). Frequency or skillfulness: How professional news media use Twitter in five Western countries. *Journalism Studies*, 16, 513–529.
- Enli, G., & Simonsen, C.-A. (2018). ‘Social media logic’ meets professional norms: Twitter hashtags usage by journalists and politicians. *Information, Communication & Society*, 21, 1081–1096.
- Evans, B. G. (1889). Cymro, Cymru, a Chymreig yn eu cysylltiad ag addysg. *Transactions of the Liverpool Welsh National Society, Fourth Session*.
- Evans, D. J. (1889). Foreign elements in the Welsh language. *Cymru Fydd*, 2(3), 140–144.
- Evans, L., Bolton, N., Jones, C., & Iorwerth, H. (n.d.). ‘Defnyddiwch y Gymraeg’: Community sport as a vehicle for encouraging the use of Welsh. *Sport in Society*.
- Evans, L., Bolton, N., Jones, C., & Iorwerth, H. (2019). ‘Defnyddiwch y Gymraeg’: Community sport as a vehicle for encouraging the use of the Welsh language. *Sport in Society*, 22, 1115–1129.

- Evans, S. W. (2006). Differential performance of items in mathematics assessment materials for 7-year-old pupils in English-medium and Welsh-medium versions. *Educational Studies in Mathematics*, 64, 145–168.
- Fabricius, A. H. (2005). Mobility, contact and an accent norm. In B. Preisler, A. Fabricius, H. Haberland, S. Kjærbeck, & K. Risager (Eds.), *The consequences of mobility* (pp. 120–134). Roskilde Universitet.
- Fägersten, K. B. (2017). English-language swearing as humor in Swedish comic strips. *Journal of Pragmatics*, 121, 175–187.
- Farrugia, M. T. (2003). The use of english as a medium of instruction in maltese mathematics classrooms: Continuing the debate. *Journal of Maltese education research*, 1, 1–14.
- Field, F. W. (2002). *Linguistic borrowing in bilingual contexts*. John Benjamins.
- Field, M. C. (2009). Changing Navajo language ideologies and changing language use. In P. V. Kroskrity & M. C. Field (Eds.), *Native American language ideologies: Beliefs, practices, and struggles in Indian country* (pp. 31–47). University of Arizona Press.
- Fiesler, C., & Proferes, N. (2018). ‘Participant’ perceptions of Twitter research ethics. *Social Media+ Society*, 4(1).
- Fillmore, C. J. (1988). The mechanisms of ‘Construction Grammar’. *Proceedings of the fourteenth annual meeting of the Berkeley Linguistics Society*, 35–55.
- Fillmore, C. J., Kay, P., & O’Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, 64, 501–538.
- Finn, E. (2017). *What algorithms want: Imagination in the age of computing*. MIT Press.
- Fiske, S. T. (2022). Twitter manipulates your feed: Ethical considerations. *Proceedings of the National Academy of Sciences*, 119.
- Forte, A., Humphreys, M., & Park, T. (2012). Grassroots professional development: How teachers use Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 6, 106–113.
- Fowkes, R. A. (1959). Phonological aspects of Welsh-English bilingualism. *General Linguistics*, 4(1), 23.
- Fowkes, R. A. (1981). Welsh naming practices, with a comparative look at Cornish. *Names*, 29, 265–272.
- Fozdar, F., & Low, M. (2015). ‘They have to abide by our laws...and stuff’: Ethnonationalism masquerading as civic nationalism. *Nations and Nationalism*, 21, 524–543.

- Freeman, M. (2007). Lady Llanover and the Welsh costume prints. *The National Library of Wales Journal*, 34, 235–252.
- Frekko, S. E. (2009). Social class, linguistic normativity and the authority of the “native Catalan speaker” in Barcelona. In N. M. Doerr (Ed.), *The native speaker concept: Ethnographic investigations of native speaker effects* (pp. 161–184). Walter de Gruyter.
- Fricke, M., & Kootstra, G. J. (2016). Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91, 181–201.
- Friesen, D. C., Chung-Fat-Yim, A., & Bialystok, E. (2016). Lexical selection differences between monolingual and bilingual listeners. *Brain and language*, 152, 1–13.
- Frost, R., & Kampf, M. (1992). Phonetic recoding of phonologically ambiguous printed words. In C. A. Fowler (Ed.), *Status report on speech research* (pp. 121–134). Haskins Laboratories.
- Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of experimental psychology: human perception and performance*, 13(1), 104.
- Fulford, T. (2024). Farmers’ boys and doomed youths: Producing the poet in the print culture of the Romantic era. *European Romantic Review*, 35, 463–477.
- Fulton, H. (2011). Negotiating Welshness: Multilingualism in Wales before and after 1066. In E. M. Tyler (Ed.), *Conceptualizing multilingualism in medieval England, c.800–c.1250* (pp. 145–70). Brepols.
- Gabrys-Barker, D. (2020). The practice of classroom code-switching, translanguaging or simply alternating languages in FL teacher talk. *Anglica Wratislaviensia*, 58, 147–164.
- García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Wiley/Blackwell.
- García, O., & Li, W. (2014). *Language, bilingualism and education*. Palgrave Macmillan.
- García, O., & Lin, A. M. Y. (2017). Translanguaging in bilingual education. In O. García, A. M. Y. Lin, & S. May (Eds.), *Bilingual and multilingual education* (3rd ed., pp. 117–130). Springer.
- García, O., & Wei, L. (2014). Translanguaging: Language, bilingualism and education (A. I. Young, Ed.). *Heritage Language Journal*, 11(3), 243–248.
- Gardner, M. (1972). *Codes, ciphers and secret writing*. Dover Publications.

- Gardner-Chloros, P., McEntee-Atalianis, L., & Paraskeva, M. (2013). Code-switching and pausing: An interdisciplinary study. *International Journal of Multilingualism*, 10, 1–26.
- Gawne, L., & McCulloch, G. (2019). Emoji as digital gestures. *Language@Internet*, 17.
- Gawne, P. (2002). Securing the future of Manx Gaelic. In P. Davey & D. Finlayson (Eds.), *Manmin revisited: Twelve essays on Manx culture and environment* (pp. 173–183). Scottish Society for Northern Studies.
- Giancarlo, M. (2001). The rise and fall of the Great Vowel Shift? The changing ideological intersections of philology, historical linguistics, and literary history. *Representations*, 76(1), 27–60.
- Goddard, C. (2011). The lexical semantics of *language* (with special reference to *words*). *Language sciences*, 33, 40–57.
- Gökçe, S., Berberoğlu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic distance and translation differential item functioning on Trends in International Mathematics and Science Study mathematics assessment items. *Journal of Psychoeducational Assessment*, 39, 728–745.
- Gómez-Jiménez, E. M. (2015). ‘Oride lesgo eckshun’: Spelling foregrounding in the experimental poetry of E. E. Cummings. *Language and Literature*, 24, 307–321.
- Goodridge, J., & Keegan, B. (2004). John Clare and the traditions of labouring-class verse. In T. Keymer & J. Mee (Eds.), *The Cambridge companion to English literature 1740–1830* (pp. 280–295). Cambridge University Press.
- Görlach, M. (1997). *The linguistic history of English: An introduction*. Macmillan.
- Görlach, M. (Ed.). (2001). *A dictionary of European Anglicisms: A usage dictionary of Anglicisms in sixteen European languages*. Oxford University Press.
- Görlach, M. (2003). *English words abroad*. John Benjamins.
- Gravell, E. (2005). Battling bards, choosing chocolates. *British Journal of General Practice*, 55, 816.
- Gries, S. T. (2005). Null-hypothesis significance testing of word frequencies: A follow-up on Kilgariff. *Corpus Linguistics and Linguistic Theory*, 1, 277–294.
- Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, 33, 69–86.
- Grosjean, F. (2008). *Studying bilinguals*. Oxford University Press.
- Gruffudd, P. (1995). Remaking Wales: Nation-building and the geographical imagination, 1925–50. *Political geography*, 14, 219–239.

- Gumperz, J. J. (1977). The sociolinguistic significance of conversational code-switching. *RELC Journal*, 8, 1–34.
- Halliday, M. A. K. (1964). Syntax and the consumer (1964). In J. Webster (Ed.), [2003] *on language and linguistics* (pp. 36–49, Vol. 3). Continuum.
- Halliday, M. A. K. (1965). Speech and situation. *English in education*, 2, 14–17.
- Halliday, M. A. K. (1978). *Language as social semiotic*. University Park Press.
- Halliday, M. A. K. (1979). Modes of meaning and modes of expression: Types of grammatical structure and their determination by different semantic functions. In J. Webster (Ed.), *On grammar* (pp. 196–218, Vol. 1). Continuum.
- Halliday, M. A. K. (1990). New ways of meaning: The challenge to applied linguistics (1990). In J. Webster (Ed.), [2003] *on language and linguistics* (pp. 139–176, Vol. 3). Continuum.
- Halliday, M. A. K. (1992). The history of a sentence (1992). In J. Webster (Ed.), [2003] *on language and linguistics* (pp. 355–374, Vol. 3). Continuum.
- Halliday, M. A. K. (1994). Appendix: Systemic theory (1994). In J. Webster (Ed.), [2003] *on language and linguistics* (pp. 433–442, Vol. 3). Continuum.
- Halliday, M. A. K. (2001). Is the grammar neutral? Is the grammarian neutral? (2001). In J. Webster (Ed.), [2003] *on language and linguistics* (pp. 271–294, Vol. 3). Continuum.
- Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th). Routledge.
- Halliday, M. A. K., & Martin, J. R. (2003). *Writing science: Literacy and discursive power*. Taylor & Francis.
- Hamann, S., & Colombo, I. E. (2017). A formal account of the interaction of orthography and perception. *Natural Language & Linguistic Theory*, 35, 683–714.
- Hamdi, S. (2017). Lexical borrowing in Arabic and the role of orthography. *International Journal of Language and Linguistics*, 4, 17–28.
- Handbook of the International Phonetic Association*. (1999). Cambridge University Press.
- Hannahs, S. J. (2013). *The phonology of Welsh*. Oxford University Press.
- Hannisdal, B. R. (2005). From public school accent to BBC English: Defining Received Pronunciation. *Målbryting*, 7, 191–210.
- Harding, J. (2008). Modernist poetry and the canon. In A. Davis & L. Jenkins (Eds.), *The Cambridge companion to modernist poetry*. (pp. 225–243). Cambridge University Press.

- Harrington, J., Palethorpe, S., & Watson, C. (2000). Monophthongal vowel changes in Received Pronunciation: An acoustic analysis of the Queen's Christmas broadcasts. *Journal of the International Phonetic Association*, 30, 63–78.
- Hasan, R. (2013). Choice, system, realisation: Describing language as meaning potential. In L. Fontaine, T. Bartlett, & G. O'Grady (Eds.), *Systemic functional linguistics: Exploring choice* (pp. 269–299). Cambridge University Press.
- Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the world's languages: A comparative handbook* (pp. 35–54). Walter de Gruyter.
- Haspelmath, M. (2023). Inflection and derivation as traditional comparative concepts. *Linguistics*. <https://doi.org/10.1515/ling-2022-0086>
- Hawke, A. (2018). Coping with an expanding vocabulary: The lexicographical contribution to Welsh. *International Journal of Lexicography*, 31, 229–248.
- Hayes, B. (2009). *Introductory phonology*. Wiley/Blackwell.
- Hedd, C. (2023). *God Sêf ddy Chwyn* [personal correspondence]. <https://www.carwpiws.cymru/collections/posteri/products/poster-god-sef-ddy-chwyn>
- Heinz, S. (2003). *Welsh dictionaries in the twentieth century: A critical analysis*. Lincom Europa.
- Henken, E. R. (1996). *National redeemer: Owain Glyndŵr in Welsh tradition*. University of Wales Press.
- Hewitt, S. (2017). Neo-speakers of endangered languages: Theorizing failure to learn the language properly as creative post-vernacularity. *Journal of Celtic linguistics*, 18, 127–154.
- Hickey, T. (2009). Code-switching and borrowing in Irish. *Journal of Sociolinguistics*, 13, 670–688.
- Hill, M. L., Pérez, B., & Irby, D. J. (2008). Street fiction: What is it and what does it mean for English teachers? *English Journal*, 97, 76–81.
- Hinrichs, L. (2004). Emerging orthographic conventions in written Creole: Computer-mediated communication in Jamaica. *AAA: Arbeiten aus Anglistik und Amerikanistik*, 29, 81–109.
- Hinton, M. (2015). Changes in Received Pronunciation: Diachronic case studies. *Research in Language*, 13, 21–37.
- Hjelmslev, L. (1942). Langue et parole. *Cahiers Ferdinand de Saussure*, 2, 29–44.
- Hjelmslev, L. (1947). Structural analysis of language. *Studia linguistica*, 1(1-3), 69–78.
- Hjelmslev, L. (1954). La stratification du langage. *Word*, 10, 163–188.

- Hjelmslev, L. (1961). *Prolegomena to a theory of language* (F. J. Whitfield, Trans.). University of Wisconsin Press.
- Hodgkin, A. (2017). *Following Searle on Twitter*. The University of Chicago Press.
- Honeybone, P., & Watson, K. (2013). Salience and the sociolinguistics of Scouse spelling: Exploring the phonology of the contemporary humorous localised dialect literature of Liverpool. *English World-Wide*, 34, 305–340.
- Horan, G. (2013). ‘You taught me language; and my profit on’t/Is, I know how to curse’: Cursing and swearing in foreign language learning. *Language and Intercultural Communication*, 13(3), 283–297.
- Hornsby, M. (2014). Colli iaith neu newid iaith? A critical sociolinguistic look at recent changes in the Welsh language. *Res Celticae*, 1, 49–62.
- Hornsby, M. (2015). *Revitalizing minority languages: New speakers of Breton, Yiddish and Lemko*. Palgrave Macmillan.
- Huang, G., & Milroy, L. (1995). Language preference and structures of code-switching. In D. Graddol & S. Thomas (Eds.), *Language in a changing Europe* (pp. 35–46). Multilingual Matters.
- Hughes, J. E. (2011). *Canllawiau iaith a chymorth sillafu* (Argraffiad Newydd). Gwasg Gomer.
- Huszára, F., Ktena, S. I., O’Brien, C., Belli, L., Schlaikjera, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, 119.
- Huws, C. F. (2011). Yr iaith Gymraeg fel model ar gyfer torri’r cylch diffyg defnydd mewn cyd-destun ieithoedd lleiafrifol. *Gwerddon*, 8, 7–27.
- Iannacci, L. (2008). Beyond the pragmatic and the liminal: Culturally and linguistically diverse students code-switching in early-years classrooms. *TESL Canada Journal*, 26, 103–123.
- Imamura, K. (2017). The pursuit of insular authenticity: The spelling reform of loanwords in Palauan. *Shima*, 12, 118–127.
- Iosad, P. (2017). The phonologisation of redundancy: Length and quality in Welsh vowels. *Phonology*, 34, 121–162.
- Irwin, M. (2011). Japanese loanword orthography from 1955. *山形大学人文学部研究年報*, 8, 39–57.
- Itô, J., & Mester, A. (1995). Japanese phonology: Constraint domains and structure preservation. In J. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 817–838). Blackwell.

- Itô, J., & Mester, A. (1999). The structure of the phonological lexicon. In N. Tsujimura (Ed.), *The handbook of Japanese linguistics* (pp. 62–100). Blackwell.
- Itô, J., & Mester, A. (2001). Covert generalizations in Optimality Theory: The role of stratal faithfulness constraints. *Studies in Phonetics, Phonology, and Morphology*, 7(2), 273–299.
- Iyengar, M. M. (2015). Not mere abstractions: Language policies and language ideologies in U.S. settler colonialism. *Decolonization: Indigeneity, Education & Society*, 3, 33–59.
- Jackson, L., & Jones, R. D. (2014). ‘We’ll keep a welcome in the hillsides’? Proximity, distance and hospitality towards migrants in Wales. *Contemporary Wales*, 27, 82–104.
- Jaffe, A. (2012). Transcription in practice: Nonstandard orthography. *Journal of Applied Linguistics*, 3, 203.
- Jaffe, A., & Walton, S. (2000). The voices people read: Orthography and the representation of non-standard speech. *Journal of Sociolinguistics*, 4, 561–587.
- Jake, J. L. (1998). Constructing interlanguage: Building a composite matrix language. *Linguistics*, 36(2), 333–382.
- James, A. (2011). English(es) in post-devolution Wales: The sociolinguistic reconstruction of late modern Valleys Voice. *AAA–Arbeiten aus Anglistik und Amerikanistik*, 36, 47–63.
- James, A. (2018). Transgressive *Cymreictod* as scripted bilingualism: Challenging Welsh language identities via selective syncretism with English. *AAA–Arbeiten aus Anglistik und Amerikanistik*, 43, 27–44.
- James, C., Scholfield, P., Garrett, P., & Griffiths, Y. (1993). Welsh bilinguals’ English spelling: An error analysis. *Journal of Multilingual & Multicultural Development*, 14(4), 287–306.
- James, E. W. (2019). Popular poetry, Methodism, and the ascendancy of the hymn. In G. Evans & H. Fulton (Eds.), *The Cambridge history of Welsh literature*. Cambridge University Press.
- Jiang, Y.-L. B., García, G. E., & Willis, A. I. (2014). Code-mixing as a bilingual instructional strategy. *Bilingual Research Journal*, 37, 311–326.
- Job, R., Peressotti, F., & Cusinato, A. (1998). Lexical effects in naming pseudowords in shallow orthographies: Further empirical data. *Journal of Experimental Psychology*, 24, 622–630.

- Johnes, G. (2020). Medium efficiency: Comparing inputs and outputs by language of instruction in secondary schools in Wales. *Cylchgrawn Addysg Cymru*, 22, 52–66.
- Johnes, M. (2012). *Wales since 1939*. Manchester University Press.
- Johnes, M. (2015). History and the making and remaking of Wales. *History*, 100, 667–684.
- Johnes, M. (2019). *Wales: England's colony?* Parthian.
- Johnson, P., & Giles, H. (1982). Values, language and inter-cultural differentiation: The Welsh-English context. *Journal of Multilingual & Multicultural Development*, 3, 103–116.
- Johnston, D. (2000). Ceidwaid yr hen iaith? Beirdd yr Uchelwyr a'r iaith Saesneg. *Y Traethodydd*, 16–24.
- Jones, B. P. (1988). Official Welsh. In M. J. Ball (Ed.), *The use of Welsh: A contribution to sociolinguistics* (pp. 172–181). Multilingual Matters.
- Jones, D. V. (1998). National Curriculum tests for mathematics in English and Welsh: Creating matched assessments. *Assessment in Education*, 5, 193–211.
- Jones, F. P. (1965). Gwerin cymru. *Y Traethodydd*, 33 (3rd series), 12–22.
- Jones, G. E. (1984). The distinctive vowels and consonants of Welsh. In M. J. Ball & G. E. Jones (Eds.), *Welsh phonology* (pp. 40–64). University of Wales Press.
- Jones, M. C. (2005). Some structural and social correlates of single word intrasentential code-switching in Jersey Norman French. *Journal of French Language Studies*, 15, 1–23.
- Jones, R. M. (1992). Beyond identity? The reconstruction of the Welsh. *Journal of British studies*, 31, 330–357.
- Jones, R., & Desforges, L. (2003). Localities and the reproduction of Welsh nationalism. *Political geography*, 22(3), 271–292.
- Jones, R., & Fowler, C. (2007). Where is Wales? narrating the territories and borders of the Welsh linguistic nation. *Regional Studies*, 41, 89–101.
- Jones, R. J., Cunliffe, D., & Honeycutt, Z. R. (2013). Twitter and the Welsh language. *Journal of Multilingual and Multicultural Development*, 34, 653–671.
- Joshi, A. (1985). Processing of sentences with intrasentential code-switching. natural language parsing ed. by david dowty; lauri kartunnen; and arnold zwicky, 190–205. In D. Dowty, L. Kartunnen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 190–205). Cambridge: Cambridge University Press.

- Juan-Garau, M., & Lyster, R. (2019). Becoming bilingual through additive immersive programs. In A. De Houwer & L. Ortega (Eds.), *The cambridge handbook of bilingualism* (pp. 213–32). Cambridge University Press.
- Kachru, B. B. (1978). Toward structuring code-mixing: An Indian perspective. *International journal of the sociology of language*, 16, 27–46.
- Kamwangamalu, N. M., & Lee, C.-L. (1991). Chinese-English code-mixing: A case of matrix language assignment. *World Englishes*, 10, 247–261.
- Karrebæk, M. S. (2003). Iconicity and structure in codeswitching. *International Journal of Bilingualism*, 7, 407–441.
- Kauffman, R. H. (2018). If it looks like a duck: Emojis, emoticons and ambiguity. *Commentator*, Spring, 26–29.
- Kaufmann, E. (2019). Ethno-traditional nationalism and the challenge of immigration. *Nations and Nationalism*, 25, 435–448.
- Kelly, L. L. (2013). Hip-hop literature: The politics, poetics, and power of hip-hop in the English classroom. *English Journal*, 102, 51–56.
- Khan, I. U., Ahmed, A., & Gul, S. (2023). Students' perception of code-switching as pedagogical tool in multilingual EFL classrooms: A social justice perspective. *Kashmir journal of language research*, 26, 187–206.
- Khan, S. S. (2019). Standardized entrance/apptitude testing in the era of Foucault's Panopticon: The bilingual student and the minority student. *Education and Urban Society*, 52, 847–871.
- King'ei, K. (1999). Swahili technical terminology: Problems of development and usage in Kenya. *AAP*, 60, 147–160.
- Kiparsky, P. (1982). *Explanation in phonology*. Foris.
- Klein, B., & Wardle, C. (2008). "These two are speaking Welsh on Channel 4!" Welsh representations and cultural tensions on Big Brother 7. *Television & New Media*, 9, 514–530.
- Knight, D., Tovey-Walsh, B., Morris, S., Davies, E., & Prosser, H. (2022). *The Geirfan wordlist: A vocabulary list for adult learners of Welsh (version 2)*. Cardiff University.
- Kohn, H. (1944). *The idea of nationalism: A study in its origins and background*. Macmillan.
- Kolers, P. A. (1966). Reading and talking bilingually. *The American Journal of Psychology*, 79, 357–376.

- Kootstra, G. J., Van Hell, J. G., & Dijkstra, T. (2012). Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition*, 15(4), 797–819.
- Koplenig, A. (2019). Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15, 321–346.
- KordiĆ, S. (2007). La langue croata serbe, bosniaque et monténégain. *Au sud de l'Est*, 3, 74–78.
- Krägeloh, C., & Neha, T. N. (2014). Lexical expansion and terminological planning in indigenous and planned languages: Comparisons between Te Reo Māori and Esperanto. *Language Problems and Language Planning*, 38, 59–86.
- Krapp, G. P. (1925). *The English language in America* (Vol. 2). The Century Co.
- Kroskrity, P. V. (2014). Borders traversed, boundaries erected: Creating discursive identities and language communities in the village of Tewa. *Language & Communication*, 38, 8–17.
- Kuzio, T. (2002). The myth of the civic state: A critical survey of Hans Kohn's framework for understanding nationalism. *Ethnic and racial studies*, (25), 20–39.
- Lagopoulos, A. P., & Boklund-Lagopoulou, K. (2021). *Theory and methodology of semiotics: The tradition of Ferdinand de Saussure*. De Gruyter Mouton.
- Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, 60, 44–51.
- Landau, J. (2010). Alphabet reform in the six independent ex-Soviet Muslim republics. *Journal of the Royal Asiatic Society*, 20, 25–32.
- Lantto, H. (2014). Code-switching, swearing and slang: The colloquial register of Basque in Greater Bilbao. *International Journal of Bilingualism*, 18(6), 633–648.
- Larsen, S. E. (1993). Patriarchal hierarchies: Notes on Hjelmslev.
- Lauro, J., & Schwartz, A. I. (2017). Bilingual non-selective lexical access in sentence contexts: A meta-analytic review. *Journal of Memory and Language*, 92, 217–233.
- Lauttamus, T. (1999). Fuzzy switch and loan types in the languages of Finnish Americans. *SKY journal of linguistics*, 12, 87–109.
- Lee, R. (2015). Implicit associations with Welsh in two educational contexts. *York Papers in Linguistics Series 2*, 14, 81–105.
- Lee, R. (2016, September). *Language ideologies in the secondary school: Attitude and identity in bilingual Wales* [PhD Thesis]. University of York.

- Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, 32, 533–550.
- Lewin, C. (2020). ‘An English monstrosity’? Evolution and reception of Manx orthography. *Studia Celtica Posnaniensia*, 5, 35–59.
- Lewis, C. (1995). Welsh soap: “Pobol y Cwm” and Welsh national identity. *Critical Survey*, 7, 152–157.
- Lewis, D. M. (2011). A discourse-constructional approach to the emergence of discourse markers in English. *Linguistics*, 49, 415–443.
- Lewis, H. (1931). *Datblygiad yr iaith Gymraeg*. Gwasg Prifysgol Cymru.
- Lewis, L. G. (2019). ‘Amlhau lleisiau’n llên’: Birth and rebirth in Welsh language literature, 1990–2014. In G. Evans & H. Fulton (Eds.), *The Cambridge history of Welsh literature* (pp. 669–695). Cambridge University Press.
- Lewis, R. E. (2008). *Wenglish: The dialect of the South Wales Valleys*. Y Lolfa.
- Li, D. C. S. (2000). Cantonese-English code-switching research in Hong Kong: A Y2K review. *World Englishes*, 19, 305–322.
- Li, L., & Gong, Q. (2018). Lettered words in Chinese dictionaries: A challenge for non-Romanised language. *Lexicography*, 5, 85–96.
- Lieber, R. (2004). *Morphology and lexical semantics*. Cambridge University Press.
- Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31, 374–397.
- Lindsey, G. (2019). *English after RP*. Palgrave Macmillan.
- Luckett, K. (1993). ‘National Additive Bilingualism’: Towards the formulation of a language plan for South African schools. *Southern African Linguistics and Applied Language Studies*, 2, 38–60.
- Lukin, A., Moore, A. R., Herke, M., Wegener, R., & Wu, C. (2011). Halliday’s model of register revisited and explored. *Linguistics and the Human Sciences*, 4, 187–213.
- Lutzky, U. (2024). “Doesn’t really answer my question...”: Exploring customer service interactions on Twitter. *International Journal of Business Communication*, 61, 92–114.
- Lynch, P. (2007). Twentieth-century Welsh literature. In S. Mac Mathúna & A. Ó Corráin (Eds.), *Celtic literatures in the twentieth century* (pp. 97–128). Languages of Slavonic Culture.

- Lyon, J. (1991). Patterns of parental language use in Wales. *Journal of Multilingual and Multicultural Development*, 12, 165–181.
- Mac Giolla Chríost, D. (2005). Prestige planning and the Welsh language: Marketing, the consumer-citizen and language behaviour. *Current Issues in Language Planning*, 6, 64–72.
- Macalister, J. (2006). The Maori presence in the New Zealand English lexicon, 1850–2000: Evidence from a corpus-based study. *English World-Wide*, 27, 1–24.
- MacAulay, D. (1992). In D. MacAulay (Ed.), *The Celtic languages* (pp. 1–8). Cambridge University Press.
- MacSwan, J. (1999). *A minimalist approach to inter-sentential code switching*. Garland.
- Madoc-Jones, I., & Parry, O. (2012). Damned if you do and damned if you don't: Language choice in the research interview. *International Journal of the Sociology of Language*, 215, 165–182.
- Madrigal, E., Jiang, X., & Roy-Chowdhuri, S. (2017). The professional Twitter account: Creation, proper maintenance, and continuous successful operation. *Diagnostic Cytopathology*, 45, 621–628.
- Makihara, M. (2007). Linguistic purism in Rapa Nui political discourse. In M. Makihara & B. B. Schieffelin (Eds.), *Consequences of contact: Language ideologies and sociocultural transformations in Pacific societies* (pp. 49–69). Oxford University Press.
- Malin, S. D. (1965). Eye dialect in “Li'l Abner”. *American Speech*, 40, 229–232.
- Mancilla-Martinez, J., & Kieffer, M. J. (2010). Language minority learners' home language use is dynamic. *Educational Researcher*, 39(7), 545–546.
- Marcolini, S., Burani, C., & Colombo, L. (2009). Lexical effects on children's pseudoword reading in a transparent orthography. *Reading and Writing*, 22, 531–544.
- Martinet, A. (1980). *Éléments de linguistique générale* (new rev. and updated ed.). Armand Colin.
- Martínez, R. A., & Morales, P. Z. (2014). ¿Puras groserías?: Rethinking the role of profanity and graphic humor in Latin@ students' bilingual wordplay. *Anthropology & Education Quarterly*, 45(4), 337–354.
- Mathieu, L. (2012). Orthographic traces in Romanian and Japanese loanwords: Enriching phonological representations. *Journal of Language Contact*, 5, 144–181.

- Matthiessen, C. M. I. M., Teruya, K., & Lam, M. (2010). *Key terms in systemic functional linguistics*. Continuum.
- May, S. (2000). Extending ethnolinguistic democracy in Europe: The case of Wales. *The Sociological Review*, 48, 147–172.
- May, S. (2012). *Language and minority rights* (2nd ed.). Routledge.
- Mayer, P., Crowley, K., & Kaminska, Z. (2007). Reading and spelling processes in Welsh–English bilinguals: Differential effects of concurrent vocalisation tasks. *Reading and Writing*, 20, 671–690.
- Mayr, R. (2010). What exactly is a front rounded vowel? An acoustic and articulatory investigation of the NURSE vowel in South Wales English. *Journal of the International Phonetic Association*, 40, 93–112.
- Mayr, R., & Davies, H. (2011). A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh. *Journal of the International Phonetic Association*, 41, 1–25.
- Mayr, R., Howells, G., & Lewis, R. (2015). Asymmetries in phonological development: The case of word-final cluster acquisition in Welsh–English bilingual children. *Journal of Child Language*, 42, 146–179.
- Mayr, R., Morris, J., Mennen, I., & Williams, D. (2017). Disentangling the effects of long-term language contact and individual bilingualism: The case of monophthongs in Welsh and English. *International Journal of Bilingualism*, 21(3), 245–267.
- McAngus, C. (2015). Party elites and the search for credibility: Plaid Cymru and the SNP as new parties of government. *The British Journal of Politics and International Relations*, 18.
- McBride, M. (2022). Nationalism and ‘sectarianism’ in contemporary Scotland. *Ethnic and Racial Studies*, 45, 335–358.
- McCulloch, G. (2019). *Because internet: Understanding the new rules of language*. Riverhead Books.
- McMahon, A. (2020). *An introduction to English phonology* (Second ed.). Edinburgh University Press.
- McPeck, T. (2017). The effect of native orthography on borrowing phonology modeling and the nature of the phonotactic fringe: Evidence from Japanese. *Tampa Papers in Linguistics*, 4.
- Meredith, C. (2012). *The book of idiots*. Seren.

- Merriman, P., & Jones, R. (2009). 'Symbols of justice': The Welsh Language Society's campaign for bilingual road signs in Wales, 1967–1980. *Journal of Historical Geography*, 35, 350–375.
- Michnowicz, J., Hyler, A., Shepherd, J., & Trawick, S. (2018). Spanish in North Carolina: English-origin loanwords in a newly forming Hispanic community. In J. Reaser, E. Wilbanks, W. Wolfram, & K. Wojcik (Eds.), *Language diversity in the New South* (pp. 289–305). University of North Carolina Press.
- Miura, H. (2015). Phonemic variations in South Wales English vowels. *Studies in the humanities*, 97, 141–160.
- Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. *Cognition*, 133, 530–534.
- Monaghan, P., & Roberts, S. G. (2019). Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition*, 186, 147–158.
- Morgan, D. D. (1999). Christianity and national identity in twentieth-century Wales. *Religion, state and society*, 27, 327–342.
- Morin, R. (2003). Derivational morphological analysis as a strategy for vocabulary acquisition in Spanish. *The Modern Language Journal*, 87, 200–221.
- Morris, J. (2014). The influence of social factors on minority language engagement amongst young people: An investigation of Welsh–English bilinguals in North Wales. *International journal of the sociology of language*, 230, 65–89.
- Morris, J., Mayr, R., & Mennen, I. (2016). The role of linguistic background on sound variation in Welsh and Welsh English. In M. Durham & J. Morris (Eds.), *Sociolinguistics in Wales* (pp. 241–272). Palgrave Macmillan.
- Morris Jones, B. (2010). *Tense and aspect in informal Welsh*. De Gruyter Mouton.
- Morys, T. (2001). Cerdd Dafod: A poet introduces a Welsh metrical tradition. *EnterText*, 2.2, 114–120.
- Moschkovich, J. (2007). Using two languages when learning mathematics. *Educational studies in mathematics*, 64, 121–144.
- Mousikou, P., Beyersmann, E., Ktori, M., Javourey-Drevet, L., Crepaldi, D., Ziegler, J. C., Grainger, J., & Schroeder, S. (2020). Orthographic consistency influences morphological processing in reading aloud: Evidence from a cross-linguistic study. *Developmental Science*, 23(6).
- Muehlmann, S. (2008). "Spread your ass cheeks" and other things that should not be said in indigenous languages. *American Ethnologist*, 35, 34–48.

- Musk, N. (2010). Code-switching and code-mixing in Welsh bilinguals' talk: Confirming or refuting the maintenance of language boundaries? *Language, Culture, and Curriculum*, 23, 179–197.
- Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.
- Myers-Scotton, C. (1992). Comparing codeswitching and borrowing. *Journal of Multilingual and Multicultural Development*, 13, 19–39.
- Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.
- Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press.
- Myers-Scotton, C., & Bolonyai, A. (2001). Calculating speakers: Codeswitching in a rational choice model. *Language in Society*, 30, 1–28.
- Neal, S., & Agyeman, J. (2006). *The new countryside? Ethnicity, nation and exclusion in contemporary rural Britain*. Policy Press.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48, 443–453.
- Newcombe, L. P. (2007). *Social context and fluency in L2 learners: The case of Wales*. Multilingual Matters.
- Ngcobo, S. (2013). Educators' attitudes towards the role of isiZulu in education: Additive rather than exclusive. *Southern African Linguistics and Applied Language Studies*, 31, 185–205.
- Nguyen, L. (2018). Borrowing or code-switching? Traces of community norms in Vietnamese–English speech. *Australian Journal of Linguistics*, 38, 443–466.
- Ní Ghearáin, H. (2011). The problematic relationship between institutionalised Irish terminology development and the Gaeltacht speech community: Dynamics of acceptance and estrangement. *Language Policy*, 10, 305–323.
- Ó Ciosáin, N. (2004). Print and Irish, 1570–1900: An exception among the Celtic languages? *Radharc*, 5, 73–106.
- Ó hIfearnáin, T. (2007). Manx orthography and language ideology in the Gaelic continuum. In J.-M. Eloy & T. Ó hIfearnáin (Eds.), *Langues proches–langues collatérales, actes du colloque international réuni á limerick* (pp. 159–170). Editions l'Harmattan.
- OED terminology. (n.d.). Retrieved 2023, from <https://www.oed.com/information/understanding-entries/oed-terminology/>

- Olausson, U. (2017). The reinvented journalist: The discursive construction of professional identity on Twitter. *Digital Journalism*, 5, 61–81.
- O'Malley, T., Allan, S., & Thompson, A. (1995). Tokens of antiquity: The newspaper press and the shaping of national identity in Wales, 1870–1900. *Tokens of antiquity: The newspaper press and the shaping of national identity in Wales, 1870–1900 Studies in Newspaper and Periodical History*, 3, 127–152.
- Orgraff yr iaith Gymraeg: Rhan 1 (3rd ed.). (1987). Gwasg Prifysgol Cymru.
- Ortega, A., Urla, J., Amorrrortu, E., Goirigolzarri, J., & Uranga, B. (2015). Linguistic identity among new speakers of Basque. *International Journal of the Sociology of Language*, 231, 85–105.
- Osmond, J. (2002). Welsh civil identity in the twenty-first century. In D. C. Harvey, R. Jones, N. McNroy, & C. Milligan (Eds.), *Celtic geographies: Old culture, new times* (pp. 69–90). Routledge.
- Ottovordemgentschenfelde, S. (2017). “organizational, professional, personal”: An exploratory study of political journalists and their hybrid brand on Twitter. *Journalism*, 18, 64–80.
- Parkwell, C. (2019). Emoji as social semiotic resources for meaning-making in discourse: Mapping the functions of the toilet emoji in Cher's tweets about Donald Trump. *Discourse, Context & Media*, 30.
- Parry, D. (1990a). The conservative English dialects of North Camarthenshire. In N. Coupland (Ed.), *English in Wales* (pp. 142–150). Multilingual Matters.
- Parry, D. (1990b). The conservative English dialects of South Pembrokeshire. In N. Coupland (Ed.), *English in Wales* (pp. 151–161). Multilingual Matters.
- Parry-Williams, T. H. (1923). *The English element in Welsh: A study of English loan-words in Welsh*. Honourable Society of Cymmrodorion.
- Paton, B. (1995). New-word lexicography and the OED. *Dictionaries: Journal of the Dictionary Society of North America*, 16(1), 79–89.
- Paugh, A. L. (2019). Negotiating language ideologies through imaginary play: Children's code choice and rescaling practices in Dominica, West Indies. *Journal of Pragmatics*, 144, 78–91.
- Paulasto, H., Penhallurick, R., & Jones, B. A. (2021). *Welsh English*. De Gruyter Mouton.
- Pavlenko, A. (2009). Language conflict in post-Soviet linguistic landscapes. *Journal of Slavic Linguistics*, 17, 247–274.

- Pearson, B. Z. (1993). Predictive validity of the Scholastic Aptitude Test (SAT) for Hispanic bilingual students. *Hispanic Journal of Behavioral Sciences*, 15, 342–356.
- Penhallurick, R. (2008). Welsh English: Phonology. In B. Kortmann & C. Upton (Eds.), *Varieties of English 1: The British Isles* (pp. 105–121). Mouton de Gruyter.
- Penhallurick, R. (1993). Welsh English: A national language? *Dialectologia et Geolinguistica*, 1, 28–46.
- Percillier, M., & Paulin, C. (2017). Postcolonial literature and world Englishes: A corpus-based approach of modes of representation of the non-standard in writing. *International Journal of Literary Linguistics*, 6.
- Pérez, A., & Duñabeitia, J. A. (2019). Speech perception in bilingual contexts: Neuropsychological impact of mixing languages at the inter-sentential level. *Journal of neurolinguistics*, 51, 258–267.
- Pérez Cañado, M. L. (2005). English and Spanish spelling: Are they really different? *The Reading Teacher*, 58, 522–530.
- Pérez-Sabater, C., & Moffo, G. M. (2019). Managing identity in football communities on Facebook: Language preference and language mixing strategies. *Lingua*, 225, 32–49.
- Pfaff, C. W. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 55, 291–318.
- Phillipson, R., & Skutnabb-Kangas, T. (2013). Linguistic imperialism and endangered languages. In T. K. Bhatia & W. C. Ritchie (Eds.), *The handbook of bilingualism and multilingualism* (2nd ed., pp. 495–516). Blackwell.
- Phipps, A. (2019). *Decolonising multilingualism*. Multilingual Matters.
- Piccardo, E. (2018). Plurilingualism: Vision, conceptualization, and practices. In P. P. Trifonas & T. Aravossitas (Eds.), *Handbook of research and practice in heritage language education* (pp. 207–225). Springer.
- Piirainen, E. (2008). Phraseology in a European framework: A cross-linguistic and cross-cultural research project on widespread idioms. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 243–258). John Benjamins.
- Pike, K. L. (1967). *Language in relation to a unified theory of the structure of human behavior*. Mouton & Co.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: Toward a typology of code-switching. *Linguistics*, 18, 581–618.

- Poplack, S. (1987). Contrasting patterns of code-switching in two communities. In E. Wande, J. Anward, B. Nordberg, L. Steensland, & M. Thelander (Eds.), *Aspects of multilingualism* (pp. 51–77). Borgströms.
- Poplack, S. (2012). What does the nonce borrowing hypothesis hypothesize? *Bilingualism: Language and Cognition*, 15, 644–648.
- Poplack, S. (2017). L'anglicisme chez nous: Une perspective sociolinguistique. *Les anglicismes: Des emprunts à intérêt variable? Recueil des actes: Colloque du réseau des Organismes francophones de politique et d'aménagement linguistiques (OPALE)*, 375–403.
- Poplack, S., & Dion, N. (2012). Myths and facts about loanword development. *Language Variation and Change*, 24(3), 279–315.
- Poplack, S., & Meechan, M. (1998). Introduction: How languages fit together in codemixing. *International journal of bilingualism*, 2(2), 127–138.
- Poplack, S., & Sankoff, D. (1984). Borrowing: The synchrony of integration. *Linguistics*, 22, 99–136.
- Poplack, S., Sankoff, D., & Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26, 47–104.
- Pound, L. (1925). The krazie for “K”. *American speech*, 1, 43–44.
- Price, A. R. (2020). *Is inverted diglossia coming to Wales? Domain use and language attitudes among Welsh-speaking youth* [PhD Thesis]. Bangor University.
- Prince, A., & Smolensky, P. (2002). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers Optimality Archive.
- Pritchard, A., & Morgan, N. (2005). Representations of ‘ethnographic knowledge’: Early comic postcards. In A. Jaworski & A. Pritchard (Eds.), *Discourse, communication, and tourism* (pp. 53–78). Multilingual Matters.
- Pritchard, A., & Morgan, N. J. (2001). Culture, identity and tourism representation: Marketing Cymru or Wales? *Tourism Management*, 22(2), 167–179.
- Pryce, H. (2011). Culture, identity, and the Medieval Revival in Victorian Wales. *Proceedings of the Harvard Celtic Colloquium*, 31, 1–40.
- Prys, D. (2003). Providing the terms: Standardizing terms for education in Wales. *Speaking in tongues: languages of lifelong learning: Proceeding of the 33rd annual conference of SCUTREA*, 139–142.
- Putter, A. (2010). Multilingualism in England and Wales, c. 1200: The testimony of Gerald of Wales. In C. Kleinheinz & K. Busby (Eds.), *Medieval multilingualism: The francophone world and its neighbours* (pp. 83–105). Brepols.

- Quintanilla, M. M. (2018). The status of nouns in Yanito: Borrowings or code-switches? *Taking Stock to Look Ahead: Celebrating Forty Years of English Studies in Spain*, 215–224.
- Ramírez, G., Chen, X., & Pasquarella, A. (2013). Cross-linguistic transfer of morphological awareness in Spanish-speaking English language learners: The facilitating effect of cognate knowledge. *Topics in Language Disorders*, 33, 73–92.
- Raudszus, H., Segers, E., & Verhoeven, L. (2021). Use of morphological and contextual cues in children's lexical inferencing in L1 and L2. *Reading and Writing*, 34, 1513–1538.
- Reid, J. (1977). God Save the Queen (Sex Pistols).
https://en.wikipedia.org/wiki/File:Sex_Pistols_-_God_Save_the_Queen.jpg
- Revers, M. (2014). The Twitterization of news making: Transparency and journalistic professionalism. *Journal of Communication*, 64, 806–826.
- Reyes, R., III. (2008). 'Cheating' as good pedagogy: Bilingual teachers defying English-only to foster student achievement. *Multicultural Perspectives*, 10, 209–213.
- Ricks, C. (2011). *Dylan's visions of sin*. Canongate Books.
- Roberts, G. T. (1996). 'Under the hatches': English Parliamentary Commissioners' views of the people and language of mid-nineteenth-century Wales. In B. Schwarz (Ed.), *The expansion of England: Race, ethnicity and cultural history* (pp. 169–194). Routledge.
- Roberts, H. G. (2009). *Embodying identity: Representations of the body in Welsh literature*. University of Wales Press.
- Rodriguez, J. R. (2017). Against ethnocentrism and toward translanguaging in literacy and English education. *Isms in language education: Oppression, intersectionality and emancipation*, 98–121.
- Rosa, J. (2018). *Looking like a language, sounding like a race: Raciolinguistic ideologies and the learning of Latinidad*. Oxford University Press.
- Rossi, E., Newman, S., Kroll, J. F., & Diaz, M. T. (2018). Neural signatures of inhibitory control in bilingual spoken production. *Cortex*, 108, 50–66.
- Rottet, K. J., & Morris, S. (2018). *Comparative stylistics of Welsh and English: Arddulleg y Gymraeg*. University of Wales Press.
- Rowlands, J. (1999). Bards, poets and unofficial poets: A bird's-eye view of the Welsh-language poetic scene. *The Poetry Ireland Review*, 62, 89–92.

- Rowlands, S. P. (2001). Houdinis of the strict metres: The poetry of Twm Morys and Emyr Lewis. *Literary Review*, 44, 381–388.
- Rüdiger, S. (2018). Mixed feelings: Attitudes towards English loanwords and their use in South Korea. *Open Linguistics*, 4, 184–198.
- Rudra, K., Rijhwani, S., Begum, R., Bali, K., Choudhury, M., & Ganguly, N. (2016). Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1131–1141.
- Russell, P. (2013). Externarum linguarum excellens: The rhetoric and reality of the languages of Gruffudd ap Cynan, ruler of Gwynedd (†1137). In J. A. Jefferson, A. Putter, & A. Hopkins (Eds.), *Multilingualism in medieval Britain (c. 1066–1520) sources and analysis* (pp. 73–87). Brepols.
- Rymes, B., & Smail, G. (2021). Citizen sociolinguists scaling back. *Applied Linguistics Review*, 12, 419–444.
- Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language variation and change*, 2(1), 71–101.
- Sarkar, M., & Winer, L. (2006). Multilingual codeswitching in Quebec rap: Poetry, pragmatics and performativity. *International Journal of Multilingualism*, 3, 173–192.
- Scheu, U. D. (2000). Cultural constraints in bilinguals' codeswitching. *International Journal of Intercultural Relations*, 24, 131–150.
- Schmidt, C. K. (2009). Loanwords in Japanese. In M. Haspelmath & U. Tadmor (Eds.), *Loanwords in the world's languages: A comparative handbook* (pp. 545–574). Walter de Gruyter.
- Schrijver, P. (2007). What Britons spoke around 400 AD. In *Britons in Anglo-Saxon england* (pp. 165–171). The Boydell Press.
- Scotton, C. M., & Okeju, J. (1972). Loan word integration in Ateso. *Anthropological Linguistics*, 14, 368–382.
- Scourfield, J., & Davies, A. (2005). Children's accounts of wales as racialized and inclusive. *Ethnicities*, 5, 83–107.
- Sebba, M. (1998). A congruence approach to the syntax of codeswitching. *International Journal of Bilingualism*, 2, 1–19.
- Sebba, M. (2000). Writing switching in British Creole. In M. Martin-Jones & K. Jones (Eds.), *Multilingual literacies* (pp. 171–188). John Benjamins.
- Sebba, M. (2003). Spelling rebellion. In J. K. Androutsopoulos & A. Georgakopoulou (Eds.), *Discourse constructions of youth identities* (pp. 151–172). John Benjamins.

- Sebba, M. (2006). Ideology and alphabets in the former USSR. *Language Problems & Language Planning*, 30, 99–125.
- Sebba, M. (2012). Orthography as literacy: How Manx was ‘reduced to writing’. In A. Jaffe, J. Androutsopoulos, M. Sebba, & S. Johnson (Eds.), *Orthography as social action: Scripts, spelling, identity and power* (pp. 161–175). Walter de Gruyter.
- Segumpan, R. G., & Abu Zahari, J. S. (2023). Reserach directions in *CommSocMed*: Quo vadis? In R. G. Segumpan & J. S. Abu Zahari (Eds.), *Mixed methods perspectives on communication and social media research* (pp. 273–274). Routledge.
- Selleck, C. L. R. (2013). Inclusive policy and exclusionary practice in secondary education in Wales. *International Journal of Bilingual Education and Bilingualism*, 16, 20–41.
- Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an “outlier” orthography. *Psychological bulletin*, 134, 584–615.
- Sheppard, L. (2018). *Y Gymru ‘ddu’ a’r ddalen ‘wen’: Aralledd ac amlddiwylliannedd mewn ffuglen Gymreig, er 1990*. University of Wales Press.
- Sheppard, L., & Byrne, A. (2017). A critical minefield: The haunting of the Welsh working class novel. In J. Goodridge & B. Keegan (Eds.), *A history of British working class literature* (pp. 385–388). Cambridge University Press.
- Shore, S. (2015). Register in Systemic Functional Linguistics. In Frog & A. Agha (Eds.), *Registers of communication* (pp. 54–76). Finnish Literature Society.
- Simon, D.-L. (2001). Towards a new understanding of codeswitching in the foreign language classroom. In R. Jacobson (Ed.), *Codeswitching worldwide* (pp. 311–342, Vol. 2). Mouton de Gruyter.
- Smith, J. J. (1993). Dialectal variation in Middle English and the actuation of the Great Vowel Shift. *Neuphilologische Mitteilungen*, 259–277.
- Smith-Christmas, C. (2017). ‘Is it really for talking?’: The implications of associating a minority language with the school. *Language Culture and Curriculum*, 30, 32–47.
- Sonesson, G. (2017). Greimasean phenomenology and beyond: From isotopy to time consciousness. *Semiotica*, 93–113.
- Spencer, L. H., & Hanley, J. R. (2003). Effects of orthographic transparency on reading and phoneme awareness in children learning to read in Wales. *British Journal of Psychology*, 94, 1–28.

- Spencer, L. H., & Hanley, J. R. (2004). Learning a transparent orthography at five years old: Reading development of children during their first year of formal reading instruction in Wales. *Journal of Research in reading*, 27, 1–14.
- Spitzmüller, J. (2012). Floating ideologies: Metamorphoses of graphic “Germanness”. In A. Jaffe, J. Androutsopoulos, M. Sebba, & S. Johnson (Eds.), *Orthography as social action: Scripts, spelling, identity and power* (pp. 255–288). Walter de Gruyter.
- Sridhar, S. N. (1978). On the functions of code-mixing in Kannada. *International Journal of the Sociology of Language*, 16, 109–118.
- Sridhar, S. N. (2008). Language modernization in Kannada. In B. B. Kachru, K. Yamuna, & S. N. Sridhar (Eds.), *Language in South Asia* (pp. 327–342). Cambridge University Press.
- Stadthagen-González, H., Parafta Couto, M. C., Párraga, C. A., & Damian, M. F. (2019). Testing alternative theoretical accounts of code-switching: Insights from comparative judgments of adjective–noun order. *International Journal of Bilingualism*, 23(1), 200–220.
- Stammers, J. R. (2010). *The integration of English-origin verbs into Welsh: A contribution to the debate over distinguishing between code-switching and lexical borrowing*. Verlag Dr. Müller.
- Stammers, J. R., & Deuchar, M. (2012). Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition*, 15, 630–643.
- Starks, D. (2005). The effects of self-confidence in bilingual abilities on language use: Perspectives on Pasifika language use in South Auckland. *Journal of Multilingual and Multicultural Development*, 26(6), 533–550.
- Staunton, M. D. (2005). Trojan horses and friendly faces: Irish Gaelic typography as propaganda. *Revue LISA*, 3, 85–98.
- Stell, G., & Yakpo, K. (2015). Elusive or self-evident? Looking for common ground in approaches to code-switching. In G. Stell & K. Yakpo (Eds.), *Code-switching between structural and sociolinguistic perspectives* (pp. 1–16). De Gruyter.
- Stevenson, A. D. (2015). ‘Why in this bilingual classroom... hablamos más Español?’ Language choice by bilingual science students. *Journal of Latinos and education*, 14, 25–39.
- Stoll, D. (1982). The Summer Institute of Linguistics and indigenous movements. *Latin American Perspectives*, 9, 84–99.

- Strandberg, J. A. E. (2020). "Nordic Cool" and writing system mimicry in global linguistic landscapes. *Lingua*, 235.
- Suppe, F. (2022). Medieval Welsh ethnic nicknames and implications for the Welsh view of their geopolitical context, 1050–1400. In C. Raffensperger (Ed.), *Authorship, worldview, and identity in medieval Europe* (pp. 327–345). Routledge.
- Sutherland, P. (2015). Writing system mimicry in the linguistic landscape. *SOAS Working Papers in Linguistics*, 17, 147–167.
- Suyuti, A. (2018). Spelling variations of standard Arabic loanwords in Indonesian and Malay. *The Asian Conference on Arts & Humanities 2018: Official Conference Proceedings*, 385–394.
- Tamargo, R. E. G., Kroff, J. R. V., & Dussias, P. E. (2016). Examining the relationship between comprehension and production processes in code-switched language. *Journal of Memory and Language*, 89, 138–161.
- Tamburelli, M. (2021a). Contested languages and the denial of linguistic rights in the 21st century. In M. Tamburelli & M. Tosco (Eds.), *Contested languages: The hidden multilingualism of Europe* (pp. 21–40). John Benjamins.
- Tamburelli, M. (2021b). Taking taxonomy seriously in linguistics: Intelligibility as a criterion of demarcation between languages and dialects. *Lingua*, 256.
- Tamir, Y. (2019). Not so civic: Is there a difference between ethnic and civic nationalism? *Annual Review of Political Science*, 22(1), 419–434.
- Tatman, R. (2015). #Go awn: Sociophonetic variation in variant spellings on Twitter. *Working Papers of the Linguistics Circle of the University of Vienna*, 25, 97–108.
- Taverniers, M. (2002). *Systemic-functional linguistics and the notion of grammatical metaphor* [PhD Thesis]. University of Gent.
- Taverniers, M. (2008). Hjelmslev's semiotic model of language: An exegesis. *Semiotica*, 171, 367–394.
- Taverniers, M. (2011). The syntax–semantics interface in systemic functional grammar: Halliday's interpretation of the Hjelmslevian model of stratification. *Journal of Pragmatics*, 43, 1100–1126.
- Templeton, S. (2012). Teaching and learning morphology: A reflection on generative vocabulary instruction. *Journal of Education*, 192, 101–107.
- Tench, P. (1990). The pronunciation of English in Abercrave. In N. Coupland (Ed.), *English in Wales* (pp. 130–141). Multilingual Matters.
- Thomas, A. R. (1987). A spoken standard for Welsh: Description and pedagogy. *International journal of the sociology of language*, 66, 99–113.

- Thomas, A. R. (1992). The Welsh language. In D. MacAulay (Ed.), *The Celtic languages* (pp. 251–345). Cambridge University Press.
- Thomas, A. (1997). Language policy and nationalism in Wales: A comparative analysis. *Nations and Nationalism*, 3, 323–344.
- Thomas, B. (2000). The riches of the past? Recording Welsh dialects. In G. H. Jenkins & M. A. Williams (Eds.), *Let's do our best for the ancient tongue: The Welsh language in the twentieth century* (pp. 421–438). University of Wales Press.
- Thomas, C. H. (1982). Registers in Welsh. *International Journal of the Sociology of Language*, 35, 87–115.
- Thomas, E. M., Apolloni, D., & Lewis, G. (2014a). The learner's voice: Exploring bilingual children's selective language use and perceptions of minority language competence. *Language and Education*, 28(4), 340–361.
- Thomas, E. M., Apolloni, D., & Lewis, G. (2014b). The learner's voice: Exploring bilingual children's selective language use and perceptions of minority language competence. *Language and education*, 28, 340–361.
- Thomas, E. M., & Roberts, D. B. (2011). Exploring bilinguals' social use of language inside and out of the minority language classroom. *Language and Education*, 25(2), 89–108.
- Thomas, E. M., & Webb-Davies, P. (2017). *Agweddau ar ddwyieithrwydd*. Coleg Cymraeg Cenedlaethol.
- Thomas, P. W. (1996). *Gramadeg y Gymraeg*. Gwasg Prifysgol Cymru.
- Tosco, M. (2021). Democracy: A threat to language diversity? In M. Tamburelli & M. Tosco (Eds.), *Contested languages: The hidden multilingualism of Europe* (pp. 41–58). John Benjamins.
- Townsend, L., & Wallace, C. (2016). *Social media research: A guide to ethics*. University of Aberdeen.
- Treffers-Daller, J. (1991). Towards a uniform approach to code-switching and borrowing. *Papers for the workshop on constraints, conditions and models*, 259–279.
- Trudgill, P., & Hannah, J. (2017). *International English* (Sixth ed.). Routledge.
- Turner, M. L. (2020). *Rhai o gymunedau'n llên: Y lleol a'r cenedlaethol* [PhD Thesis]. Prifysgol Bangor.
- Tyler, R. L. (2018). Migrant culture maintenance: The Welsh experience in Martins Ferry, Belmont County, Ohio, 1900–1940. *Ohio History*, 125, 70–94.

- Urrieta, L., Jr. (2004). *Assistencialism* and the politics of high-stakes testing. *The Urban Review*, 36, 211–226.
- Üstünel, E., & Seedhouse, P. (2005). Why that, in that language, right now? Code-switching and pedagogical focus. *International Journal of Applied Linguistics*, 15, 302–325.
- Van Avermaet, P. (2009). Fortress Europe? Language policy regimes for immigration and citizenship. In G. Hogan-Brun, C. Mar-Molinero, & P. Stevenson (Eds.), *Discourses on language and integration* (pp. 15–44). John Benjamins.
- Van Haneghan, J. (2021). Exploratory factor analysis of text. In A. J. Onwuegbuzie & R. B. Johnson (Eds.), *The Routledge reviewer's guide to Mixed Methods Analysis* (pp. 25–36). Routledge.
- van Kesteren, R., Dijkstra, T., & de Smedt, K. (2012). Markedness effects in Norwegian–English bilinguals: Task-dependent use of language-specific letters and bigrams. *The quarterly journal of experimental psychology*, 65, 2129–2154.
- Van Vugt, W. E. (2006). *British Buckeyes: The English, Scots, and Welsh in Ohio, 1700–1900*. Kent State University Press.
- Van-Hagen, S. (2017). “But genius is the special gift of God!”: The reclamation of “natural genius” in the late-eighteenth-century verses of Ann Yearsley and James Woodhouse. In J. Goodridge & B. Keegan (Eds.), *A history of British working class literature* (pp. 55–69). Cambridge University Press.
- Vasileva, N. M., & Ammosova, O. N. (2019). On spelling of loanwords in the Yakut language. *Amazonia Investiga*, 8, 28–37.
- Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PloS one*, 11.
- Vendelin, I., & Peperkamp, S. (2006). The influence of orthography on loanword adaptations. *Lingua*, 116, 996–1007.
- Virues-Ortega, J., Martin, N., Schnerch, G., García, J. Á. M., & Mellichamp, F. (2015). A general methodology for the translation of behavioral terms into vernacular languages. *Behavioural Analysis*, 38, 127–135.
- Wagner, A., Marusek, S., & Yu, W. (2020). Sarcasm, the smiling poop, and E-discourse aggressiveness: Getting far too emotional with emojis. *Social Semiotics*, 30, 305–311.
- Walpole, J. R. (1974). Eye dialect in fictional dialogue. *College composition and communication*, 25, 191–196.

- Waltermire, M., & Valtierrez, M. (2019). Spontaneous loanwords and the question of lexical proficiency among Spanish–English bilinguals. *Hispania*, 102(3), 409–422.
- Walters, J. R. (1999, September). *A study of the segmental and suprasegmental phonology of Rhondda Valleys English* [PhD Thesis]. University of Glamorgan.
- Walters, J. R. (2001). English in Wales and a ‘Welsh Valleys’ accent. *World Englishes*, 20, 285–304.
- Wasserscheidt, P. (2014). Constructions do not cross languages: On cross-linguistic generalizations of constructions. *Constructions and Frames*, 6(2), 305–337.
- Wasserscheidt, P. (2016). Construction grammar and code-mixing. *Cognitive Perspectives on Bilingualism*, 65–91.
- Wasserscheidt, P. (2020). Explaining code-switching: Matrix language models vs. bilingual construction grammar. *Književni Jezik*, 31, 57–87.
- Watkins, D., Biggs, J., & Regmi, M. (1991). Does confidence in the language of instruction influence a student’s approach to learning? *Instructional Science*, 20, 331–339.
- Webb, H. (1995). *Harri Webb: Collected poems* (M. Stephens, Ed.). Gwasg Gomer.
- Wells, J. C. (1982a). *Accents of English 1: An introduction*. Cambridge University Press.
- Wells, J. C. (1982b). *Accents of English 2: The British Isles*. Cambridge University Press.
- Western Institute for Endangered Language Documentation. (n.d.). *Americanist transcription system*. http://wioldoc.org/?page_id=344
- Wijeratnem, S., Balasuriya, L., Sheth, A., & Doran, D. (2016). EmojiNet: Building a machine readable sense inventory for emoji. In E. Spiro & Y.-Y. Ahn (Eds.), *Social informatics: 8th International Conference, SocInfo 2016 Proceedings, Part I* (pp. 527–542). Springer.
- Willans, F. (2010). Classroom code-switching in a Vanuatu secondary school: Conflict between policy and practice. *International Journal of Bilingual Education and Bilingualism*, 14, 23–38.
- Williams, C. H. (1976). Non-violence and the development of the Welsh Language Society, 1962–c.1974. *Welsh History Review*, 8, 426–455.
- Williams, C. (1994). *Arfarniad o ddulliau dysgu ac addysgu yng nghyd-destun addysg uwchradd ddwyieithog* [PhD Thesis]. Prifysgol Cymru, Bangor.
- Williams, C. (2007). Revisiting the rural/race debates: A view from the Welsh countryside. *Ethnic and racial studies*, 30, 741–765.

- Williams, C. (2015). Experiencing rural Wales. In C. Williams, N. Evans, & P. O'Leary (Eds.), *A tolerant nation? Revisiting ethnic diversity in a devolved Wales* (pp. 251–276). University of Wales Press.
- Williams, D. G. (2014). Beyond invisibility: Uses of comparison in Welsh cultural nationalism. *Studies on National Movements*, 2.
- Williams, E. (2009). Language attitudes and identity in a North Wales town: “something different about Caernarfon”? *International Journal of the Sociology of Language*, 2009(195), 63–91.
- Willis, D. (2017). Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun *chdi*. *Journal of Linguistic Geography*, 5, 41–66.
- Willis, D. (2020). Using social-media data to investigate morphosyntactic variation and dialect syntax in a lesser-used language: Two case studies from Welsh. *Glossa*, 5, 1–33.
- Willis, P. (1988). Is the Welsh verbal noun a verb or a noun? *Word*, 39, 201–224.
- Wodak, R., Krzyżanowski, M., & Forchtner, B. (2012). The interplay of language ideologies and contextual cues in multilingual interactions: Language choice and code-switching in European Union institutions. *Language in Society*, 41, 157–186.
- Wohlgemuth, J. (2009). *A typology of verbal borrowings*. Walter de Gruyter.
- Wooldridge, I. (2005). *George Orwell's Animal Farm*. Dramatic Publishing.
<https://www.dramaticpublishing.com/animal-farm-ab6000>
- Xia, X. (2010). Application and normalization of lettered words in Chinese character system. *Asian Social Science*, 6, 24–27.
- Zappavigna, M. (2012). *Discourse of Twitter and social media*. Continuum.
- Zenner, E., & Van De Mierop, D. (2017). The social and pragmatic function of English in weak contact situations: Ingroup and outgroup marking in the Dutch reality TV show *Expeditie Robinson*. *Journal of Pragmatics*, 113, 77–88.
- Zibin, A. (2019). A phonological analysis of English loanwords inflected with Arabic morphemes in urban Jordanian spoken Arabic. *SAGE Open*, 9, 1–13.
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological bulletin*, 131, 3–29.
- Zimmer, B. (2008). “Ghoti” before Shaw. Retrieved April 23, 2008, from
<https://languagelog.ldc.upenn.edu/nll/?p=81>

Appendix A

Transcription conventions

In referring to the speech sounds of Welsh and English, this study uses the symbols of the International Phonetic Alphabet (IPA), with the addition of the non-IPA symbol ɨ, explained below. For vowel sounds, lexical set labels are also used, based on those introduced by Wells (1982a, pp. xviii–xiv) to compare accents of English. The lexical sets offer keywords to exemplify particular vowel sounds. This facilitates comparison of accents by providing a simple way to discuss phonemic contrasts which are present in some accents and absent in others, regardless of the phonetic realization of the phonemes in question. For example, British English Received Pronunciation (RP) [ɒ] is the vowel of both the lexical sets **LOT** and **CLOTH**, whereas General American (GA) has [ɑ:] in **LOT** and [ɔ:] in **CLOTH** (McMahon, 2020, p. 87). We can therefore say that GA has a **LOT CLOTH** split, and that RP does not, without needing to specify that RP **LOT CLOTH** vowel is different from both GA **LOT** and GA **CLOTH**.

A.i The vowels of English

The standard list of English vowels distinguishes 27 lexical sets, labelled by keywords, which were chosen to optimize descriptions of the distinctions between RP and GA (Wells, 1982a, p. xviii). The list may therefore fail to capture distinctions which are important for comparisons with other varieties, such as Welsh English (WE). The expanded list of lexical sets developed for this study takes into account systematic differences in WE varieties, drawing on a number of descriptive accounts (B. Collins & Mees, 1990; Connolly, 1981, 1990; Coupland, 1988, 1990; Durham, 2022; Iosad, 2017; Mayr et al., 2017; Miura, 2015; Morris et al., 2016; Parry, 1990a, 1990b;

Penhallurick, 2008; Tench, 1990; Trudgill & Hannah, 2017; Wells, 1982b)¹.

RP is the traditional label given for a Standard British English reference accent, which was never the primary acquired accent of more than a small percentage of the British population, but which was nonetheless generally taken as the norm for “correct” or “standard” British English pronunciation (Fabricius, 2005) and for teaching foreigners to speak English (Hannisdal, 2005). I have instead chosen to use the label “Prestige Southern-England English” (PSEE) as an umbrella term for the reference variety of British English described here, covering both RP and its late-twentieth- and twenty-first-century descendant, called Standard Southern British English (SSBE) by some linguists (Harrington et al., 2000; Lindsey, 2019; McMahon, 2020).

Although sometimes described as a non-geographic British accent (Hannisdal, 2005), RP is more precisely “an accent *of England* that is regionless” (Harrington et al., 2000, 65, *italics mine*). Although it may have been spoken by people living outside England as a result of its status as a nationwide ideal and a necessity for access to certain prestigious careers (Fabricius, 2005; Hannisdal, 2005), RP has traditionally been associated with centres of cultural and political power based in southern England, notably the parliament, the monarchy, the BBC, and the Universities of Oxford and Cambridge (Hannisdal, 2005; Harrington et al., 2000; Hinton, 2015). Its phonological basis, too, lies closest to the regional accents of the south-east of England (Hannisdal, 2005).

In contrast to RP’s status as a national standard, SSBE is slowly losing RP’s association with cultural and political prestige and, therefore, beginning to be recognized as a geographically-limited variety of southern English, although it still has significant prestige as the accent of younger English aristocracy and royalty (Lindsey, 2019). Using an umbrella label to cover RP and SSBE means that both the older national standard and its modern descendant can be included when we discuss non-Welsh English influences on loanword assimilation and adaptation by Cymrophones. The name “Prestige Southern-England English” acknowledges the role of social and political pressure in establishing these varieties at the top of the

¹ Where reasonable, the precise symbols used in the original sources have been harmonized so that differences in transcription do not lead to unnecessarily detailed lists of pronunciations with only minor variations. In particular, disyllabic sequences with a medial close vowel symbol have been normalized to contain a glide instead, so (for example) /eɪə/ and /erə/ would both be normalized to /ejə/.

accent hierarchy, and highlights the cultural Englishness of “regionless” RP.

The cultural prestige of PSEE pronunciations, and their prevalence in media and public discourse, means that they provide one potential target for the production of English-origin ELUs in Welsh. At least some Cymrophone commentators in the early twentieth century were keen for their compatriots to adopt an “English” accent when speaking English, rather than one of the stigmatized Welsh accents (E. D. Jones, 1933, pp. 50–51). Schoolteachers in Wales in the late-nineteenth and earlier twentieth centuries are also most likely to have used RP as their model, teaching the prestige accent of English to schoolchildren who were often acquiring the language exclusively through schooling and who therefore had no other models of pronunciation available to them (Walters, 2001).

As English became a primary language for Welsh people, distinctively Welsh accents of English developed. These accents have been influenced by the RP-like English taught in schools in the earlier period, by the accents of English regions closely neighbouring Wales, and by the Welsh language itself (Paulasto et al., 2021; Walters, 2001).

have always co-existed with prestige models of “correct” English pronunciation, nonetheless, and these regional accents also provide a potential target for the production of English-origin ELUs by Cymrophones. The second type of pronunciation represented here is therefore geographically Welsh.

The phonemes used to label the lexical sets in the list of English vowels are chosen with two aims. The first is to maximize the distinctiveness of the **WE** phoneme labels, even if this sometimes means choosing a label for **WE** which represents a geographically-limited realization. The **WANT** vowel is therefore labelled /ʌ/ for **WE**, to distinguish it from the **LOT** vowel, although the **WANT LOT** split (recorded by Connolly (e.g. 1981) in southern Wales) is not fully distributed throughout the country. The second aim is to preserve distinctions between PSEE and Welsh. On the whole, the phonemic labels proposed by Lindsey (2019) for SSBE will be used for PSEE, except where this might reduce the distinctions between PSEE and **WE** or Welsh transcriptions. So, for example, although the **TRAP** vowel is now most commonly a lower [a] in SSBE (Lindsey, 2019; McMahon, 2020) the symbol used here is the higher /æ/ characteristic of RP, to preserve a clear distinction from the Welsh vowel /a/. I also retain the more traditional symbol /ɜ:/ for the vowel of **NURSE**, rather than adopting Lindsay’s /ə:/, since it is important to distinguish between English schwa, as a default vowel to which unstressed vowels may be reduced, and Welsh schwa, which is a “full” vowel (Hannahs, 2013).

Alongside the phonemic labels, a range of recorded phonetic realizations of each vowel is given, one for PSEE and one for WE. There is no one canonical Welsh accent which can stand as representative of Welsh English (WE) as a whole (Coupland, 1990), and a range of documented realizations of each vowel is therefore listed below. The PSEE phones (from accounts of RP and of SSBE) are based on data from Fabricius (2005), Harrington et al. (2000), Hinton (2015), Lindsey (2019), and McMahon (2020). The WE phones are summarized from (B. Collins & Mees, 1990; Connolly, 1981, 1990; Coupland, 1988, 1990; Durham, 2022; Iosad, 2017; Mayr et al., 2017; Miura, 2015; Morris et al., 2016; Parry, 1990a, 1990b; Paulasto et al., 2021; Penhallurick, 2008; Tench, 1990; Trudgill & Hannah, 2017; Wells, 1982b)².

CLOTH encodes a distinction from **LOT** present in GA, but not in either PSEE or WE. **LOT**, in this study, will therefore be taken to cover both Wells' **LOT** and his **CLOTH**. The lax realization of the *happy* vowel in RP, [ɪ], is now distinctly old-fashioned (Lindsey, 2019). There seems no good reason to distinguish *happy* from **FLEECE** in this study, since the two sounds have essentially identical realizations for modern PSEE speakers, and seem not to have been distinct in WE in any of the recorded data. **FLEECE** will therefore be used to represent both Wells' **FLEECE** and his *happy*.

Following Walters (2001), I treat the **GOAT** and **FACE** vowels as separate phonemes from the **KNOW** and **BAIT** vowels. The former are usually monophthongs in WE, whereas the latter are realized as diphthongs, as in PSEE. A number of other additions to the list represent sounds which are realized with vowels in WE and with glides in PSEE, or vice versa. The **TUNE** vowel is /jʊu/ in PSEE: a combination of a glide and a diphthong. In Welsh Englishes, it is more usually realised as a simple diphthong: /ɪu/. Wells does not include /jʊu/ as distinct from /ʊu/, seeing the former as a combination of consonant and /ʊu/ rather than as a separate vowel phoneme (Wells, 1982a, p. 206). However, Wells's (1982a) discussion of the process termed 'Early Yod Dropping' suggests that distinguishing **TUNE**, **GOOSE**, and **FLEW**, and including the initial [j] in the PSEE **TUNE** phoneme, is reasonable.

Early Yod Dropping saw a development in southern English accents from the

² Where reasonable, the precise symbols used in the original sources have been harmonized so that differences in transcription do not lead to unnecessarily detailed lists of pronunciations with only minor variations. In particular, disyllabic sequences with a medial close vowel symbol have been normalized to contain a glide instead, so (for example) /eɪə/ and /erə/ would both be normalized to /ejə/.

Middle English diphthong [ɪʊ] to a glide + vowel sequence, [jʊ:]. Subsequently, the /j/ was lost in a number of contexts, so that words such as “blew” and “chew” have the same vowel as **GOOSE** in PSEE (Wells, 1982a). In Welsh Englishes, however, [ɪʊ] did not generally develop into [jʊ:], meaning that there was never a [j] to be lost in the relevant phonological contexts. As a result, **TUNE** and **FLEW** both contain the phoneme /ɪʊ/ in many Welsh varieties (Wells, 1982a, p. 207). The **GOOSE**~**FLEW** pair is therefore split in WE and merged in PSEE, while **FLEW**~**TUNE** is merged in WE and split in PSEE. The effects of Early Yod Dropping also motivate the addition of **POOR** to the list of lexical sets. Wells’ original **CURE** was intended to illustrate only the diphthong /ʊə/ in RP, without the preceding glide. However, Welsh Englishes tend to produce /jʊə/ as a distinct sequence, not as a combination of /j/ and /ʊə/: WE **CURE** is /ɪwə/, while **POOR** is /u[w]ə/.

Like **POOR**, the added item **BEER** was a diphthong in RP, and tends towards a long monophthong in SSBE (Lindsey, 2019), but is generally a disyllabic sequence with a central glide in WE. In some Welsh Englishes it is realized identically with **NEAR**, which appears in Wells’ original list. However, there is a notable WE variant for certain words which have /ɪə/ in PSEE, which realizes the vowel as [(j)œ:]. In some WE varieties, “nearly” is realized as [ˈnjœ:li:] or [ˈnœ:li:], rhyming with “surly” ([ˈsœ:li:]) rather than with “severely” ([sɛvˈɪjəli:]). The words “ear” and “year” are realized as /jœ:/ in such varieties (as are “here” and “hear” if the initial /h/ is realized with a zero variant) (Mayr, 2010; Walters, 2001). **BEER** is therefore added to represent the vowel of “severely”, as distinct from that of “nearly”, although these are only separate phonemes in some varieties of WE.

The vowels of **FIRE** and **TOWER** are consistently realized as two syllables, with a central glide, in Welsh Englishes. There is some justification for treating them as single phonemes, however, when we consider that the most common PSEE realizations of these sounds are either monophthongs or diphthongs, rather than triphthongs (although this is the traditional representation) or a V-C-V sequence with a central glide. This is true both for RP and for SBEE, and with a similar pattern as is found for **SQUARE**, **NEAR**, and **POOR**: where RP realized these phonemes as a centering diphthong, SBEE is developing a long monophthong (Lindsey, 2019).

Three unstressed vowels are distinguished by Wells: the vowels in the final syllables of **LETTER**, **COMMA**, and **HAPPY**. These represent most examples of vowels in unstressed syllables in PSEE. Many Welsh English accents resist the reduction or elision of unstressed vowels (e.g. Coupland, 1990, p. 246), perhaps as a result of influence from Welsh, which generally does not permit schwa in an unstressed

syllable (Wells, 1982b, p. 387). In most cases, however, these realizations are most easily represented by one of the short vowel phonemes, since the finer phonetic details of their production is not of immediate relevance. Nonetheless, WE resistance to vowel reduction has implications for both phonological and orthographic assimilation of English-origin ELUs, and is discussed in Section 4.i.d of the main text.

A.ii The vowels of Welsh

The Welsh graphemes <a>, <e>, <i>, <o>, <u>, <w> and <y> each represents (at least) two vowel sounds. There is disagreement about the correct phonemic representation of the vowels corresponding by each grapheme (such as the pair of vowels in “nen” and “sêr”, conventionally a “short” and “long” <e>), and particularly about whether these vowels differ in quantity, in quality, or both Hannahs (2013, p. 23). Like Iosad (2017), therefore, I consider it useful at this stage to have a way to refer to these related sets of vowels in general terms, without distinguishing the various phonemes in terms of the exact quality or quantity of their phonetic realizations. For convenience, therefore, I will refer to the set of phonemes conventionally represented by one of the seven graphemes <a>, <e>, <i>, <o>, <u>, <w>, and <y> as “vowel sets”.

Hannahs (2013) recognizes two phonemes in each vowel set (apart from <y>, which has one). P. W. Thomas (1996, p. 759) finds that the vowels of the <i> and <w> sets are triplets rather than pairs: a lax, short vowel, and a tense vowel with distinct long and short variants. Although in each case the tense and lax short vowels in Thomas’ system are in complementary distribution, with the short vowels in open syllables being tense [i] and [u], and those in other positions being lax [ɪ] and [ʊ], Thomas nonetheless appears to consider them as separate phonemes rather than as allophones. Hannahs and Thomas agree, however, in distinguishing the vowels in each set primarily by quantity.

G. E. Jones (1984) describes vowel sets for southern varieties of Welsh whose members differ primarily in quality, rather than quantity as in Hannahs’ and Thomas’ accounts. The paired items in Jones’ sets are generally distinguished by tongue height (or tongue retraction, in the case of the <a> set), leading G. E. Jones (1984, p. 53) to identify two series of vowels: an “open” series and a “close” series. The “short” vowels of Hannahs’ and Thomas’ vowel sets correspond to Jones’ “open” vowels, and their “long” vowels to Jones’ “close” vowels. Jones’ “open” vowels are always short. The “close” vowels are always long in stressed syllables. Elsewhere, quantity is not a distinctive characteristic, and the “close” vowels may be realized as short in some

contexts (Iosad, 2017; G. E. Jones, 1984).

Although Jones finds a definite qualitative distinction between the phonemes of each vowel set in southern Welsh, he acknowledges that this distinction is reduced, and at times absent, in many northern varieties, which instead have quantity as their major distinctive characteristic (G. E. Jones, 1984, p. 57). Nonetheless, Mayr and Davies (2011) and Bell et al. (2023) find differences of both quantity and quality in the paired phonemes of the vowel sets produced by their northern Welsh respondents, except for the <u> and <a> sets, whose members differ only in quantity. Since quantity appears to be a more stable distinction than quality according to most accounts, this study will refer to “short” and “long” vowels in Welsh.

For this study, Welsh will be described using symbols which mark a difference of quality within each vowel set other than the <u> set, largely following G. E. Jones (1984) for southern Welsh, and Bell et al. (2023) for northern Welsh. Length will be marked on monophthongs, and occasionally also on the constituents of diphthongs, since this helps to highlight similarities and differences between the vowels of English and Welsh. The resulting system of phonemic symbols should not, however, be taken as a positive assertion that there is always a strong difference in quality between the phonemes of every vowel set. There may be significant, moderate, or almost no variation in quality amongst the actual realizations by individual speakers of phonemes in a given set.

The phonemic symbols used for Welsh are primarily those of Jones and of Bell et al., with a few changes in order to create a symmetrical system of phonemic symbols for the Welsh vowels³. Firstly, I borrow the symbol [ɪ], representing a near-close central high vowel, from the Americanist transcription system developed by the Western Institute for Endangered Language Documentation (*WIELD*, year). This is used to represent the short vowel of the <u> set, as a counterpart to near-close [i] for the short vowel of the <i> set.

Secondly, I use [ɐ] and [a] for the short and long vowels of the <a> set, where Jones has [a] and [ɑ] and Bell et al. have [ɐ] and [ɛ:]. In reality, it seems likely that there is little difference in quality between these vowels for most speakers of Welsh (Iosad, 2017). A difference in quality may be found in some varieties of southern Welsh (Mayr & Davies, 2011), but it is generally much less dramatic than the

³ Jones' choice of symbols has also been modernized, by replacing the deprecated “ω” with the now-standard “o” to represent the near-close near-back rounded vowel (*Handbook of the International Phonetic Association*, 1999).

differences which hold between the vowel pairs of the <e>, <i>, <o>, and <w> sets⁴. Nonetheless, using /ɐ/ and /a/ for the <a> vowels preserves a symmetry with the other sets, which are also represented as pairs of tense and lax vowels. The phonemic symbols which will be used from this point onwards to describe Welsh vowels are therefore as follows:

- <a> /ɐ/~/a:/
- <e> /ɛ/~/e:/
- <i> /ɪ/~/i:/
- <o> /ɔ/~/o:/
- <u> /ʊ/~/u:/
- <w> /ʊ/~/u:/
- <y> /ə/

The first phoneme in each pair is the “short” vowel of the set, and the second the “long”.

Evidence for the diphthongs of Welsh listed below is drawn from the same sources as already cited for the monophthongs.

⁴ See e.g. Mayr and Davies (2011), Mayr et al. (2017) and Bell et al. (2023) for acoustic analyses of these vowels.

A.iii Lexical sets

A.iii.a English vowels

Items in small capital letters are those of the standard list introduced by Wells; items set in *italic capitals*, with a preceding asterisk, are my additions. Items marked with an obelus are those omitted from the list used for this study.

Two English labels retained in this study have two syllables: **LETTER**, **COMMA**. The vowel represented is the vowel of the final syllable. The Welsh label **BYWYD**, in contrast, represents the vowel of its first syllable.

Table A.1: Lexical sets for English vowels with phonemic and phonetic details.

	PSEE		WE	
Short Vowels				
KIT	/ɪ/	[ɪ ~ i]	/ɪ/	[ɪ ~ e]
DRESS	/ɛ/	[ɛ ~ e]	/ɛ/	[ɛ ~ æ]
TRAP	/æ/	[æ ~ a]	/ɐ/	[æ ~ a(:) ~ ɐ ~ ɑ(:)]
LOT	/ɔ/	[ɒ ~ ɔ]	/ɔ/	[ɔ ~ ɒ ~ ɑ]
† CLOTH	/ɔ/	[ɒ ~ ɔ(:)]	/ɔ/	[ɔ ~ ɒ]
* WANT	/ɔ/	[ɒ ~ ɔ]	/ʌ/	[ʌ ~ ɔ ~ ɒ]
STRUT	/ʌ/	[ʌ]	/ə/	[ə ~ ʌ ~ ʊ]
FOOT	/ʊ/	[ʊ ~ ɵ]	/ʊ/	[ʊ ~ ɵ ~ ɤ ~ ʌ]
Reduced Vowels				
LETTER	/ɝ/	[ɝ ~ ə]	/ɝ/	[ɝ ~ ə ~ ɪ ~ ɛr]
COMMA	/ə/	[ə]	/ə/	[ɐ ~ ə]
† happy	/ii/	[ii ~ i]	/i:/	[i:]
R-linking Monophthongs				
BATH	/ɑ:/	[ɑ:]	/ɐ/	[a ~ ɐ ~ ɑ:]
PALM	/ɑ:/	[ɑ:]	/a:/	[a: ~ æ: ~ ɛ:]
START	/ɑ:/	[ɑ:]	/a:/	[a(:) ~ æ: ~ ɛ: ~ ɑ]
THOUGHT	/o:/	[ɔ: ~ o:]	/ɒ:/	[o: ~ ɒ: ~ ʌ:]
NORTH	/o:/	[ɔ: ~ o:]	/ɒ:/	[ɔ:(r) ~ ɒ:(r)]
FORCE	/o:/	[ɔ: ~ o:]	/o:/	[o:(r) ~ ɔ:(r) ~ oə]
NURSE	/ɜ:/	[ə: ~ ɜ: ~ ɐ:]	/œ:/	[ə: ~ ø: ~ œ: ~ ɜ: ~ ʌr]
R-linking Diphthongs				

NEAR	/ɪə/	[ɪə ~ ɪ:]	/jœ:/	[i(j)ə ~ (j)ɜ: ~ (j)œ:]
* <i>BEER</i>	/ɪə/	[ɪə ~ ɪ:]	/ijə/	[i(j)ə]
SQUARE	/ɛə/	[ɛə ~ ɛ:]	/e:/	[ɛ: ~ e: ~ ɛə ~ eə ~ ɛr]
* <i>LAIR</i>	/ɛə/	[ɛə ~ ɛ:]	/ejə/	[ɛ: ~ e: ~ ejə]
* <i>FIRE</i>	/aə/	[aɪə ~ aə ~ ɑjə ~ a:]	/əjə/	[əjə ~ ʌjə ~ ajə]
* <i>TOWER</i>	/ɑə/	[ɑʊə ~ aə ~ awə ~ ɑ:]	/əwə/	[əwə ~ ʌwə ~ awə]
* <i>POOR</i>	/ʊə/	[ʊə ~ o:]	/uwə/	[uwə ~ ʌ: ~ o:]
* <i>BREWER</i>	/ʊə/	[ʊə]	/ɪwə/	[ɪwə ~ uwə]
Yod-initial Vowels				
* <i>TUNE</i>	/jʊu/	[jʊu ~ ju:]	/ɪu/	[(j)ɪu ~ (j)eu]
CURE	/jʊə/	[jʊə ~ jɔ:]	/ɪwə/	[ɪwə ~ juwə ~ jʌ:]
Diphthongs > /i/				
FLEECE	/i:/	[i:]	/i:/	[i: ~ iə]
FACE	/ei/	[eɪ ~ ei]	/e:/	[e:]
* <i>BAIT</i>	/ei/	[eɪ ~ ei]	/ei/	[e: ~ ei]
PRICE	/ai/	[aɪ ~ ai]	/əi/	[əi ~ ʌi ~ ei]
CHOICE	/oi/	[oi]	/ɒi/	[ɒi ~ ɒi ~ ʌi]
Diphthongs > /u/				
MOUTH	/au/	[aʊ ~ au]	/əu/	[əu ~ ʌu ~ au ~ ɛu]
GOAT	/əu/	[əu ~ əʊ ~ ɒʊ ~ eʊ]	/o:/	[o: ~ oə]
* <i>KNOW</i>	/əu/	[əʊ ~ əu ~ ɒʊ ~ eʊ]	/ou/	[o: ~ ou ~ əu]
GOOSE	/u:/	[ʊu ~ ʊu ~ u:]	/u:/	[u: ~ ɪu]
* <i>FLEW</i>	/u:/	[(j)ʊu ~ (j)u:]	/ɪu/	[ɪu ~ eu]

A.iii.b Welsh vowels

Table A.2: Lexical sets for Welsh vowels with phonemic and phonetic details.

Short Vowels		
PRIN	/ɪ/	[ɪ ~ i]
LLYN	/ɛ/	[ɛ ~ ɛ]
NEN	/ɛ/	[ɛ ~ e]
RHAN	/ɐ/	[ɐ ~ a]

LLON	/ɔ/	[ɔ ~ o]
CRWN	/ʊ/	[ə ~ ʊ ~ u]
YN	/ə/	[ə]
Long Vowels		
GWIN	/i:/	[i:]
CLUN	/i:/	[i: ~ i:]
GÊN	/e:/	[e:]
GLÂN	/a:/	[a: ~ æ: ~ ε:]
BÔN	/o:/	[o:]
CŴN	/u:/	[u:]
Diphthongs		
SAIN	/ei/	[ai ~ ei]
GWAUN	/ai/	[ai ~ ei ~ ɔi]
HAEN	/a:i/	[a:i ~ ei ~ a: ~ æ:]
TROI	/ɔi/	[ɔi ~ oi]
POEN	/o:i/	[ɔi ~ ɔi ~ o:i ~ o: ~ we:]
CWYN	/vi/	[vi ~ ʊi ~ u:i ~ wi:]
EIN	/əi/	[əi ~ ei]
CREU	/əi/	[əi ~ əi ~ ei]
LLIW	/iu/	[iu ~ iu]
DRYW	/iu/	[iu ~ iu]
RHEW	/eu/	[e:u ~ εu]
GLAW	/əu/	[əu ~ a:u]
BYWYD	/əu/	[əu ~ ʌu]

Appendix B

Additional postcards



Figure B.3

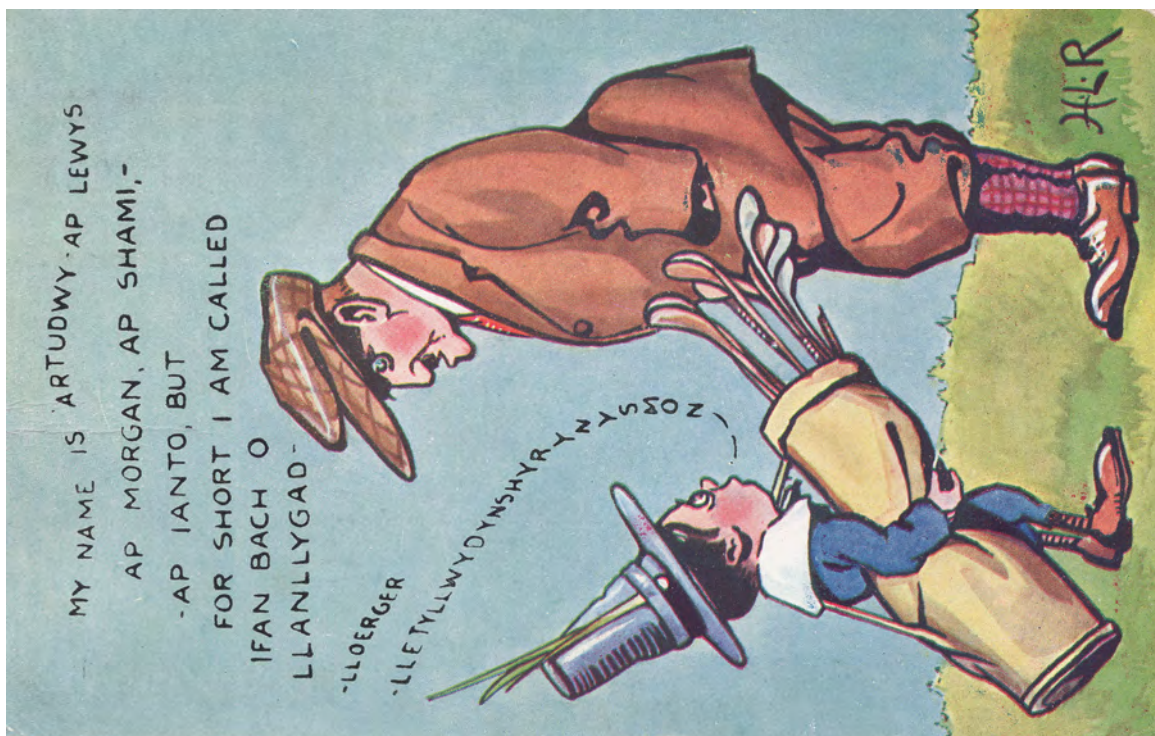


Figure B.4



Figure B.5

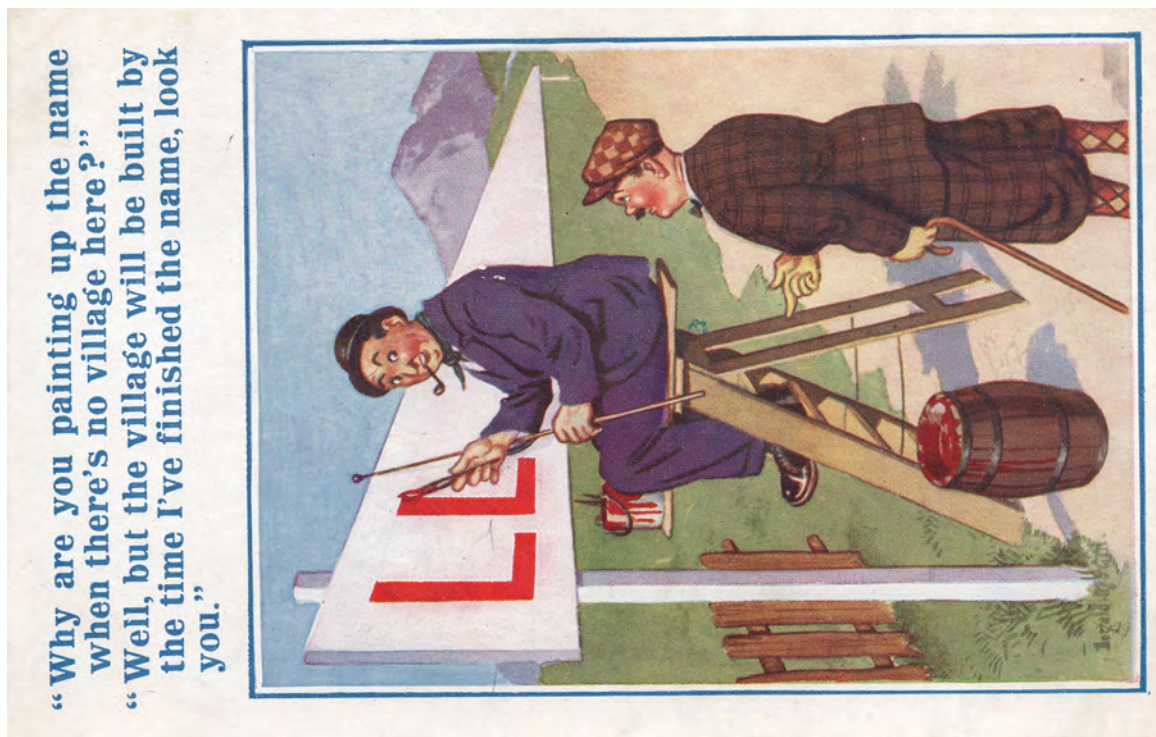


Figure B.6

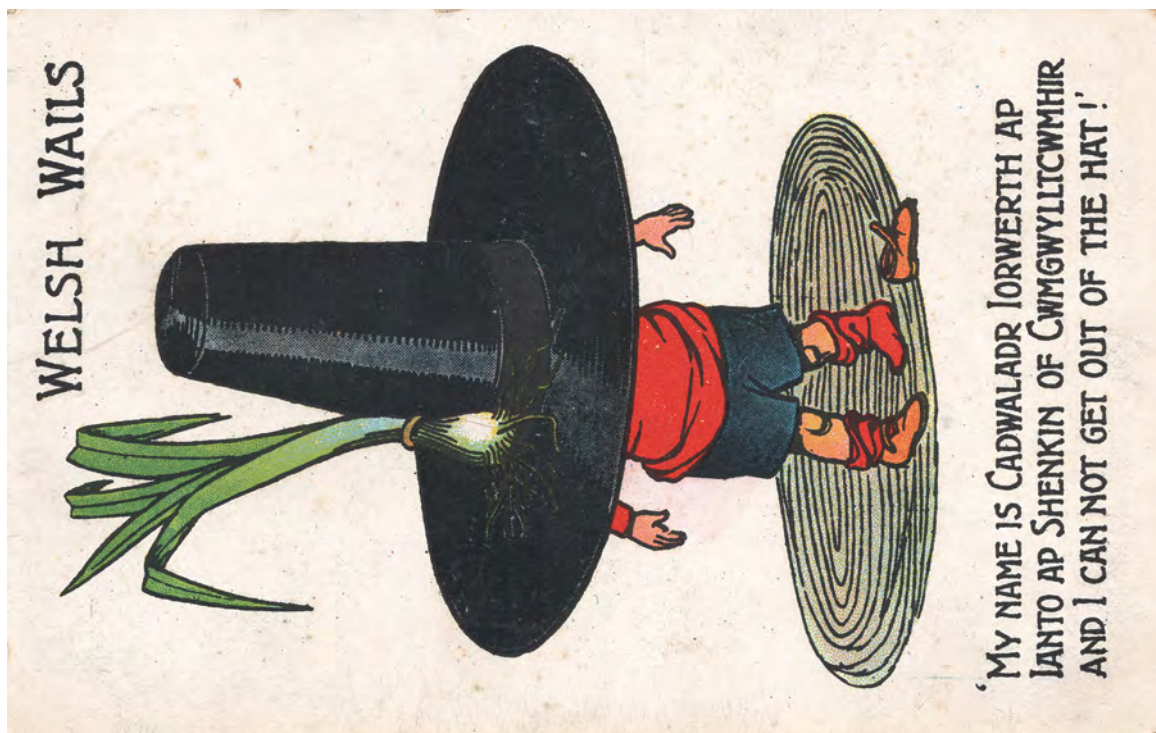


Figure B.7

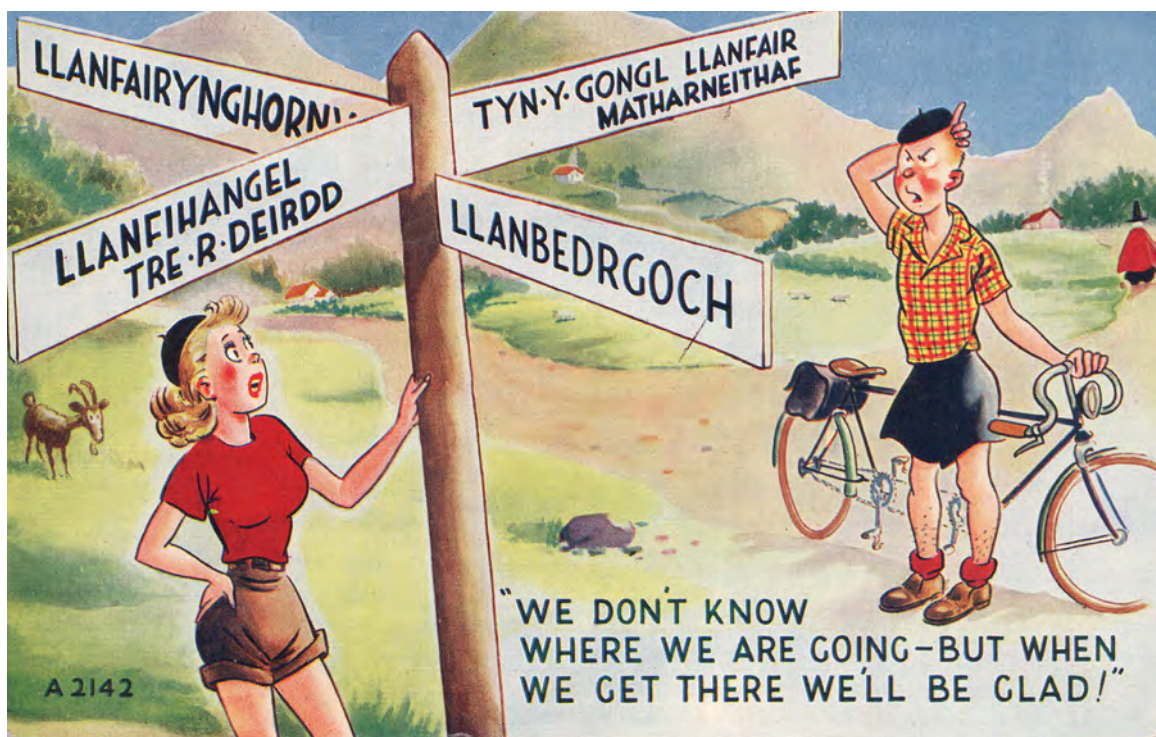


Figure B.8



Figure B.9



Figure B.10

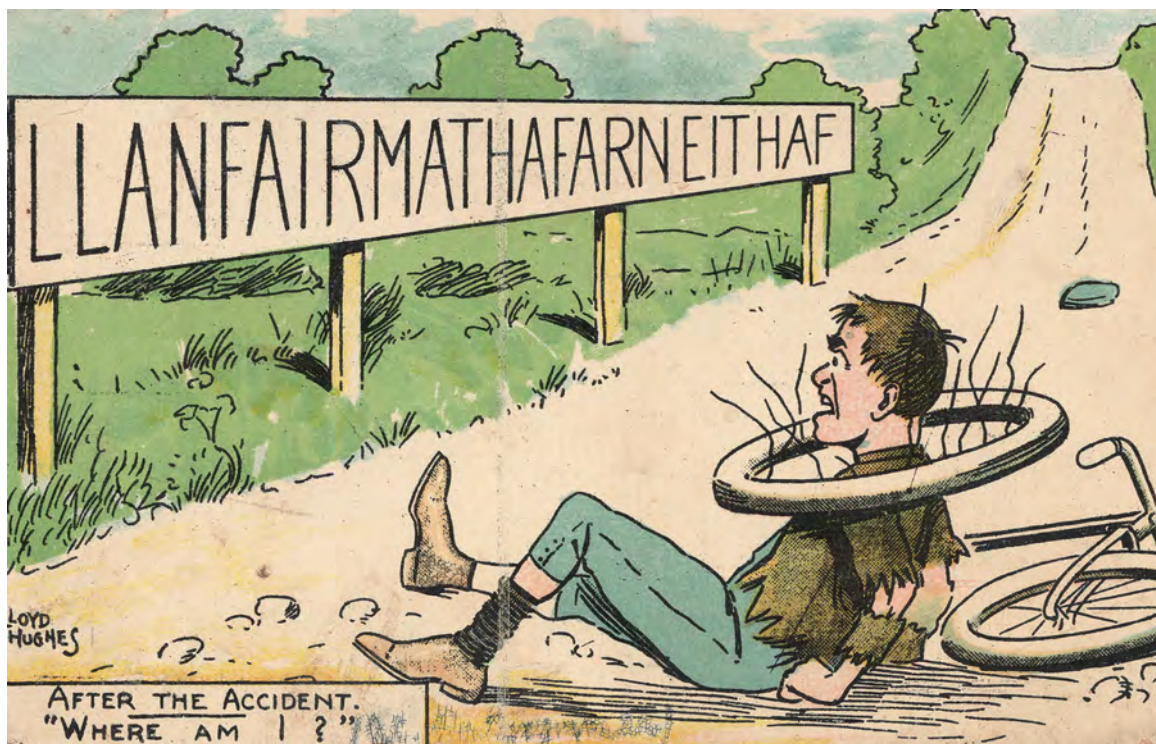


Figure B.11



Figure B.12

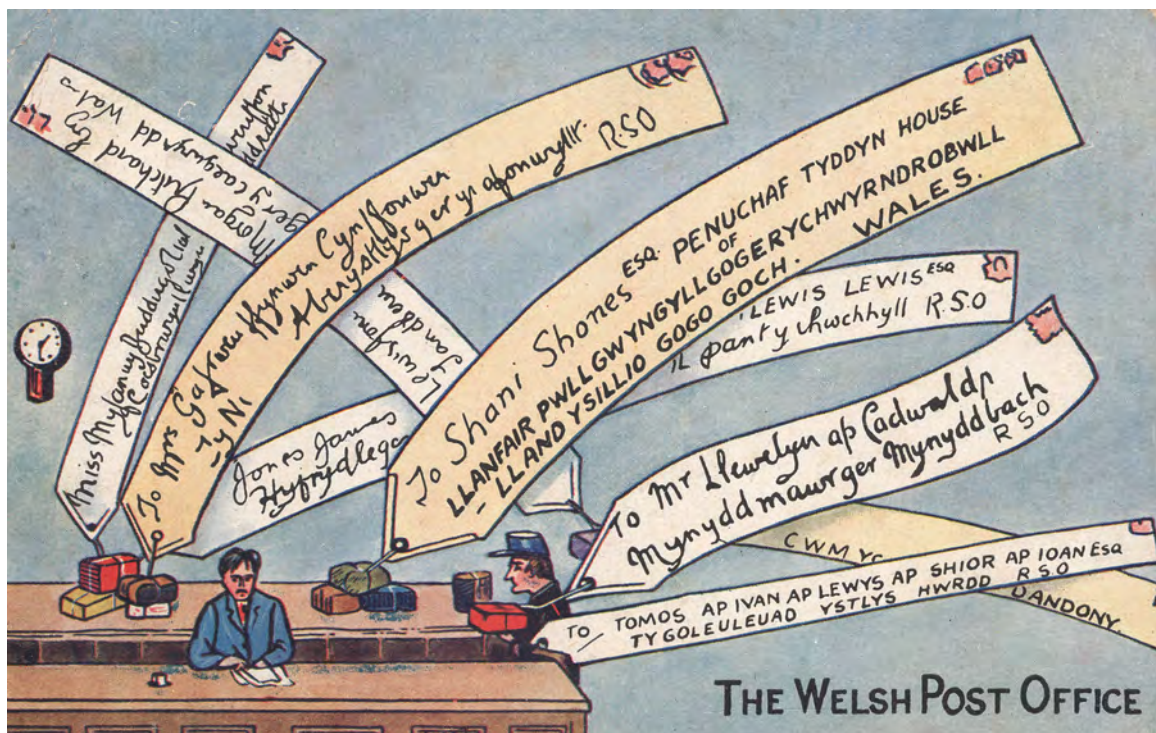


Figure B.13

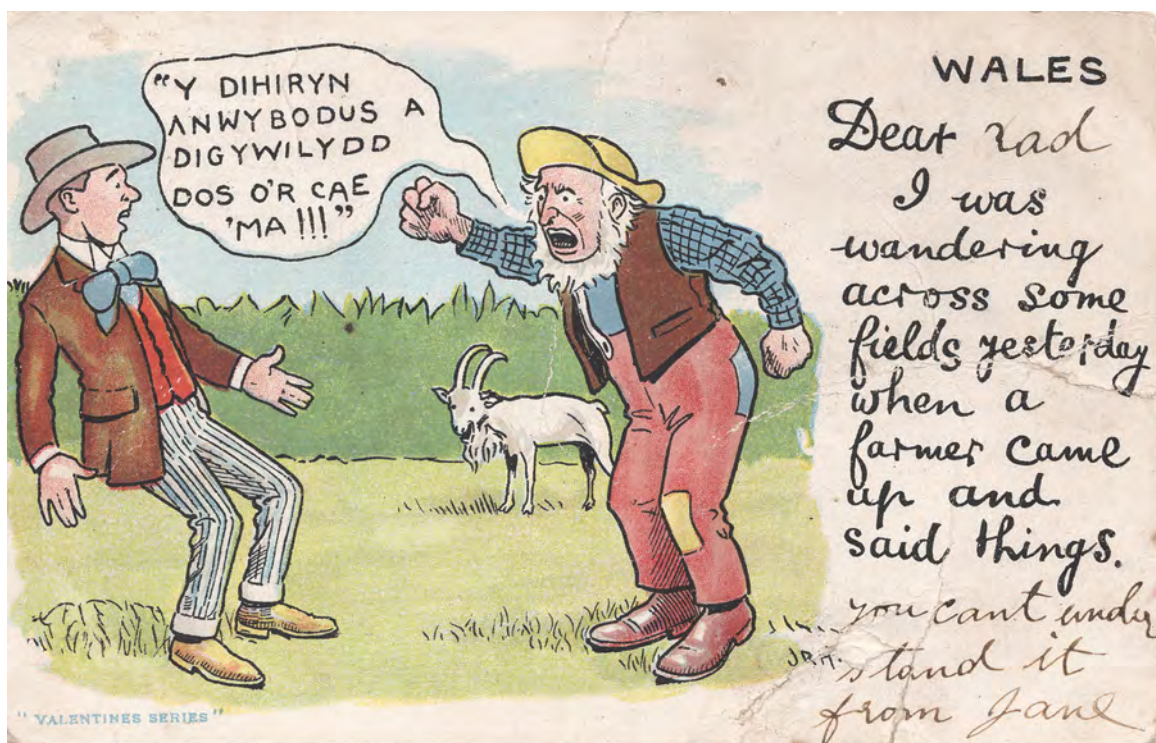


Figure B.14

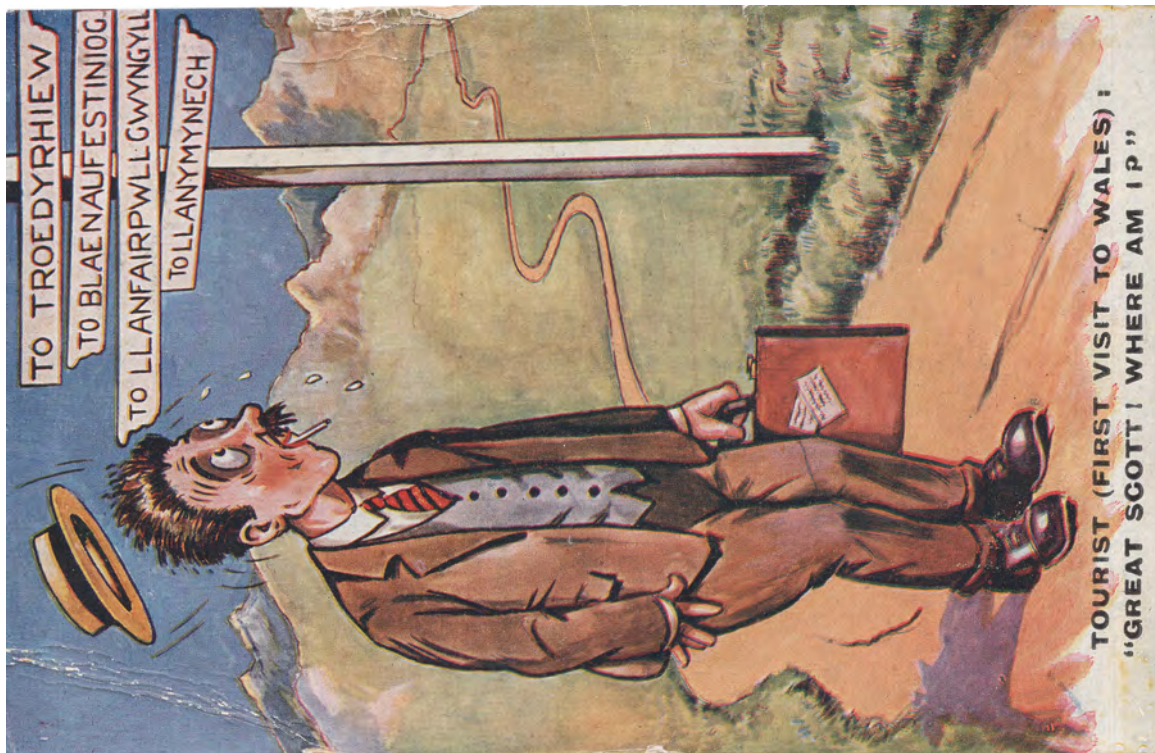


Figure B.15



Figure B.16



Figure B.17

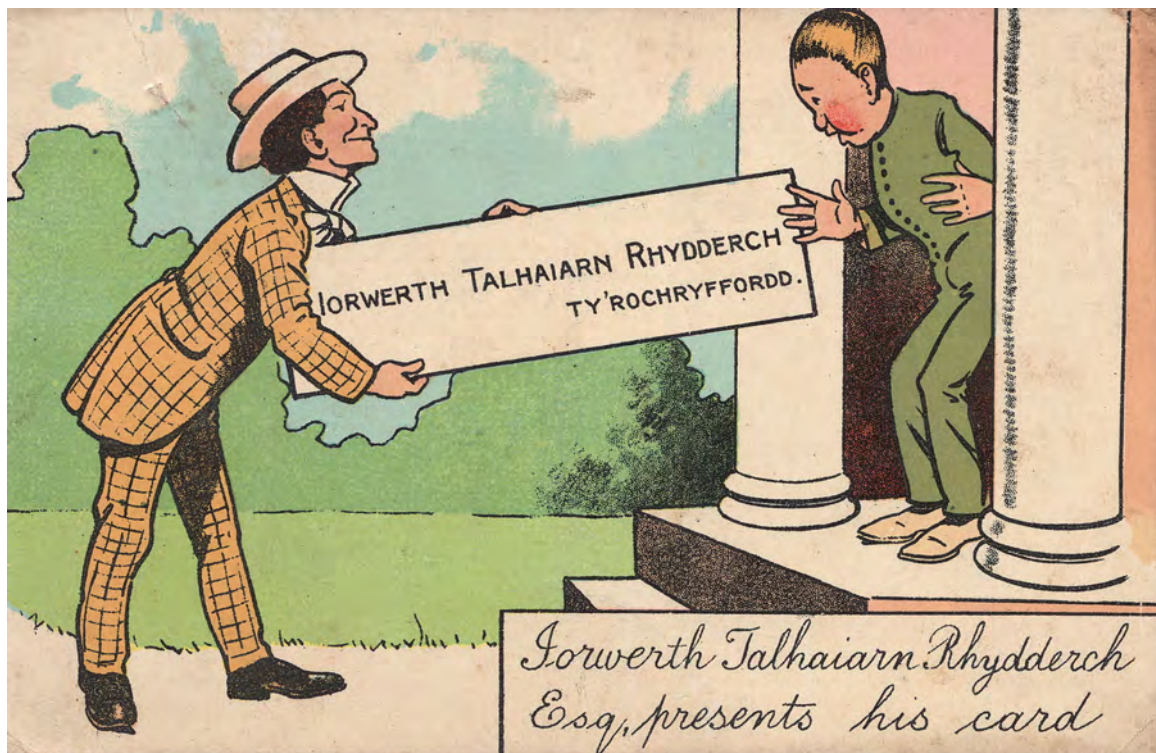


Figure B.18

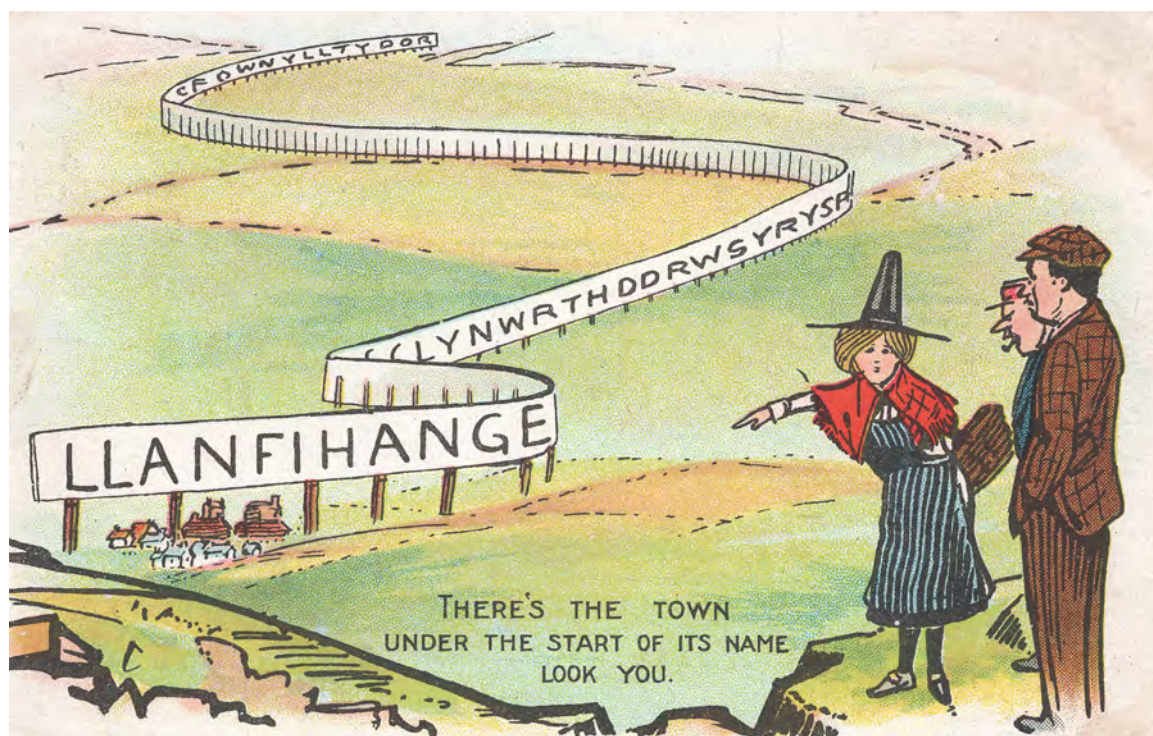


Figure B.19



Figure B.20



Figure B.21

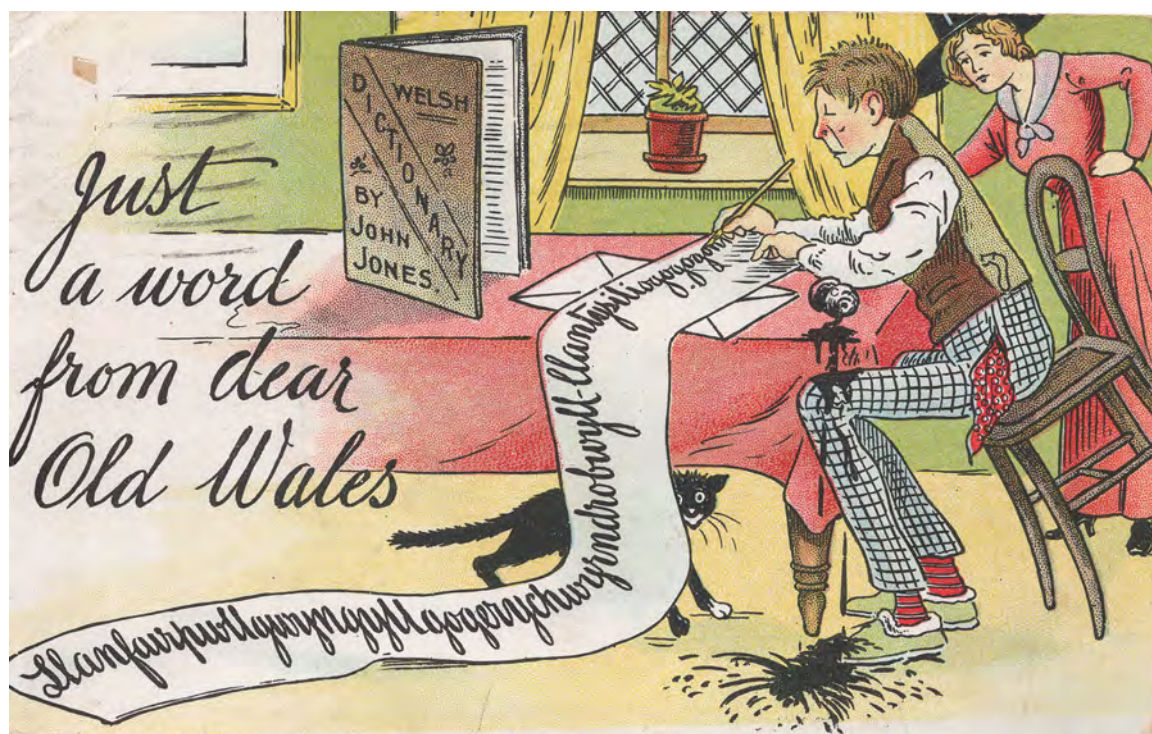


Figure B.22



Figure B.23



Figure B.24

Appendix C

Comparison of linguistic features in GPC, Survey, and Twitter datasets

C.i Distribution of orthographic variants for vowel and consonant phonemes

Notes on data representation and pre-processing

Phonemes in these graphs are ordered by ascending sum total of orthographic variants across all four datasets. Only phonemes which feature in two or more lexemes in both the Survey and Twitter datasets are included, since patterns of use for phonemes occurring in only a single lexeme are too likely to be affected by the specific features of that lexeme.

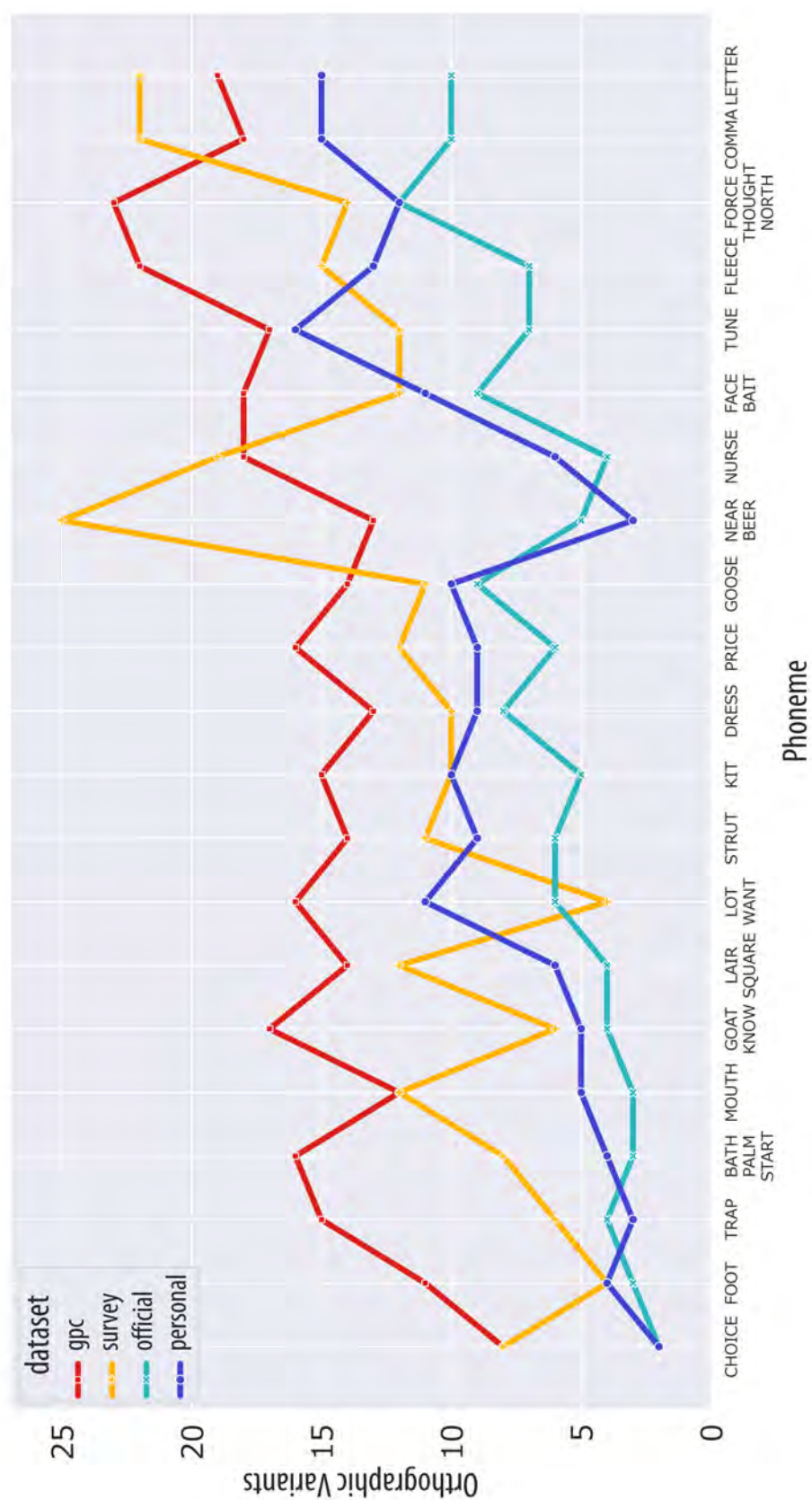


Figure C.1: Number of Welsh orthographic variants found for each vowel phoneme, separated by dataset.

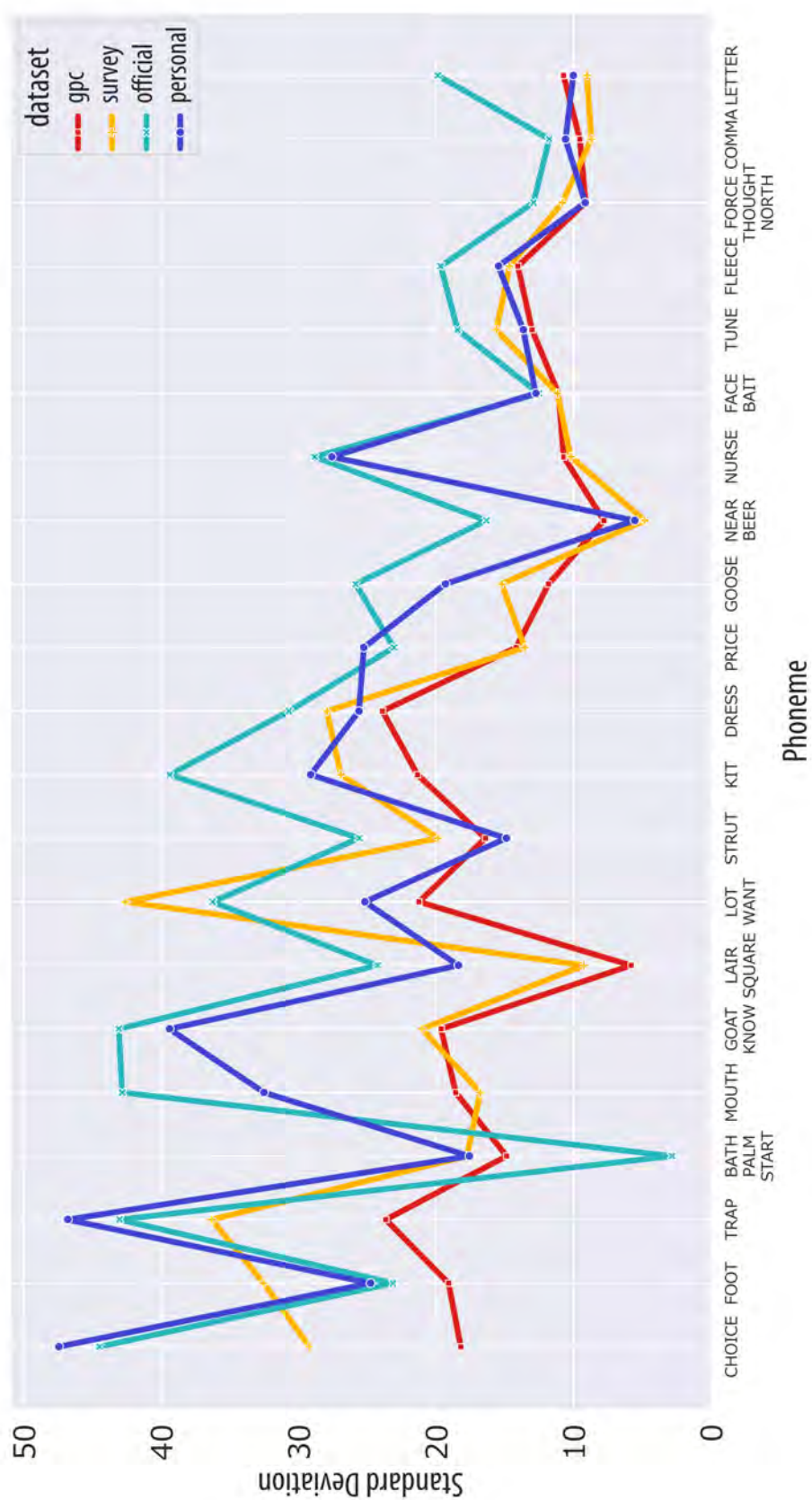


Figure C.2: Standard deviation of the distribution of each Welsh orthographic variant as a proportion of the total instances of a phoneme, separated by dataset.

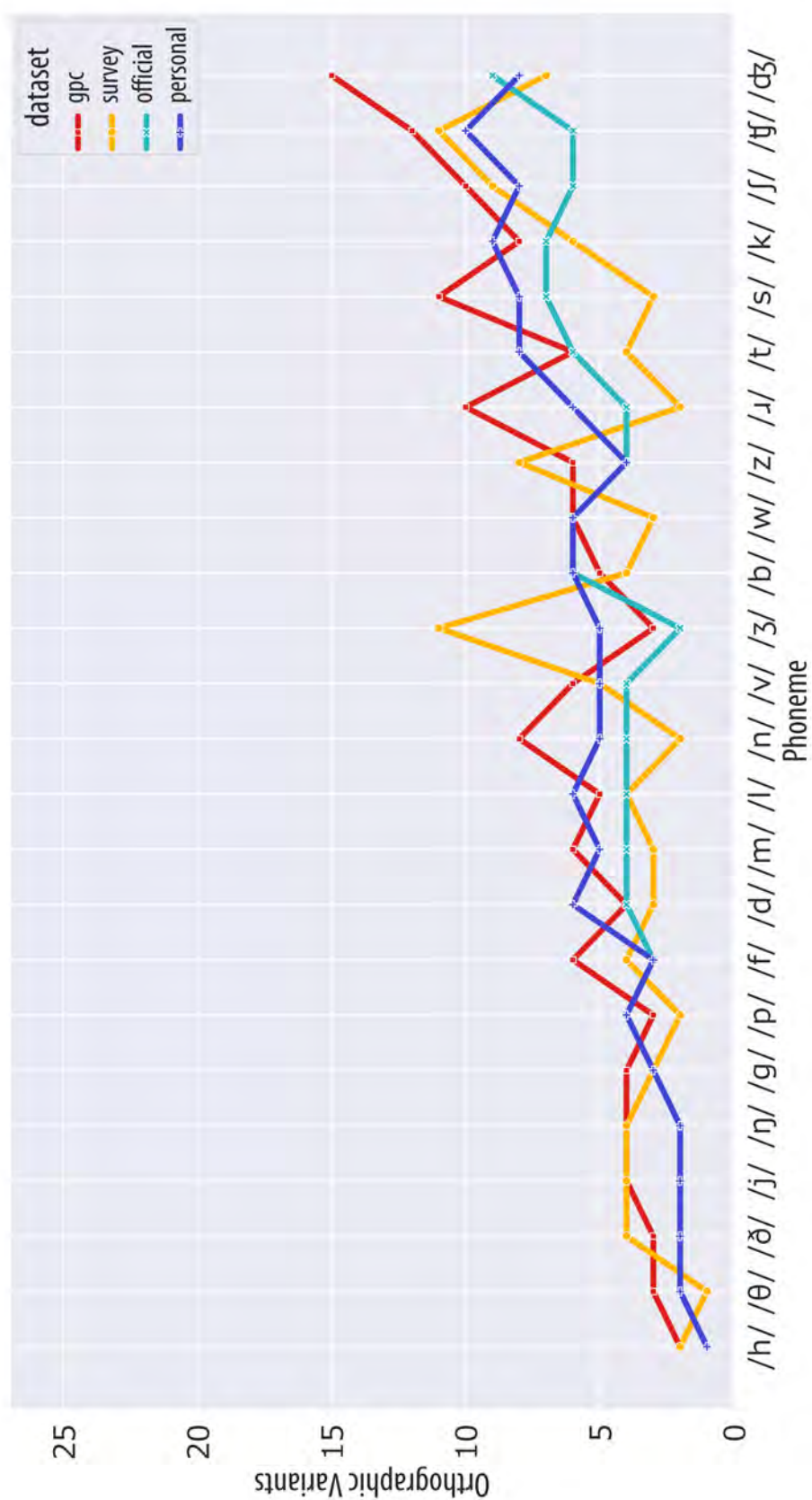


Figure C.3: Number of Welsh orthographic variants found for each consonant phoneme, separated by dataset.

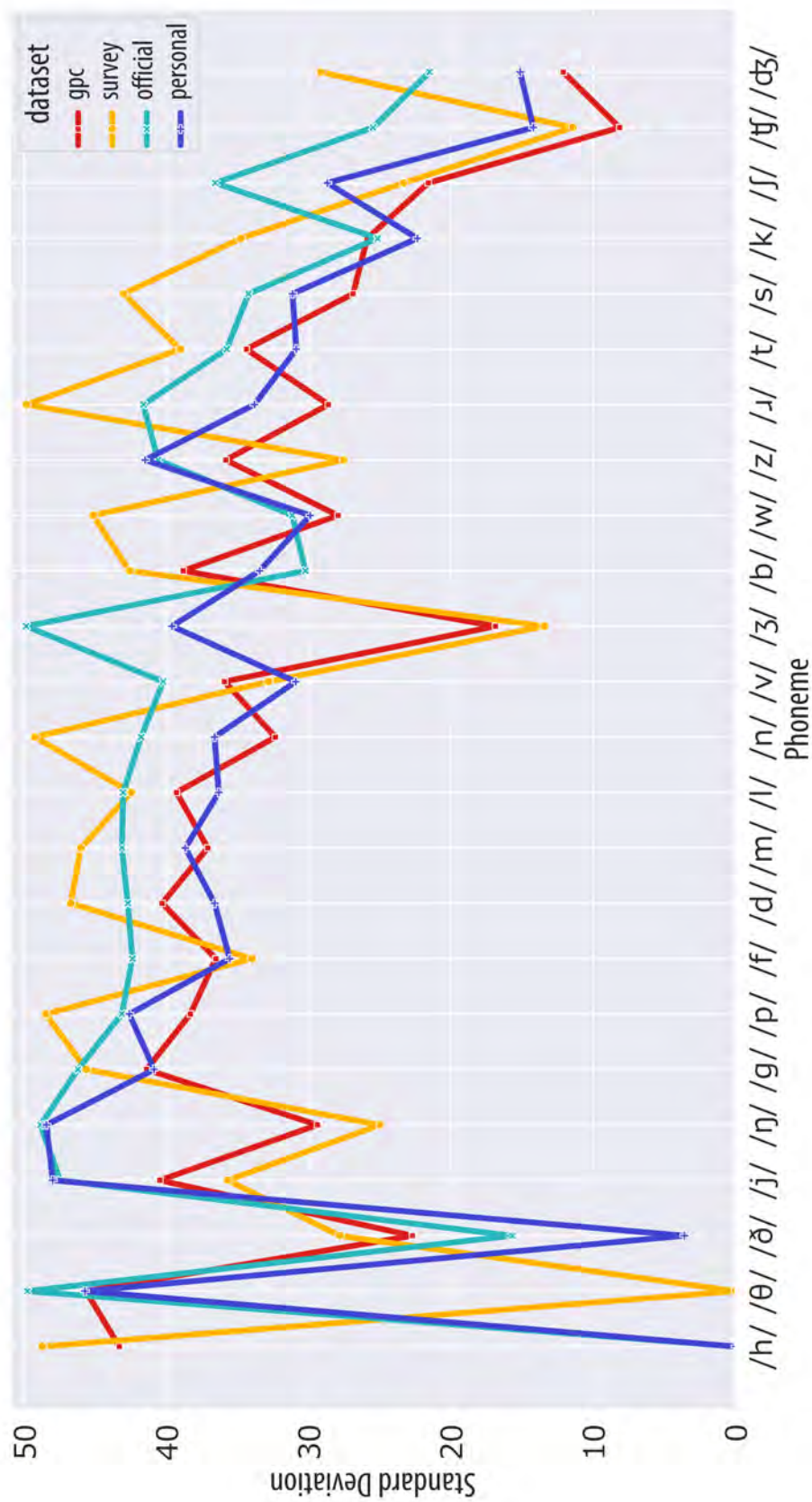


Figure C.4: Standard deviation of the distribution of each Welsh orthographic variant as a proportion of the total instances of a phoneme, separated by dataset.

C.ii Patterns of diacritic use in the representation of vowel sounds

Notes on data representation

The barcharts showing diacritic use separate vowel sounds into seven categories. Following Lindsey (2019, p. 24), I classify the **GOOSE** and **FLEECE** vowels as diphthongs; however, since they are particularly narrow diphthongs which are interpreted by others as monophthongs, I have separated them into their own class. The classes, and the phonemes belonging to them, are as follows:

- **unstressed:** /ə/, /ə/ (**LETTER, COMMA**)
- **short:** /ɪ/, /ɛ/, /æ/, /ɔ/, /ʌ/, /ʊ/ (**KIT, DRESS, TRAP, LOT~CLOTH, STRUT, FOOT**)
- **long:** /ɜ:/, /ɑ:/, /o:/ (**NURSE, PALM~BATH~START, THOUGHT~NORTH~FORCE**)
- **centering diphthong:** /ɪə/, /ɛə/, /aə/, /ʊə/, /jʊə/ (**NEAR~BEER, SQUARE~LAIR, FIRE, TOWER, POOR, CURE**)
- **closing diphthong:** /eɪ/, /aɪ/, /oɪ/, /aʊ/, /əʊ/, /jʊə/ (**FACE~BAIT, PRICE~AYE, CHOICE, MOUTH, GOAT~KNOW, TUNE**)
- **narrow diphthong:** /i:/, /u:/ (**FLEECE, GOOSE~FLEW**)
- **glide:** /j/, /w/

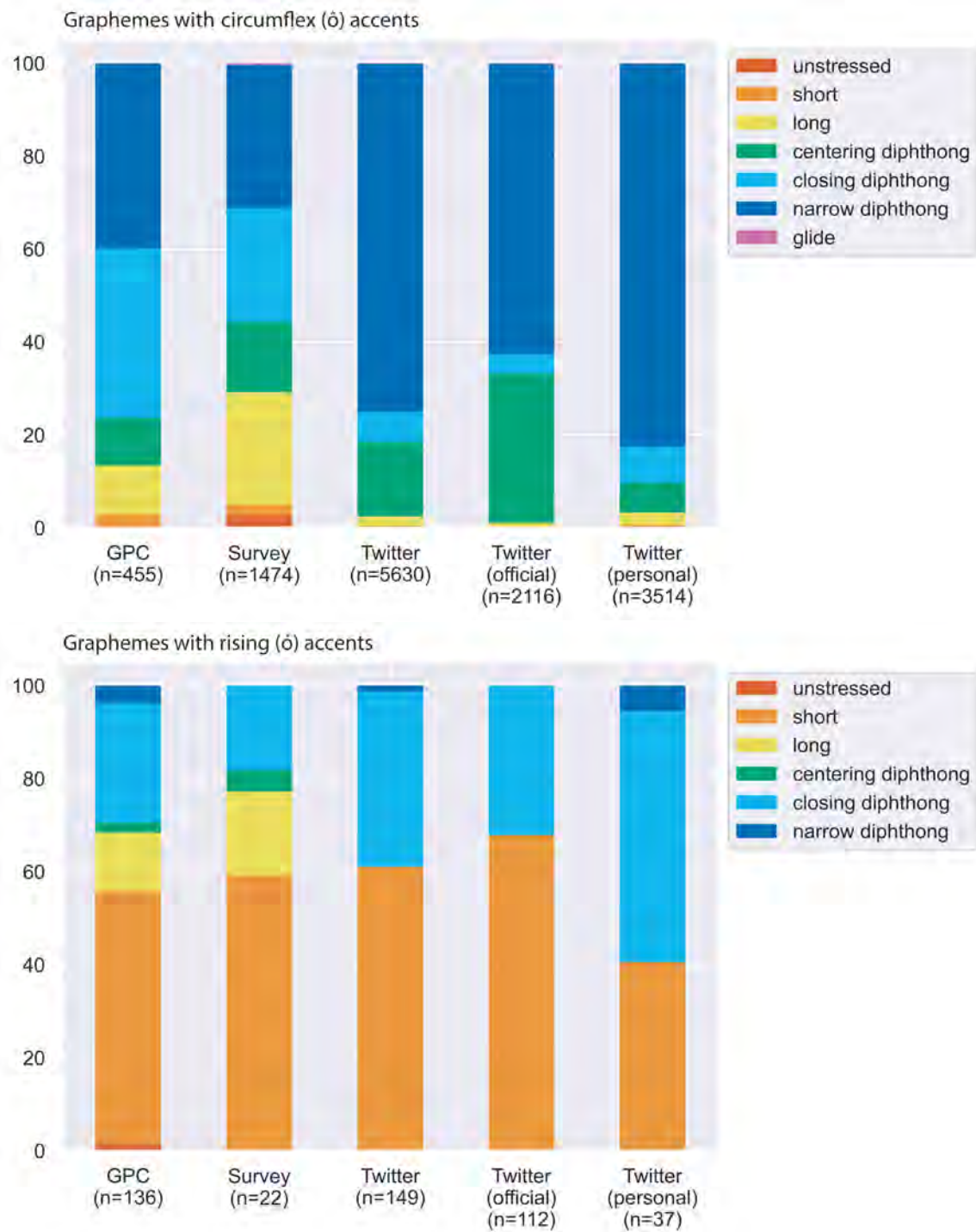


Figure C.5: Stacked barcharts showing the proportion of accented characters representing different vowel types in this study's datasets.

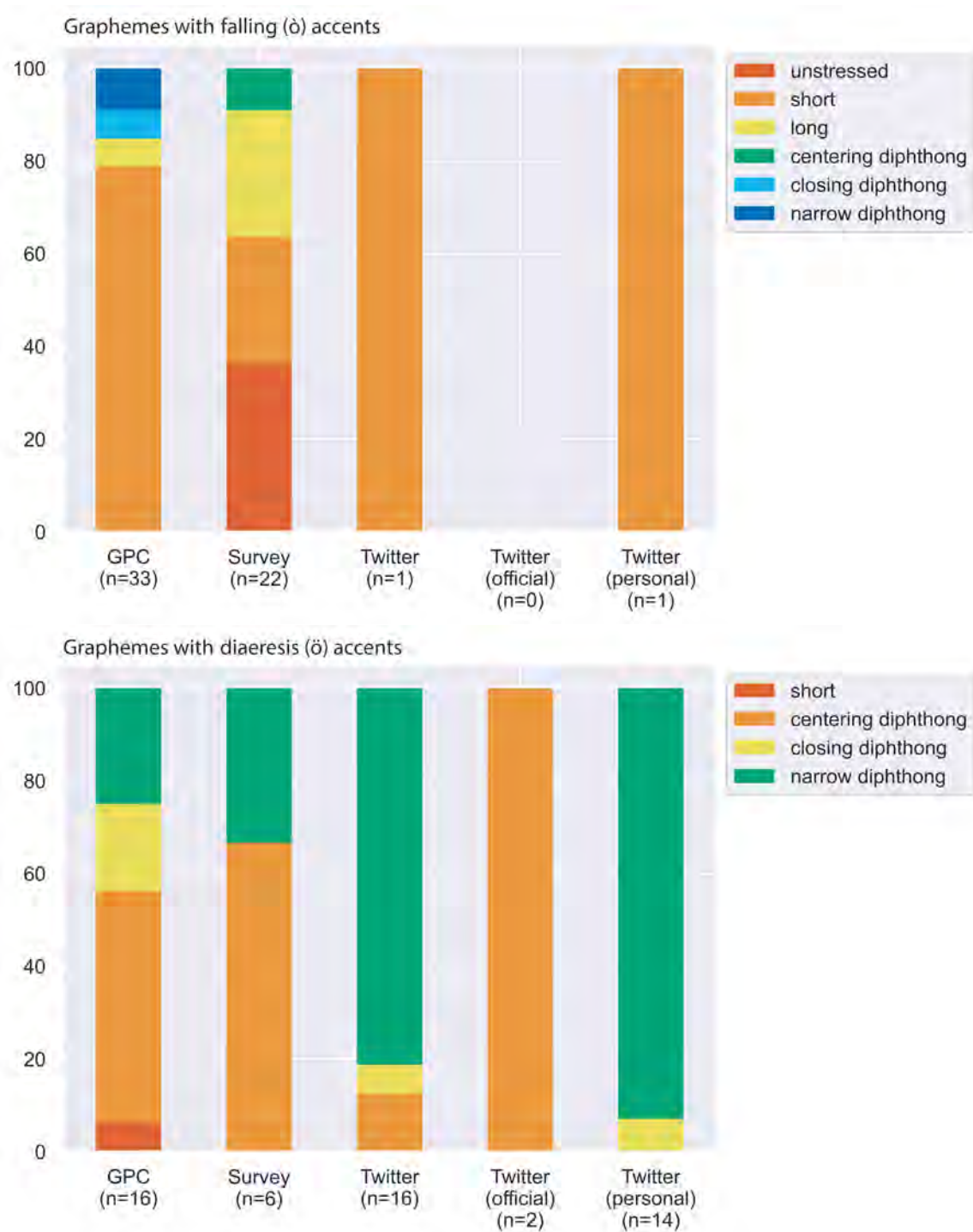


Figure C.5: Stacked barcharts showing the proportion of accented characters (cont.).

C.iii Phoneme-grapheme distributions

Notes on data representation and pre-processing

In the figures in this appendix, the colours assigned to graphemes are kept consistent within charts representing the same phoneme. This means, for example, that the grapheme <g> will always be shown in bright green in every chart representing the phoneme /k/. However, the same grapheme may be shown in a different colour in charts for other phonemes: in charts representing the phoneme /g/, for example, the grapheme <g> is shown in orange.

The distribution of graphemes within a given phoneme across the different datasets can be broadly compared by looking at all the charts for the same phoneme. Nonetheless, graphemes which make up less than one percent of the total instances of the phoneme are generally not clearly visible. The key printed to the right of each chart therefore lists every grapheme alongside its percentage, giving precise figures to complement the rough sketch of grapheme distributions provided by the pie charts.

In the survey data, respondents were specifically requested to adapt all lexemes unless doing so was unnecessary or impossible. The variant forms produced by the respondents can therefore all be interpreted as the result of a deliberate attempt to apply Welsh orthographic norms to an English-origin ELU. When the variants are unadapted, this is most likely not because the respondents intended to produce the English word (i.e. to switch languages to English for the given lexeme), but because they either believe that the unadapted variant conforms to Welsh norms, or because they have had difficulty finding a Welsh orthographic alternative, or because they have failed to notice that some aspect of the Standard English form does not conform to Welsh orthography.

In the Twitter data, in contrast, it is likely that many unadapted lexemes represent code-switches into English, either single-word code-switches, or parts of longer English-language stretches of discourse, such as quotations, sayings or proverbs, company or brand names, titles of artistic works, and so on.

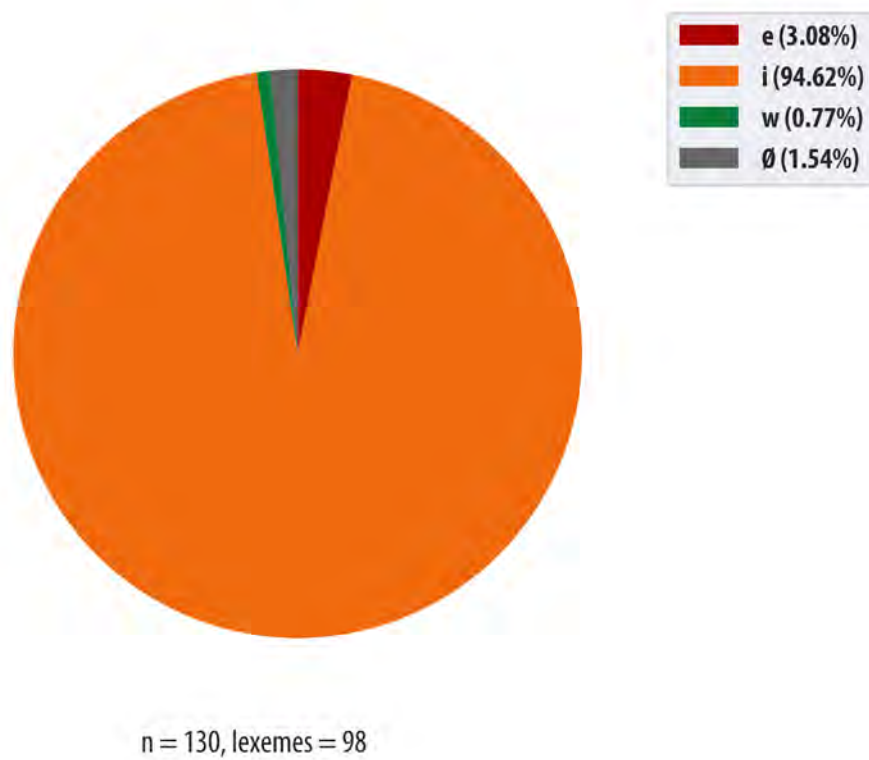
Since the aim of these charts is to show the distribution of graphemes matched to phonemes in variant forms with some level of adaptation to the Welsh system of orthography, the Twitter dataset is preprocessed to remove many unadapted variant forms. If a variant form is identical to the Standard English form, it has been excluded from the data used to compile this appendix, unless it is listed in at least one of the dictionaries consulted for this study. Variants such as <football>, <music>, and <conservative>, which appear in no dictionaries, are excluded. Examining the data

shows that these unadapted forms are often used in one of the contexts described above, for example in the (English-language) names of sporting clubs and tournaments, cultural or artistic events, and political organizations.

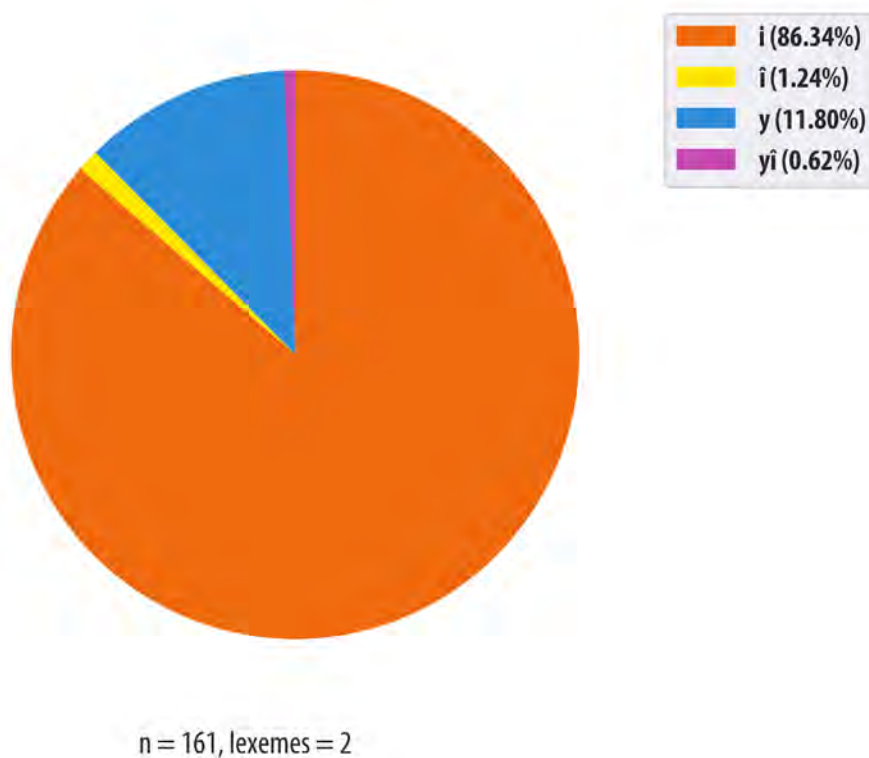
Variants such as <help> and <agenda>, however, are counted for the purposes of these charts, since they are listed in at least one Welsh dictionary. This provides some evidence that they are established loanwords and that their orthography remains unadapted either because it conforms to Welsh norms, or because the usual Welsh pronunciation of the lexeme has been adapted to conform to the English spelling. The spelling of <help>, for example, is unadapted because a Welsh pronunciation of this sequence of graphemes, /hɛlp/, matches the standard English pronunciation. The spelling of <agenda>, on the other hand, reflects a Welsh pronunciation /ɛ'gɛndɛ/, which matches the orthography (reinterpreted according to Welsh grapheme-to-phoneme rules) rather than the phonology of the Standard English form.

It cannot be assumed that all the remaining variants from the Twitter dataset are wholly or partially adapted to Welsh orthography, since forms which represent non-standard English spellings will not have been excluded. For example, this means that the data include non-standard spellings or typographic errors, such as the variant <desparate> for the lexeme desperate, which are common amongst non-Cymrophones and which show no particular adaptation to Welsh orthography. On balance, however, it seemed more reasonable to include any and all variants with differences from standard English orthography, rather than attempting to formulate rules to identify which forms are truly an attempt to adapt a variant and which are errors or non-standard forms that might equally have come from a non-Cymrophone.

In total, 18,245 unadapted variant forms, or 5.7% of the total variants, were removed from the Twitter corpus before producing the graphs in this appendix.

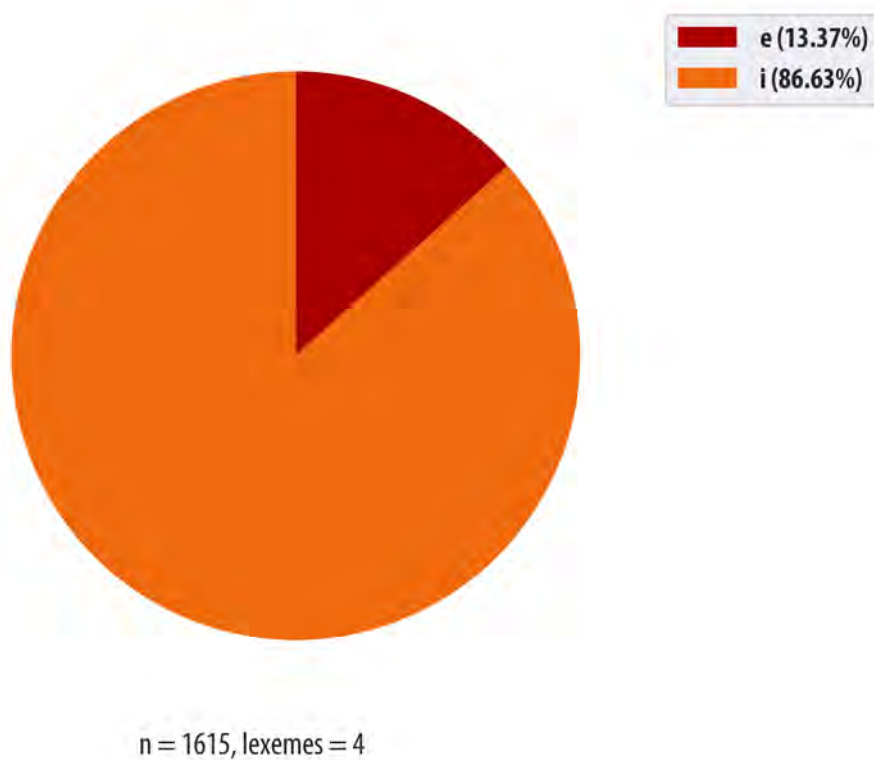


(a) GPC data

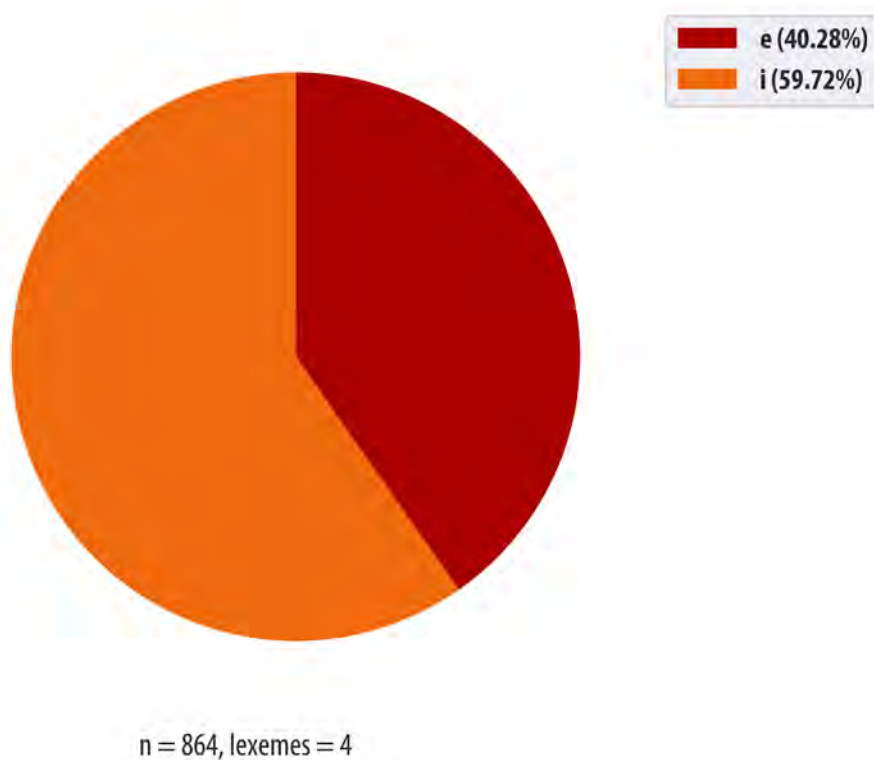


(b) Survey data

Figure C.6: Distribution of graphemes for phoneme /j/

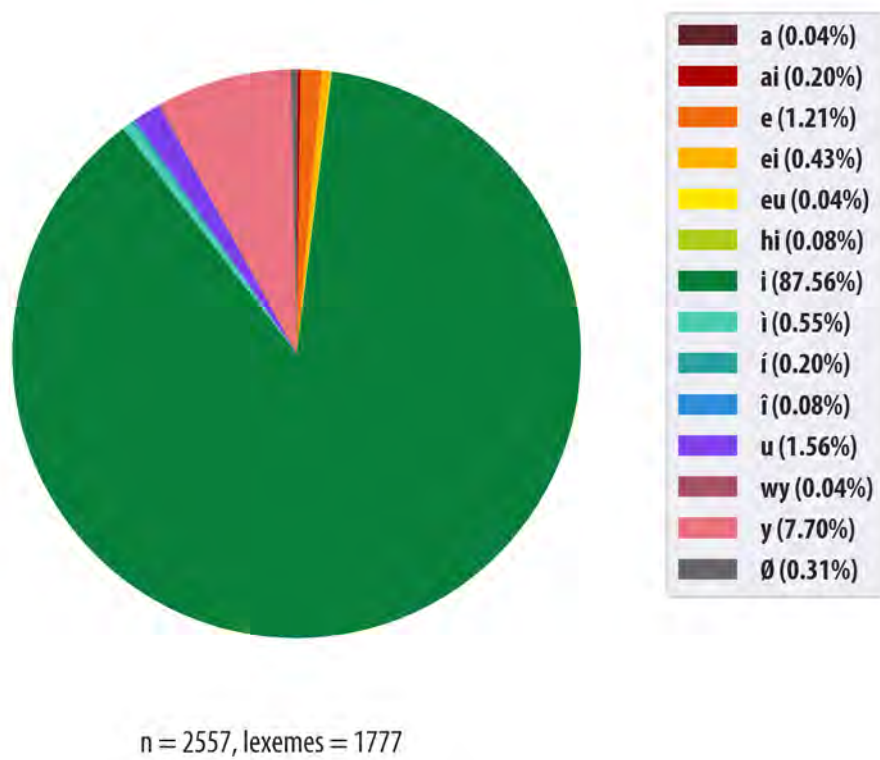


(c) Twitter (personal) data

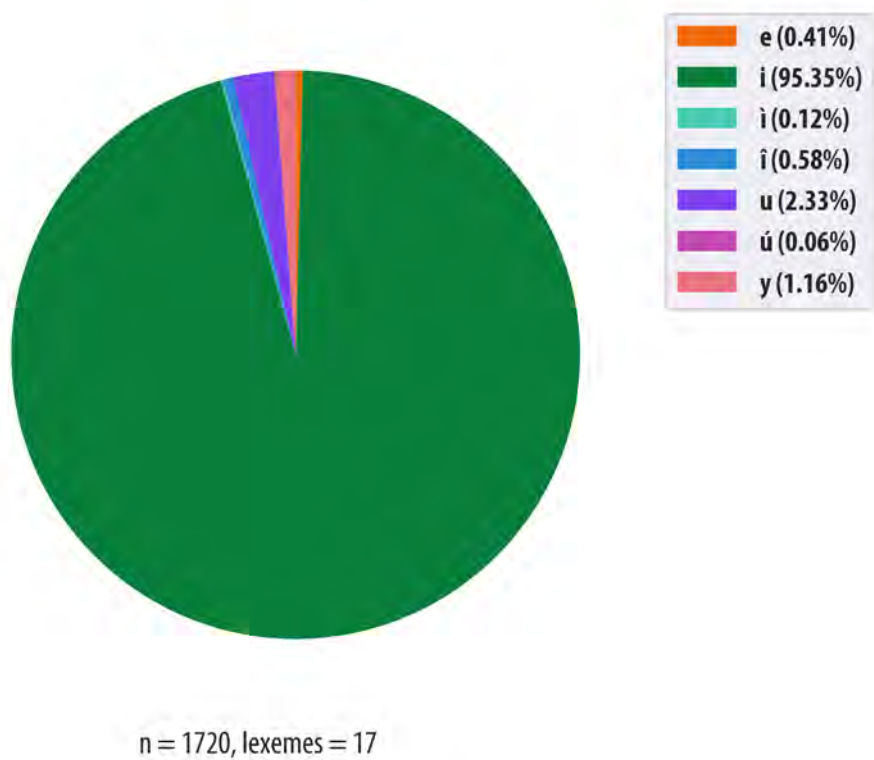


(d) Twitter (official) data

Figure C.6: Distribution of graphemes for phoneme /j/ (cont.)

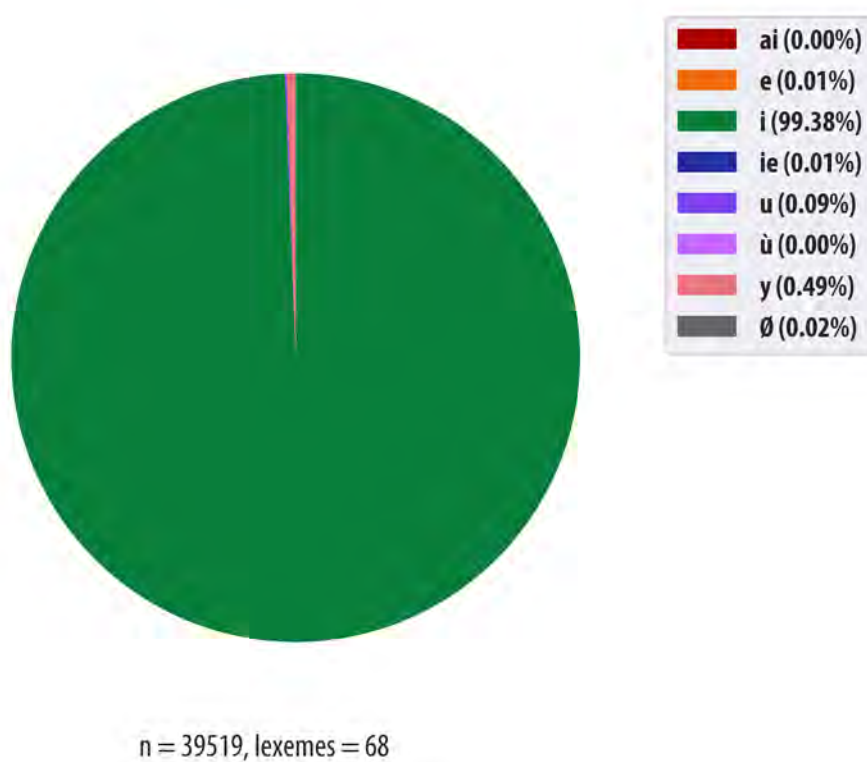


(a) GPC data

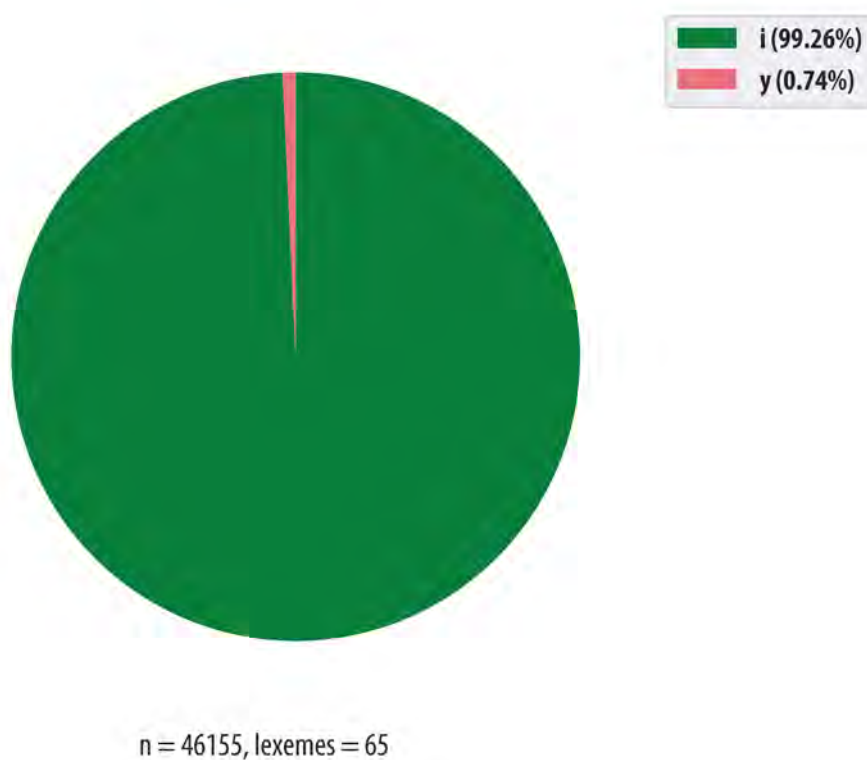


(b) Survey data

Figure C.7: Distribution of graphemes for phoneme /ɪ/ (**KIT**)

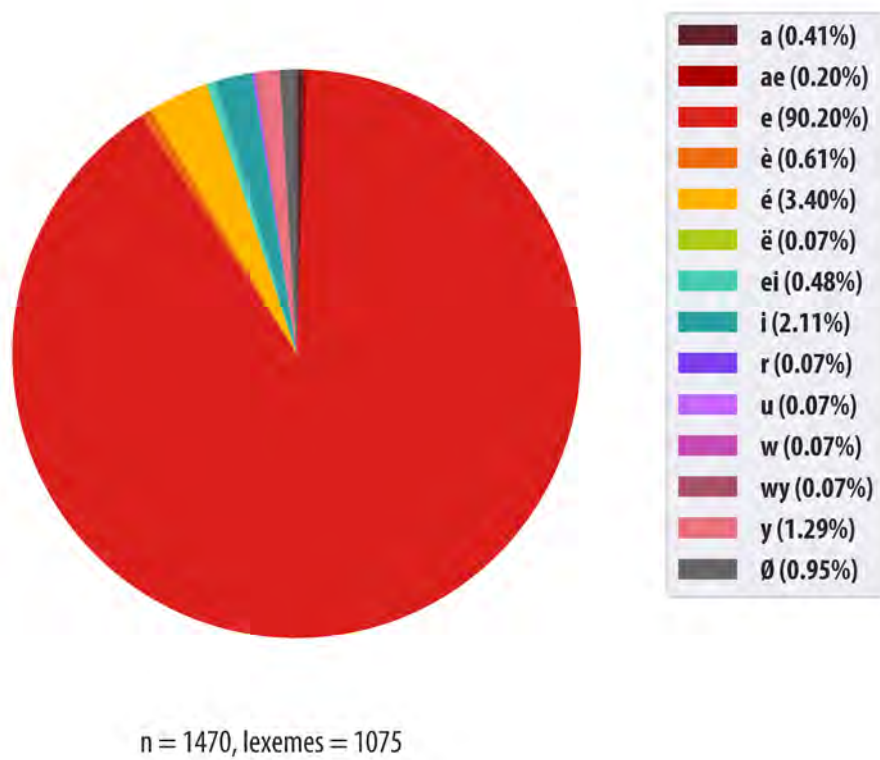


(c) Twitter (personal) data

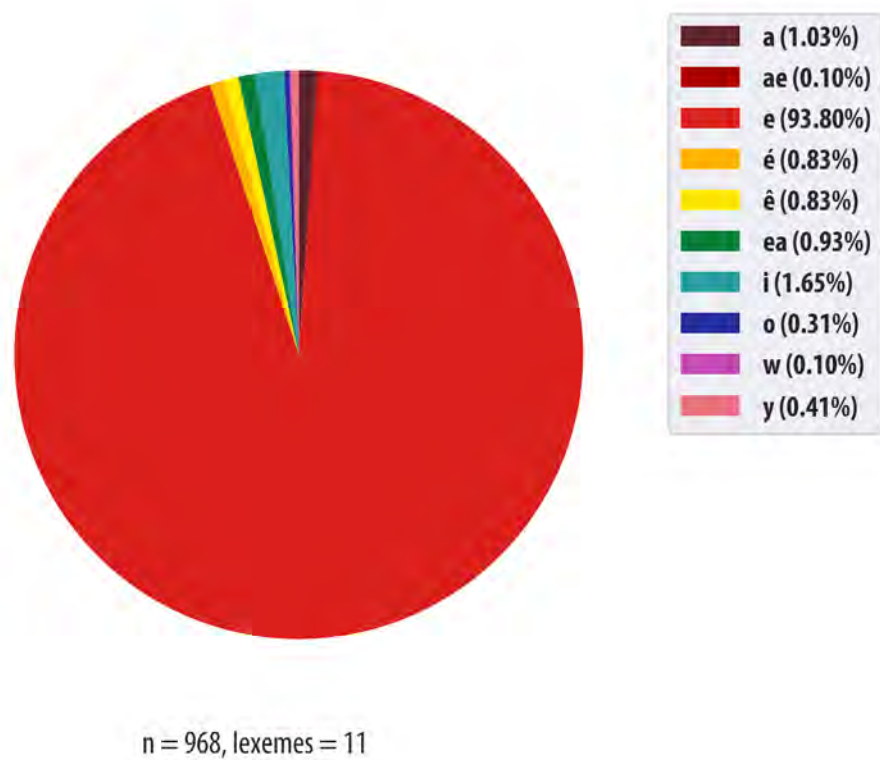


(d) Twitter (official) data

Figure C.7: Distribution of graphemes for phoneme /i/ (**KIT**) (cont.)

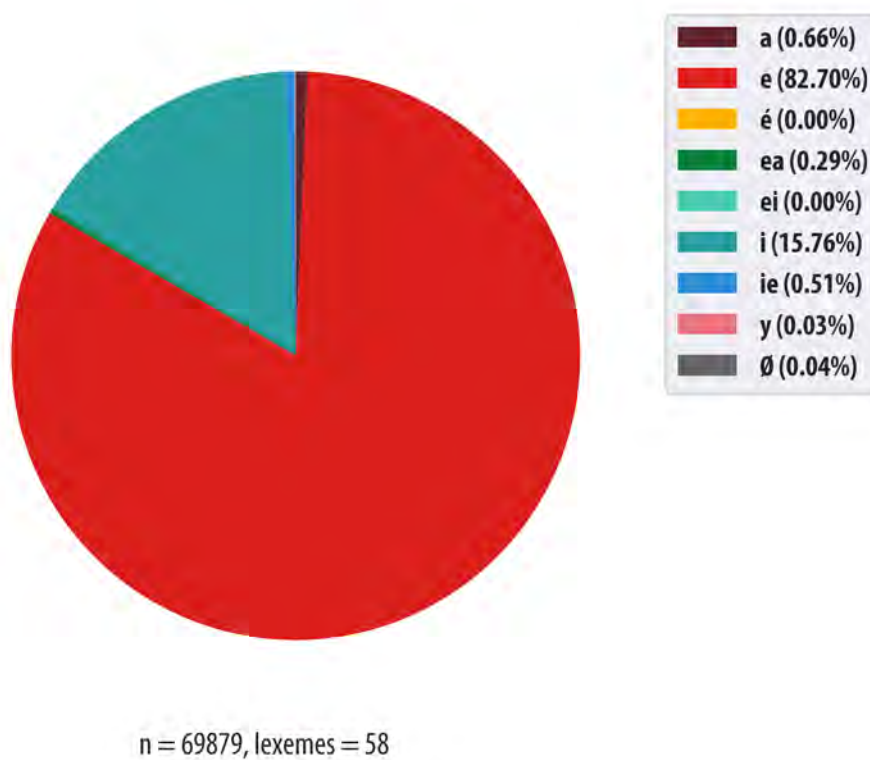


(a) GPC data

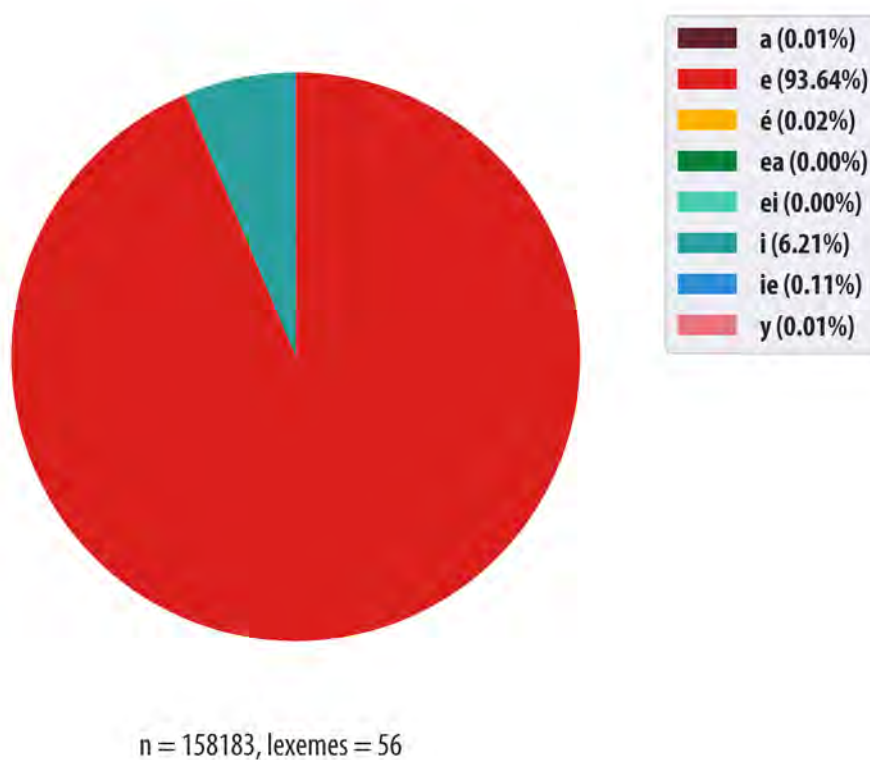


(b) Survey data

Figure C.8: Distribution of graphemes for phoneme /ɛ/ (**DRESS**)

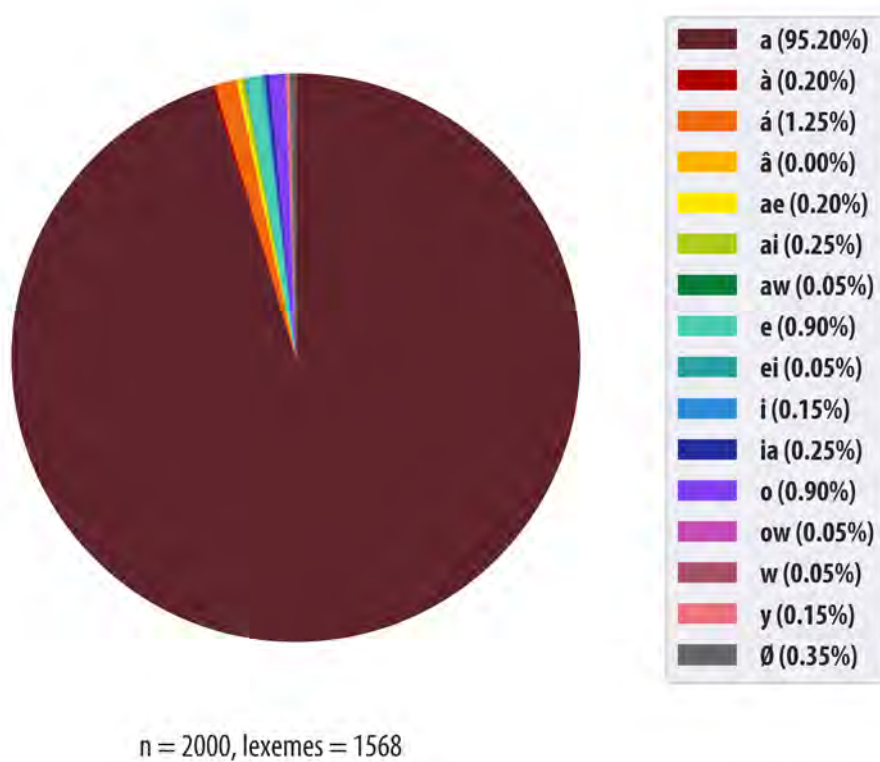


(c) Twitter (personal) data

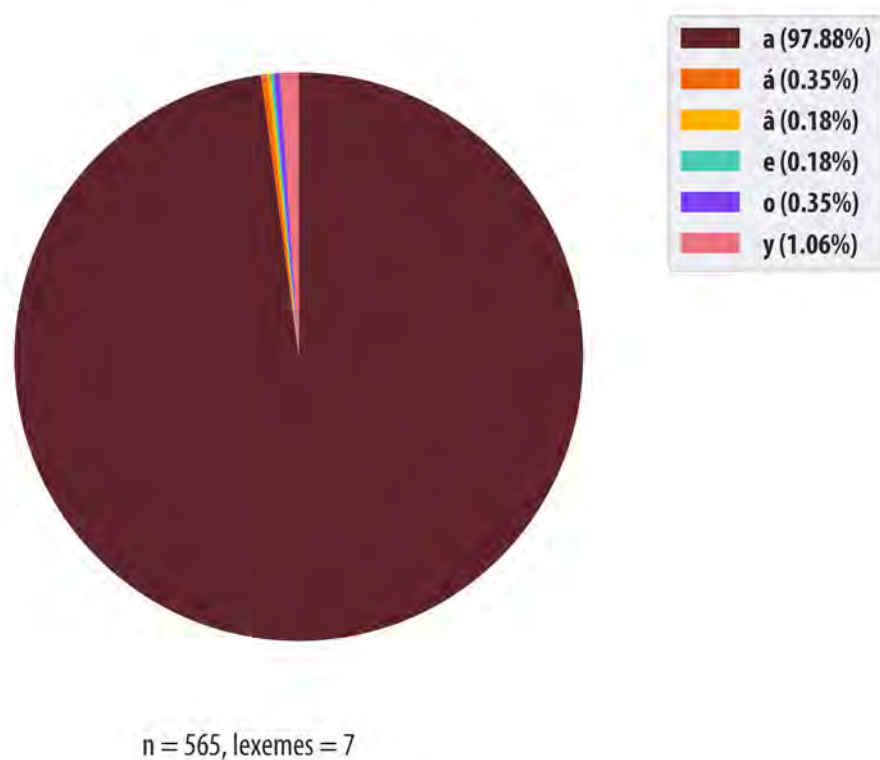


(d) Twitter (official) data

Figure C.8: Distribution of graphemes for phoneme /ɛ/ (**DRESS**) (cont.)

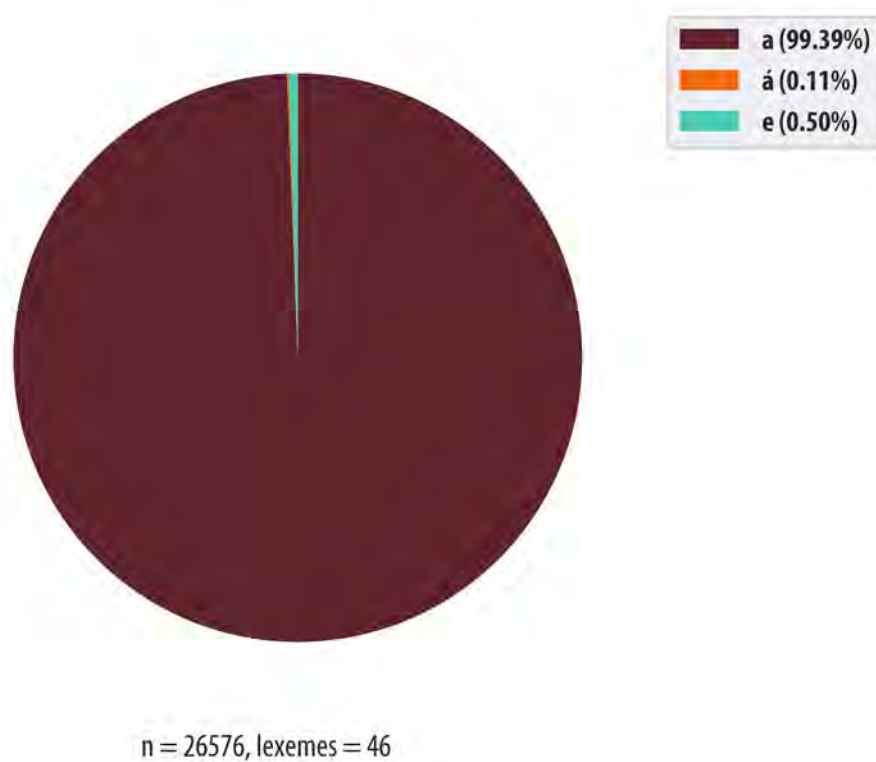


(a) GPC data

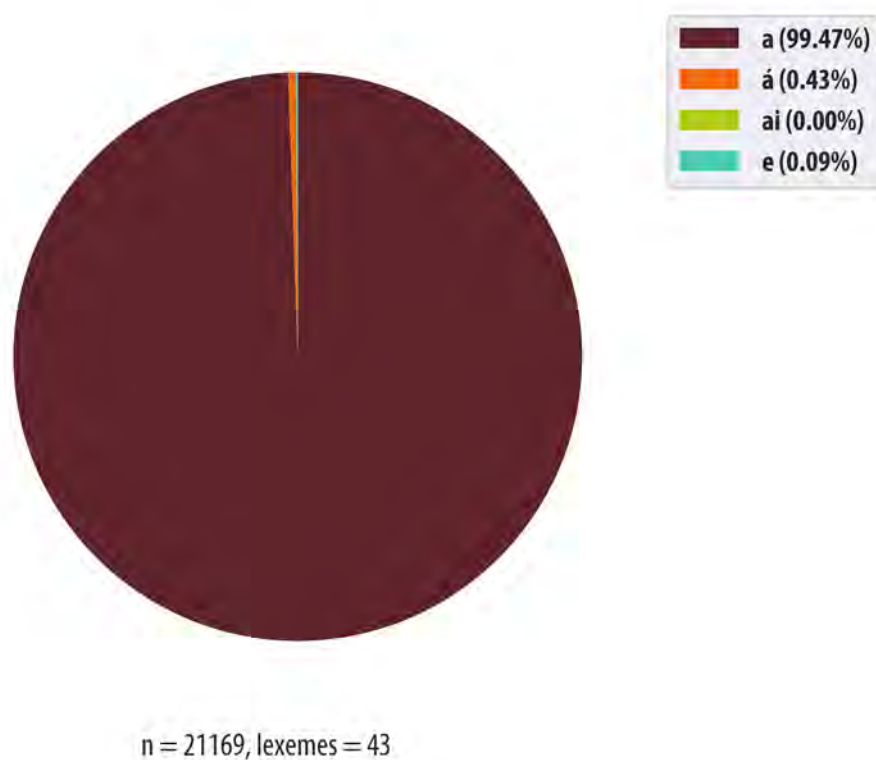


(b) Survey data

Figure C.9: Distribution of graphemes for phoneme /æ/ (**TRAP**)

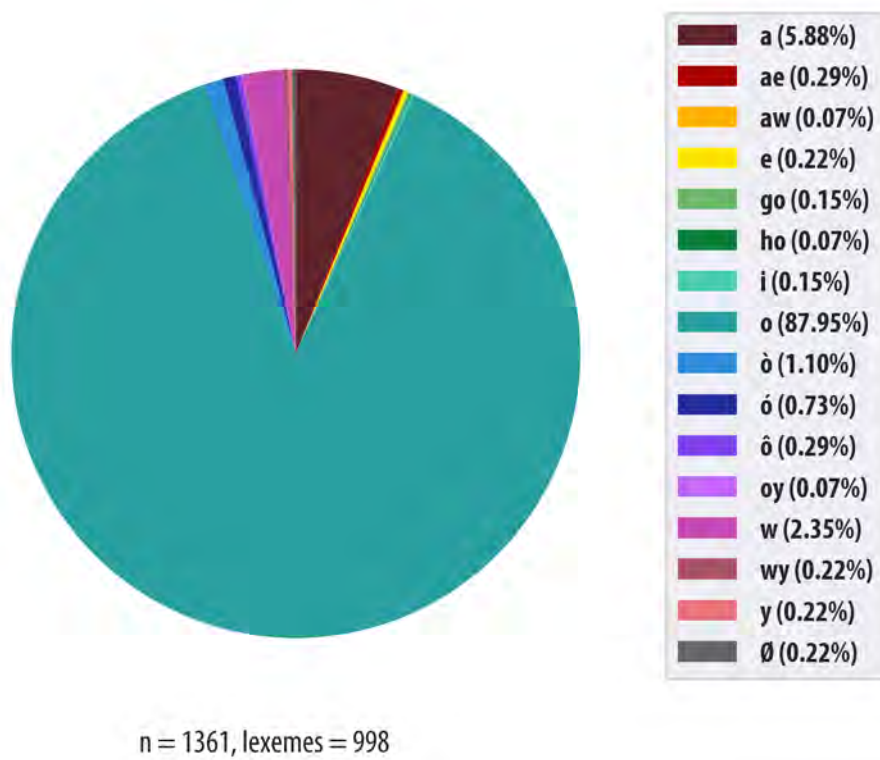


(c) Twitter (personal) data

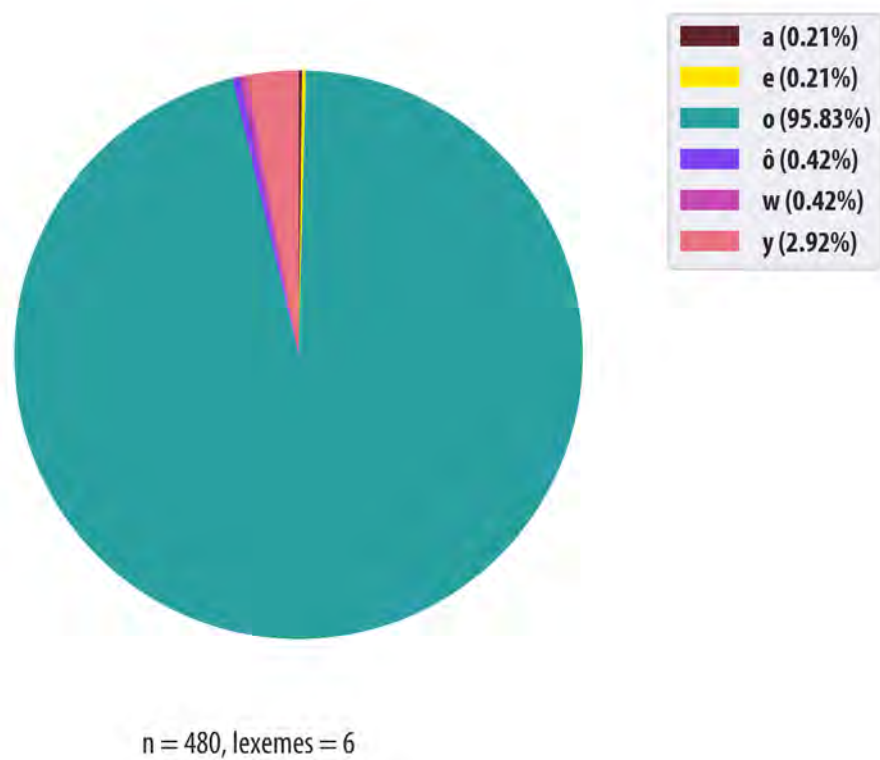


(d) Twitter (official) data

Figure C.9: Distribution of graphemes for phoneme /æ/ (**TRAP**) (cont.)

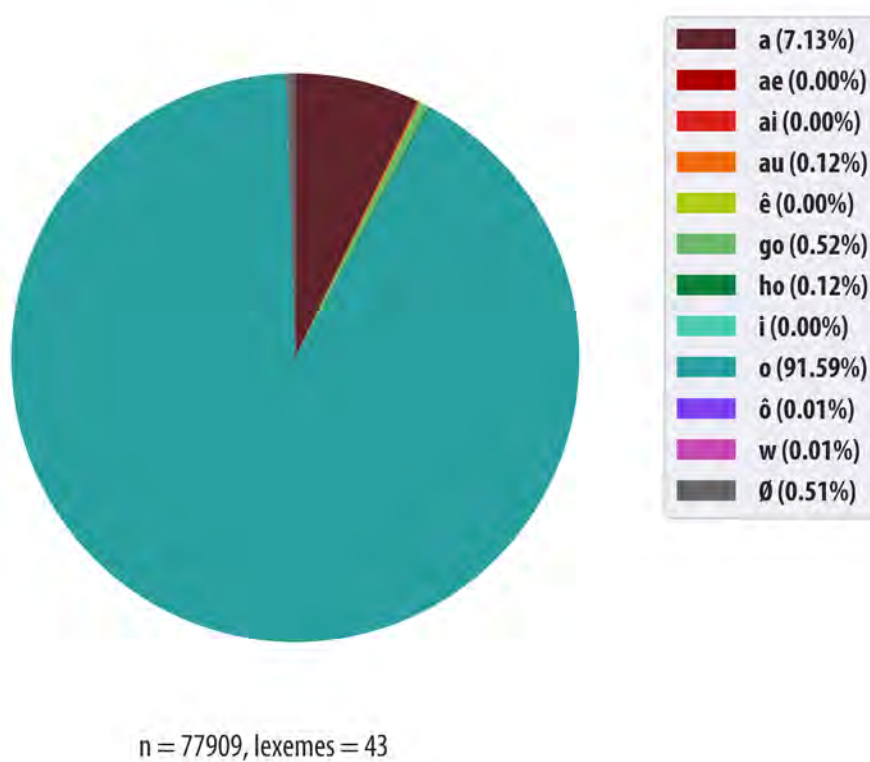


(a) GPC data

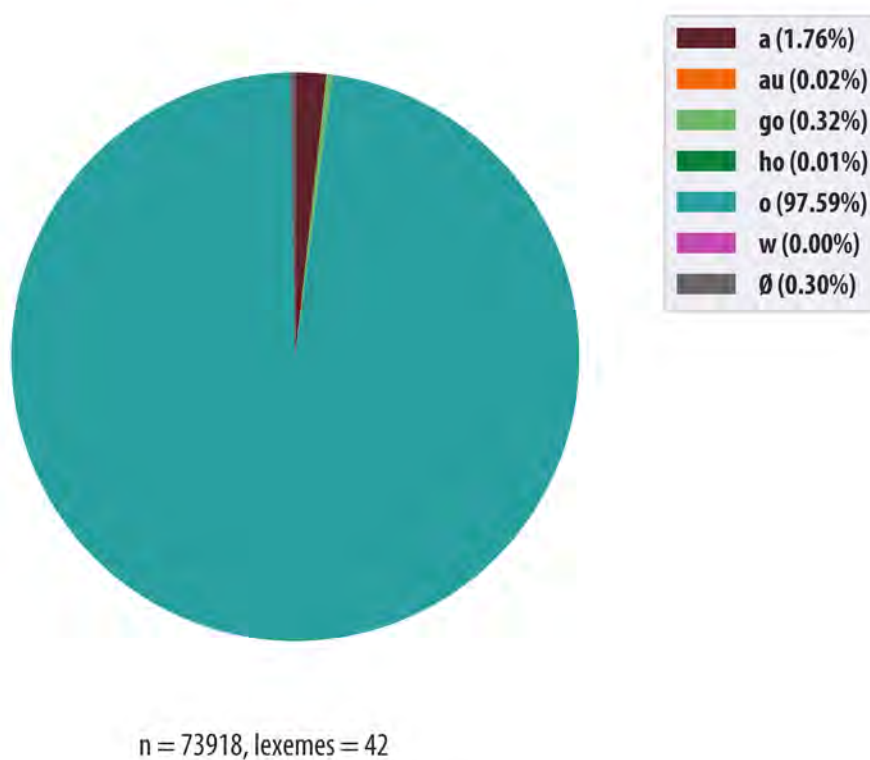


(b) Survey data

Figure C.10: Distribution of graphemes for phoneme /ɔ/ (LOT~WANT)

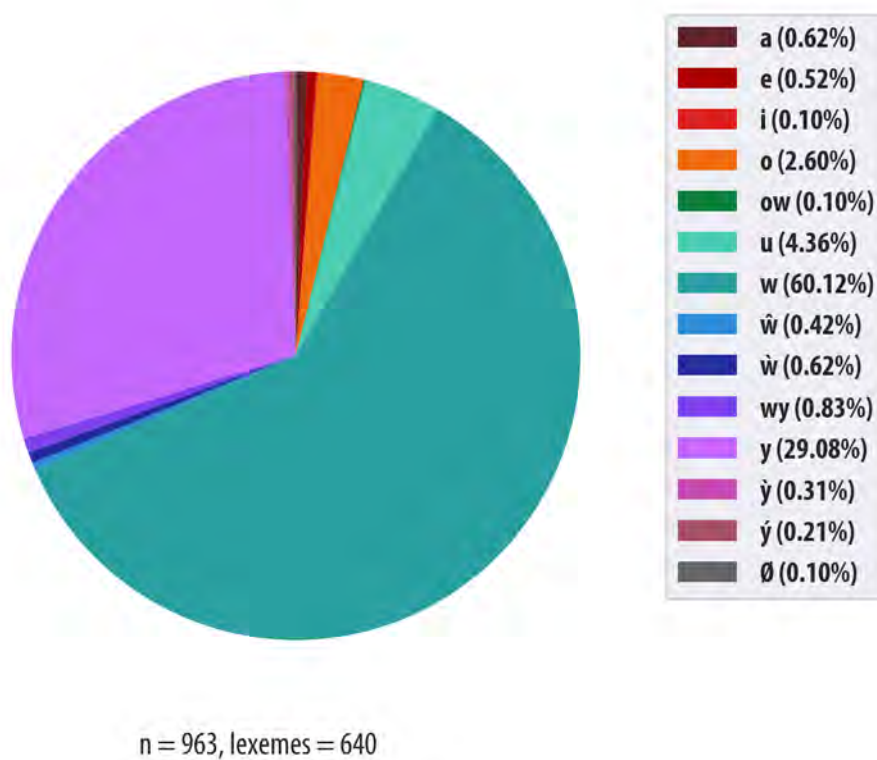


(c) Twitter (personal) data

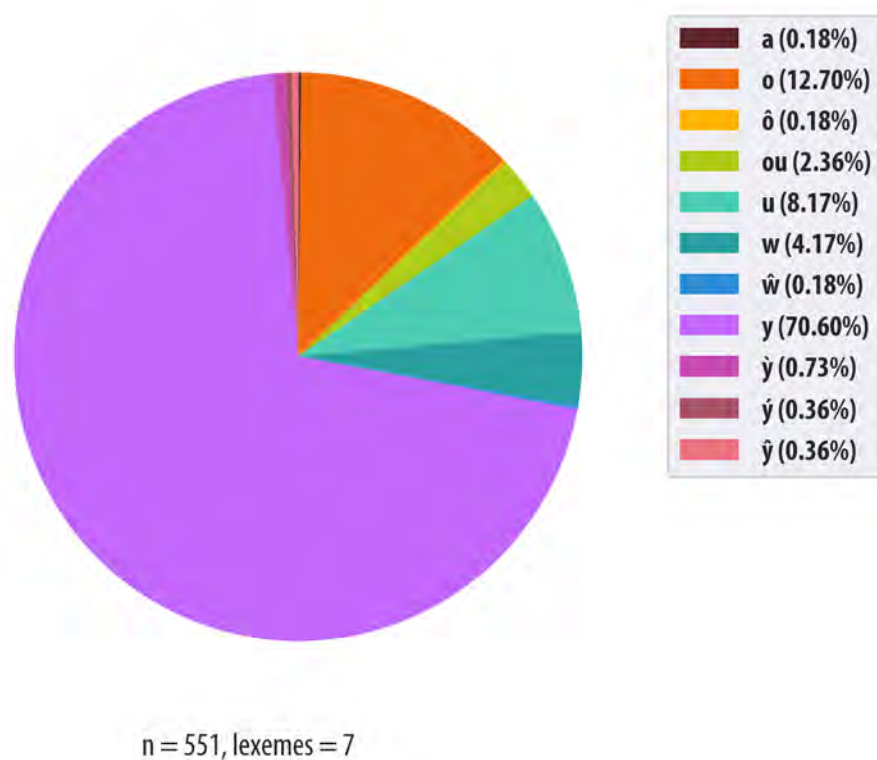


(d) Twitter (official) data

Figure C.10: Distribution of graphemes for phoneme /ɔ/ (**LOT**~**WANT**) (cont.)

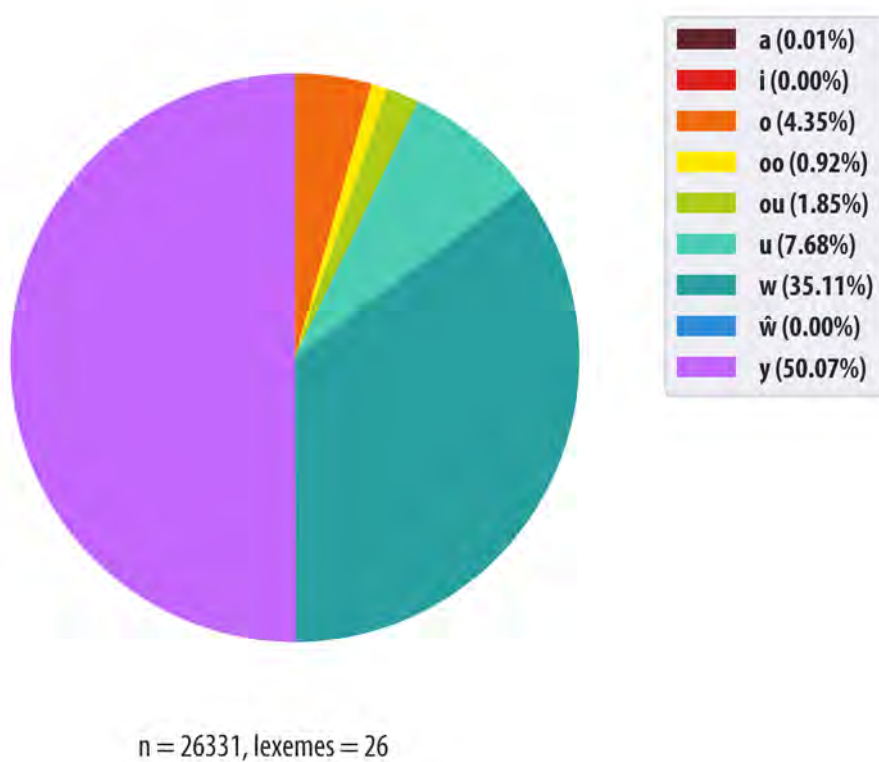


(a) GPC data

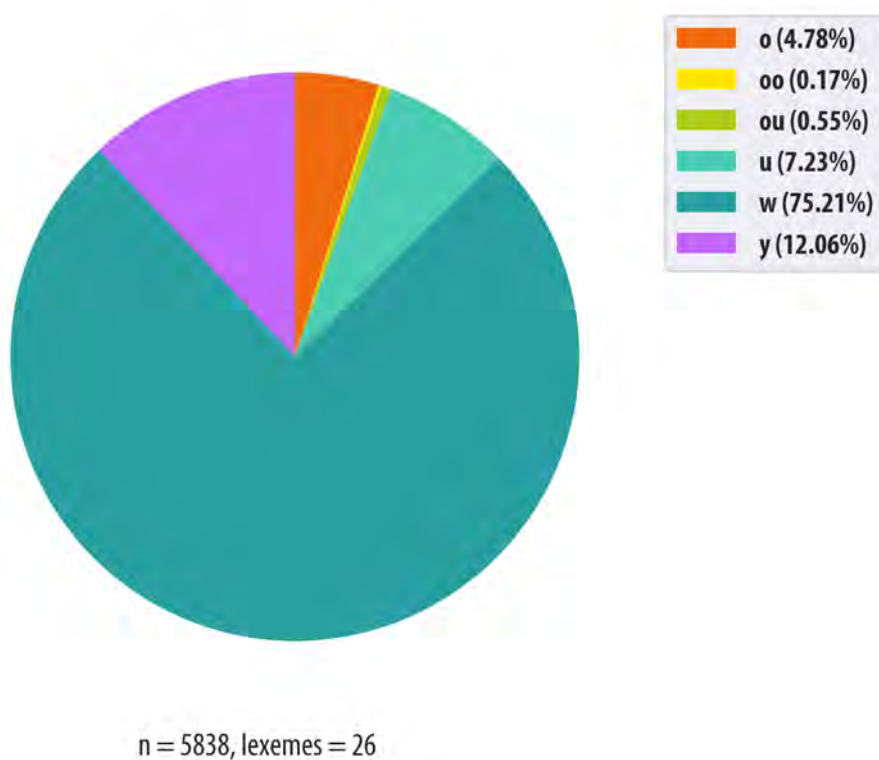


(b) Survey data

Figure C.11: Distribution of graphemes for phoneme /ʌ/ (**STRUT**)

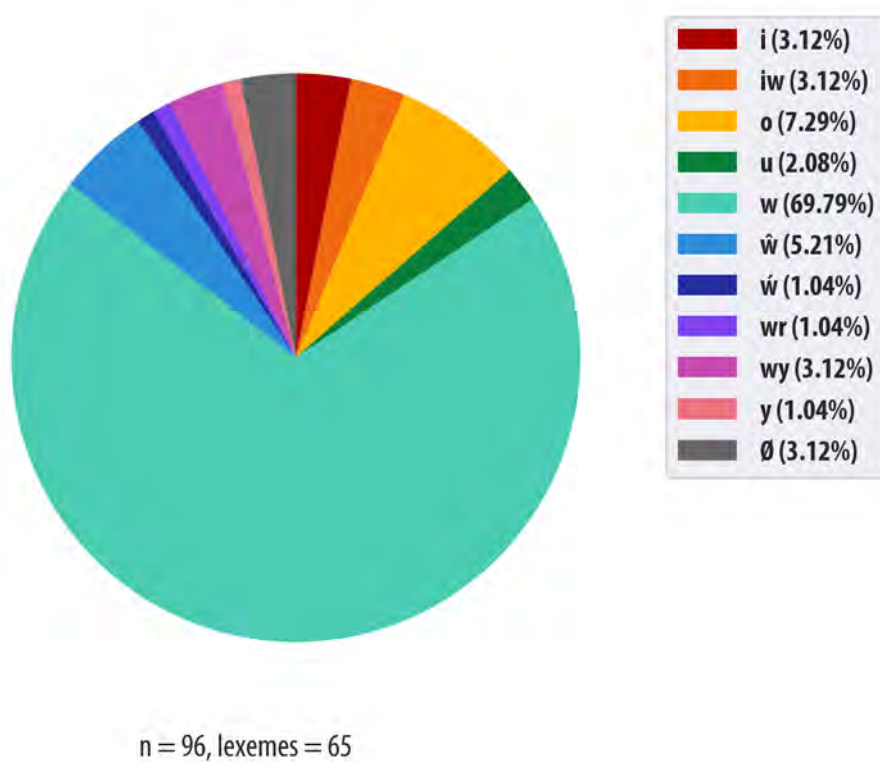


(c) Twitter (personal) data

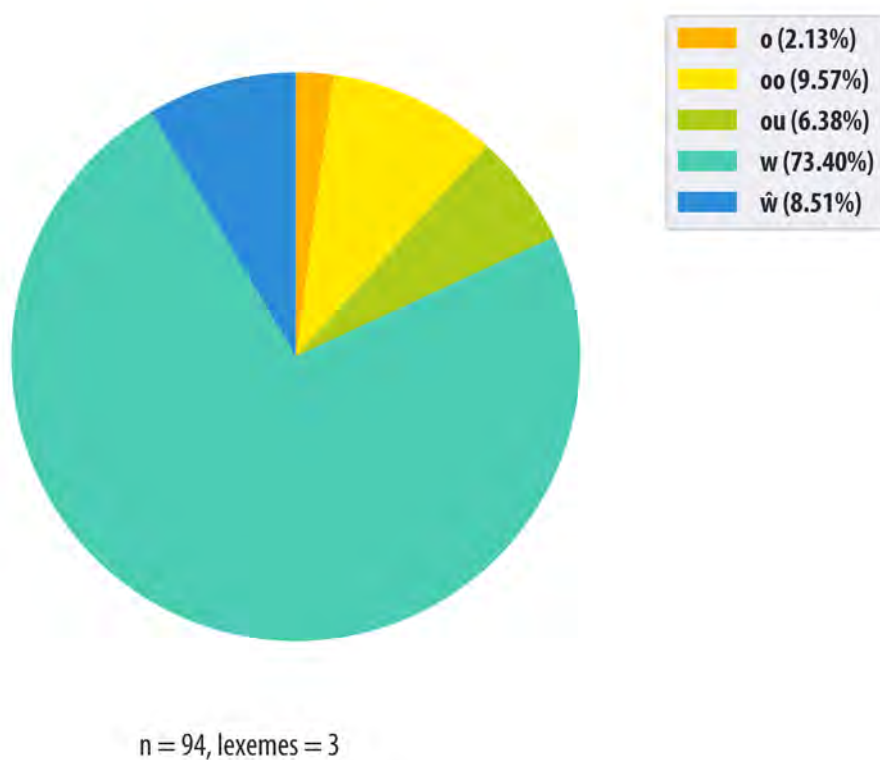


(d) Twitter (official) data

Figure C.11: Distribution of graphemes for phoneme /ʌ/ (**STRUT**) (cont.)

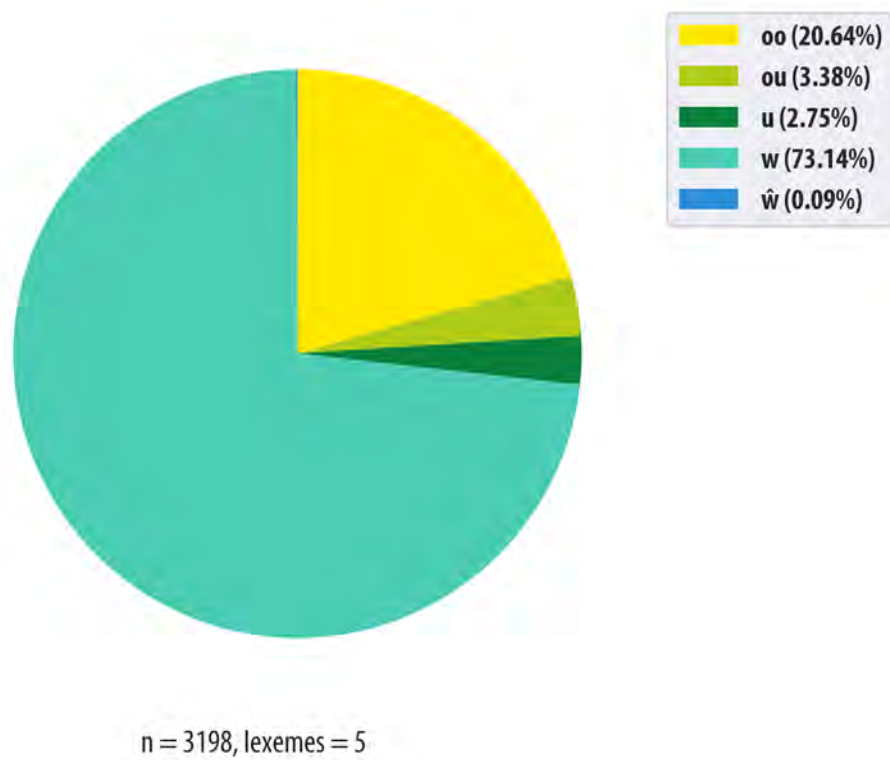


(a) GPC data

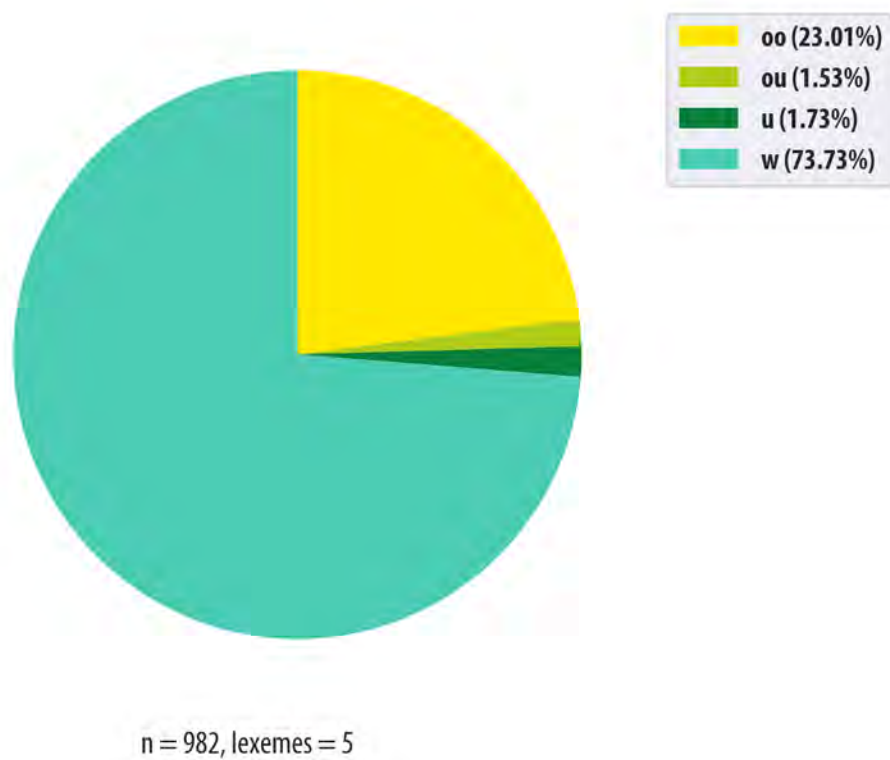


(b) Survey data

Figure C.12: Distribution of graphemes for phoneme /ʊ/ (**FOOT**)

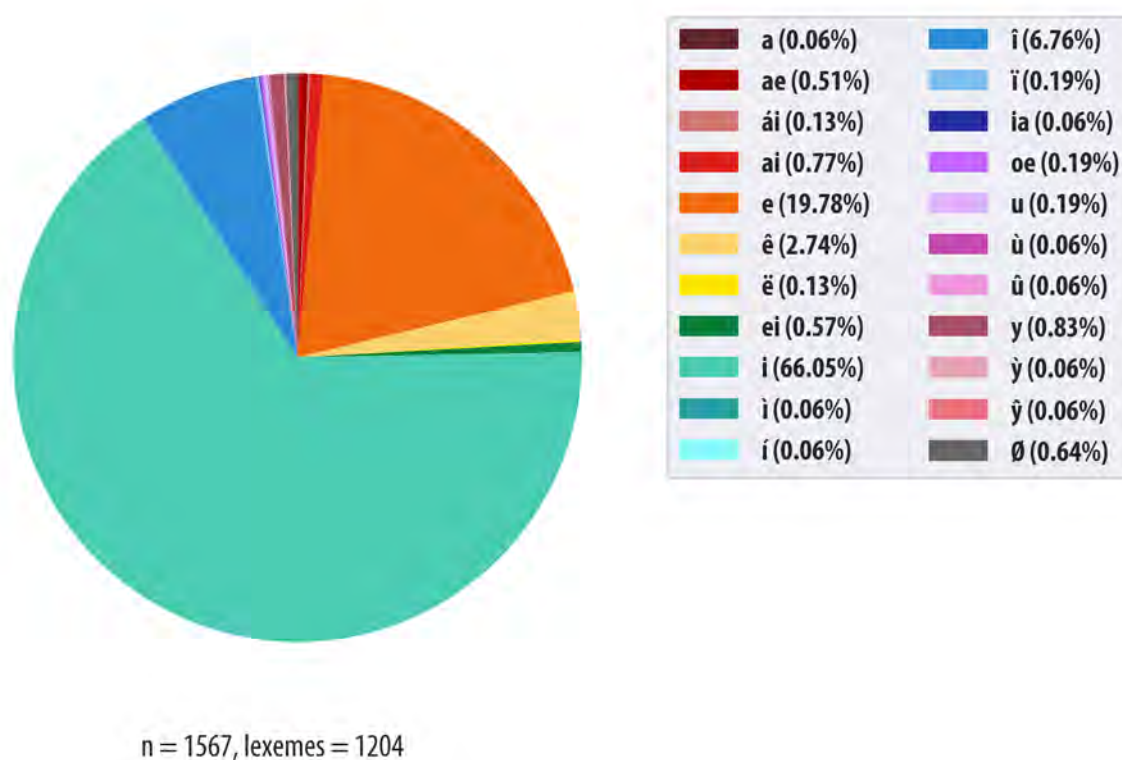


(c) Twitter (personal) data

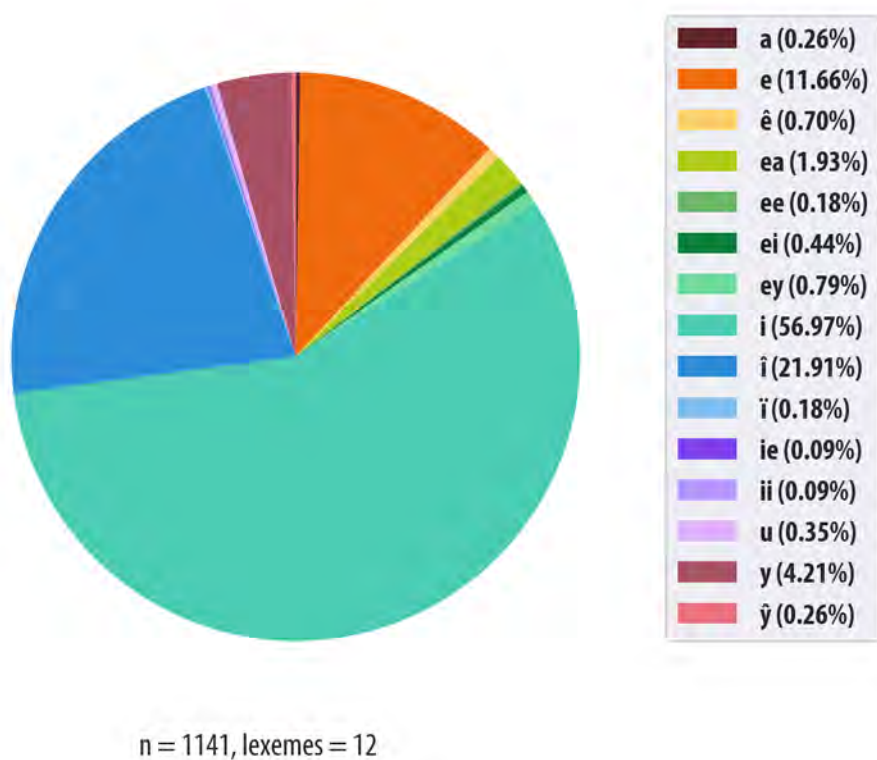


(d) Twitter (official) data

Figure C.12: Distribution of graphemes for phoneme /ʊ/ (**FOOT**) (cont.)

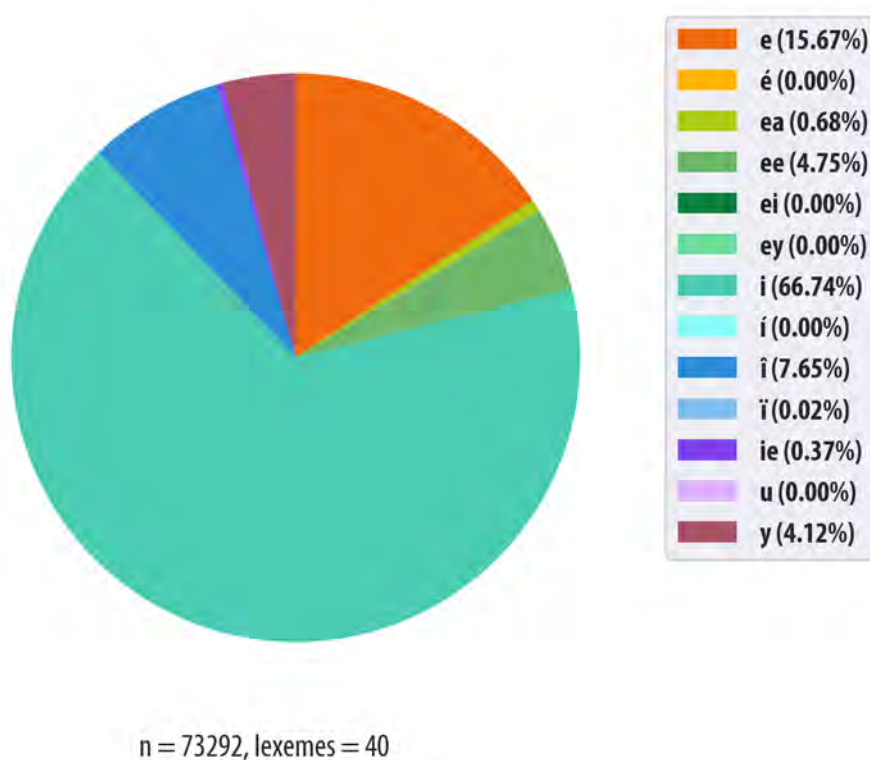


(a) GPC data

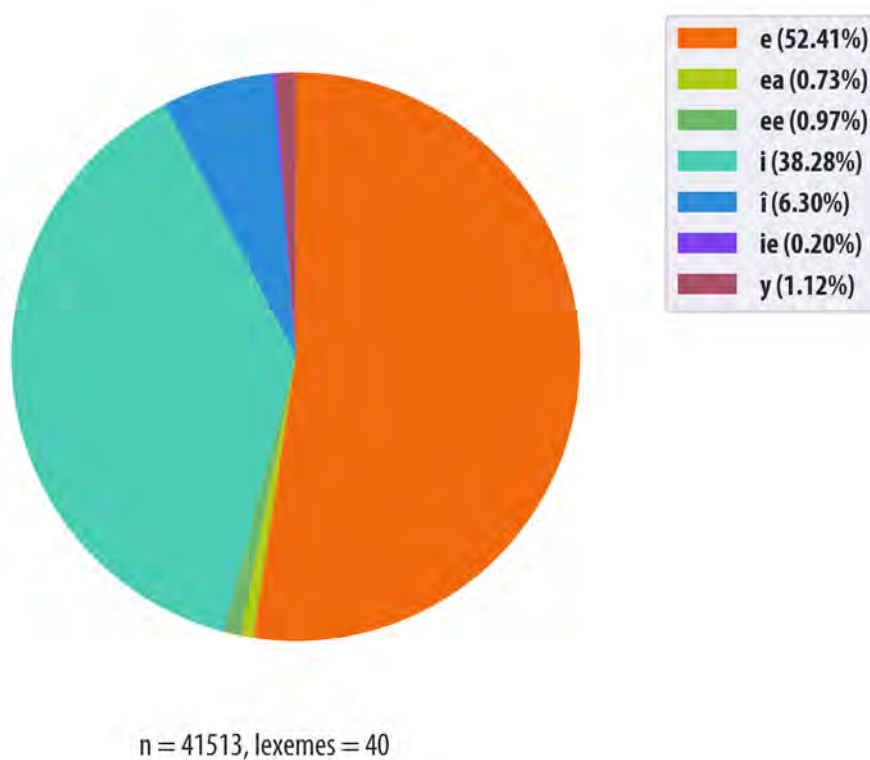


(b) Survey data

Figure C.13: Distribution of graphemes for phoneme /i:/ (FLEECE)

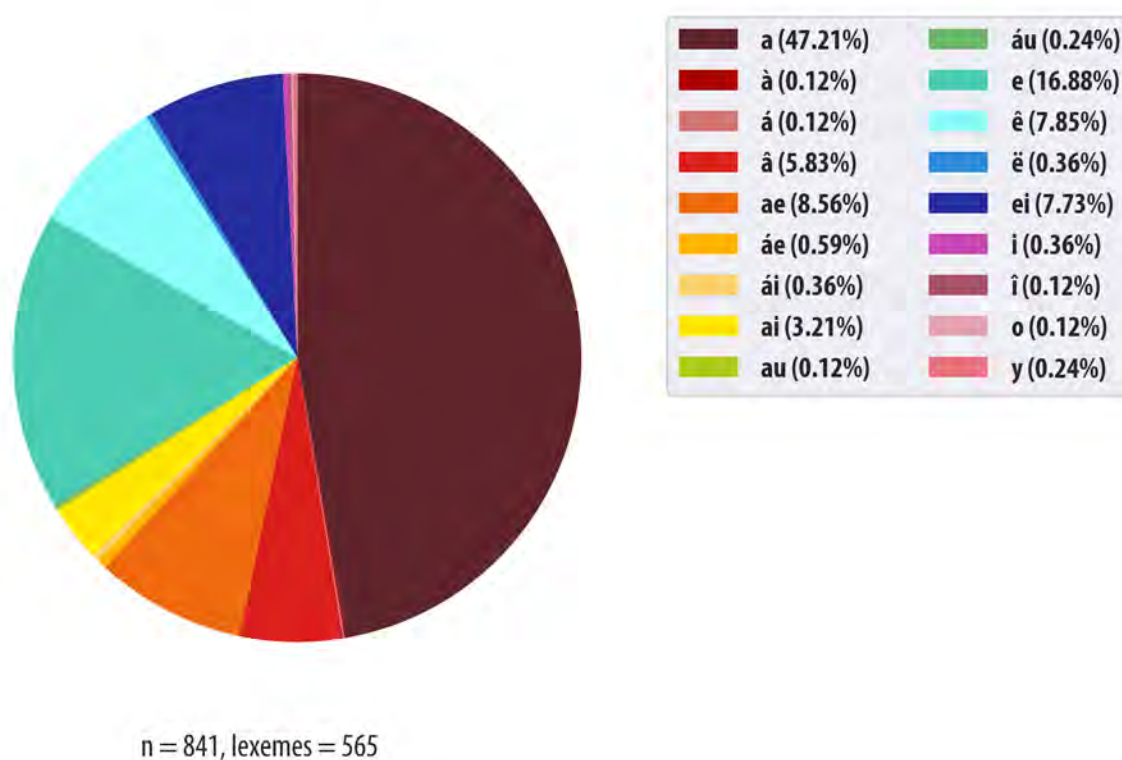


(c) Twitter (personal) data

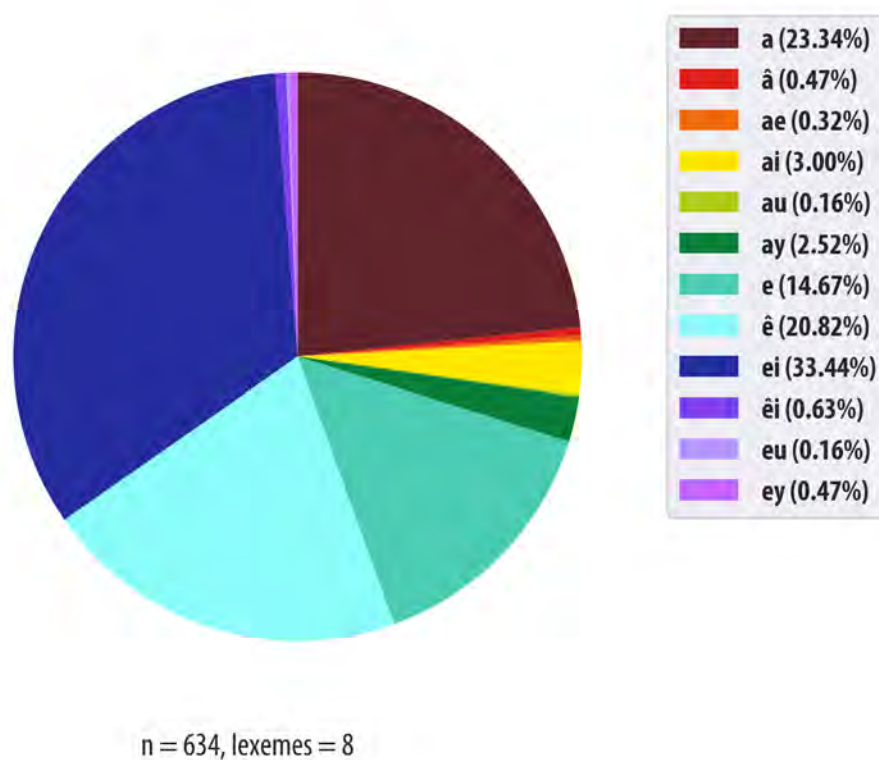


(d) Twitter (official) data

Figure C.13: Distribution of graphemes for phoneme /i/ (FLEECE) (cont.)

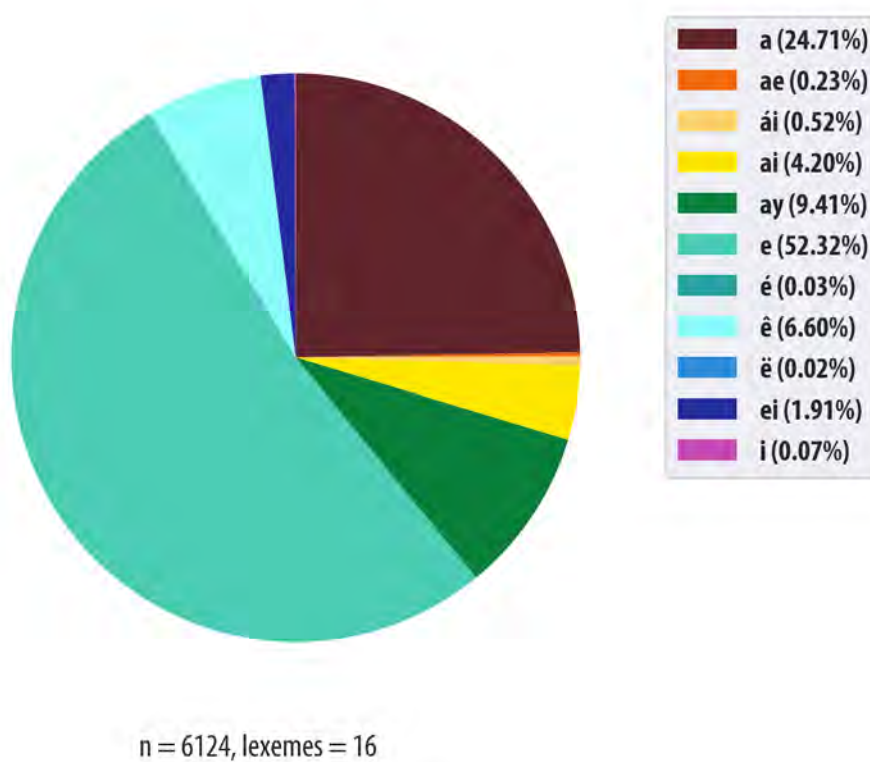


(a) GPC data

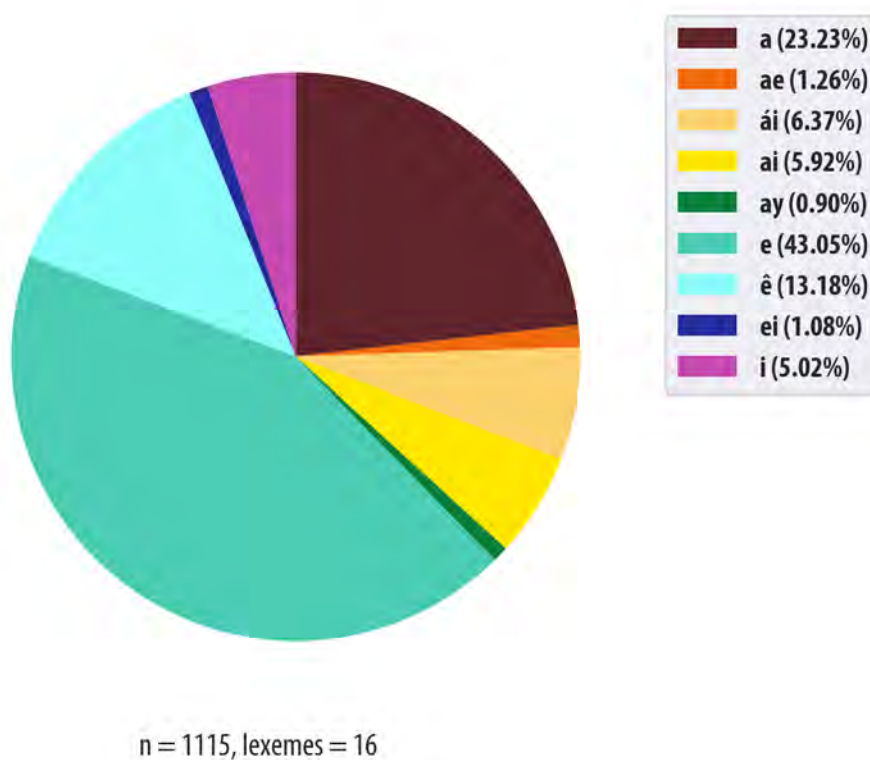


(b) Survey data

Figure C.14: Distribution of graphemes for phoneme /ei/ (**FACE**~**BAIT**)

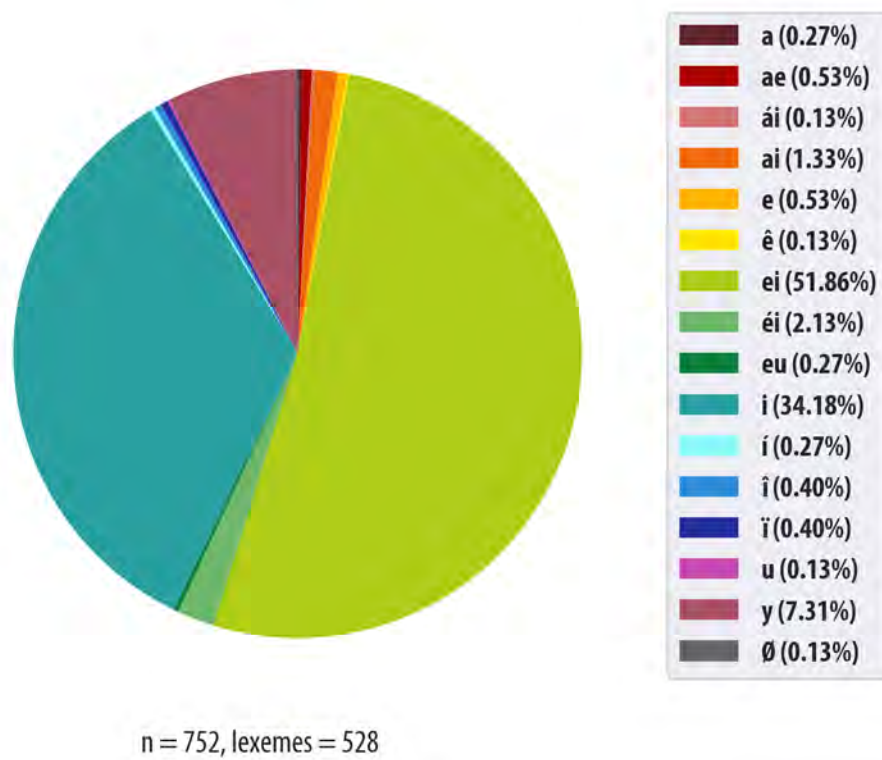


(c) Twitter (personal) data

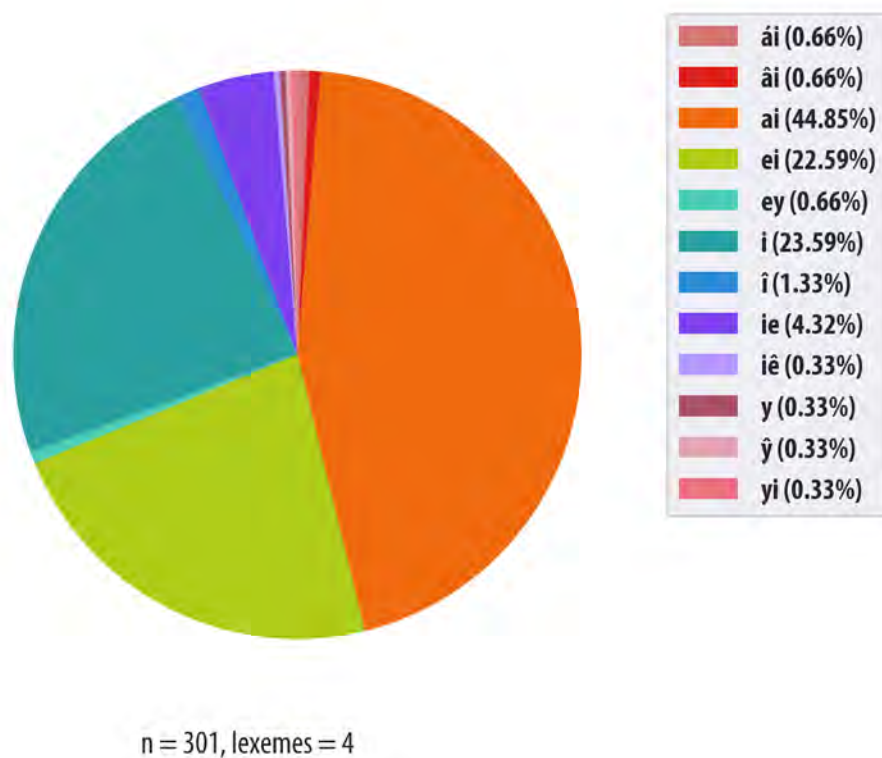


(d) Twitter (official) data

Figure C.14: Distribution of graphemes for phoneme /ei/ (**FACE**~**BAIT**) (cont.)

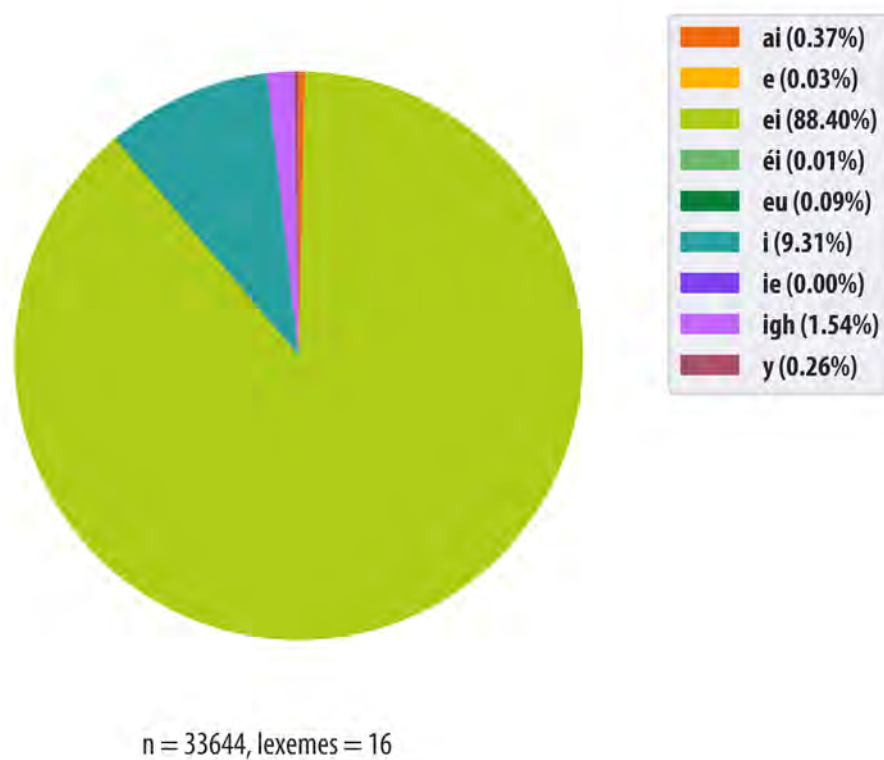


(a) GPC data

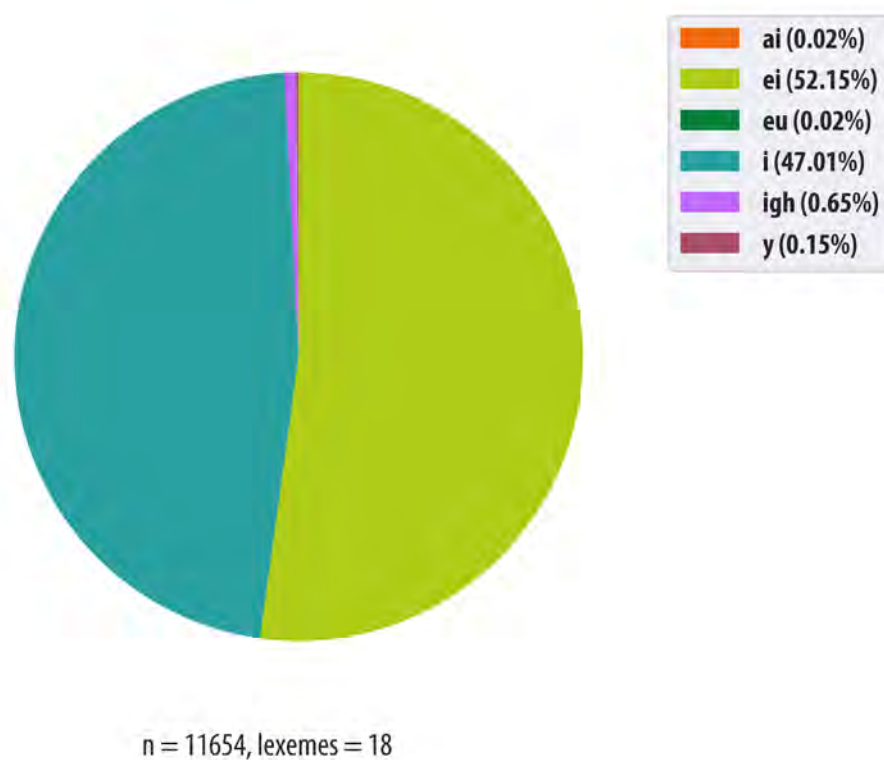


(b) Survey data

Figure C.15: Distribution of graphemes for phoneme /ai/ (**PRICE**)

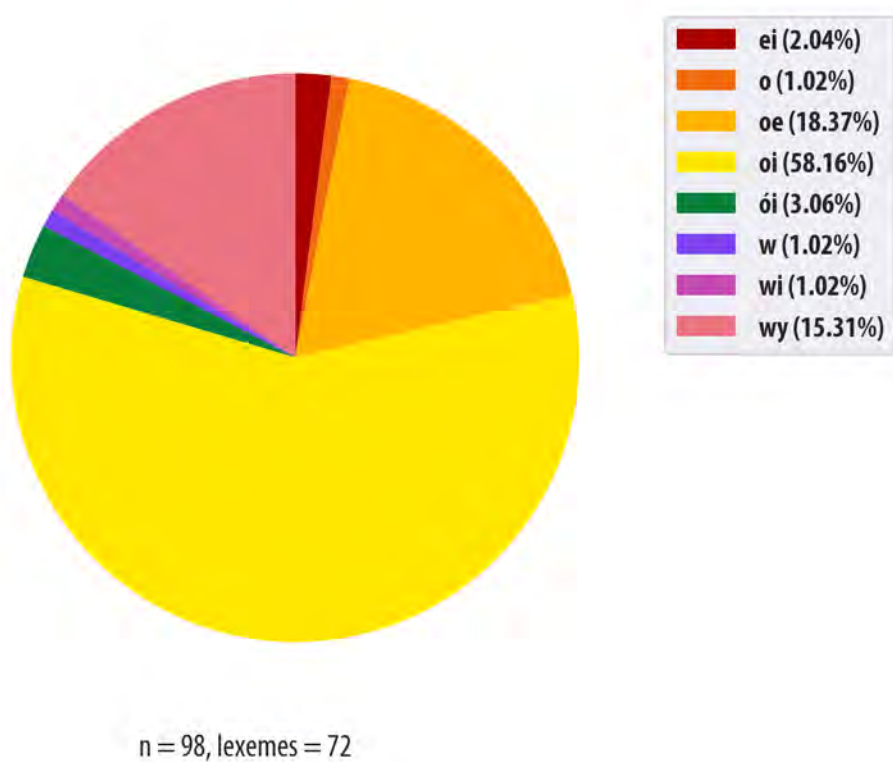


(c) Twitter (personal) data

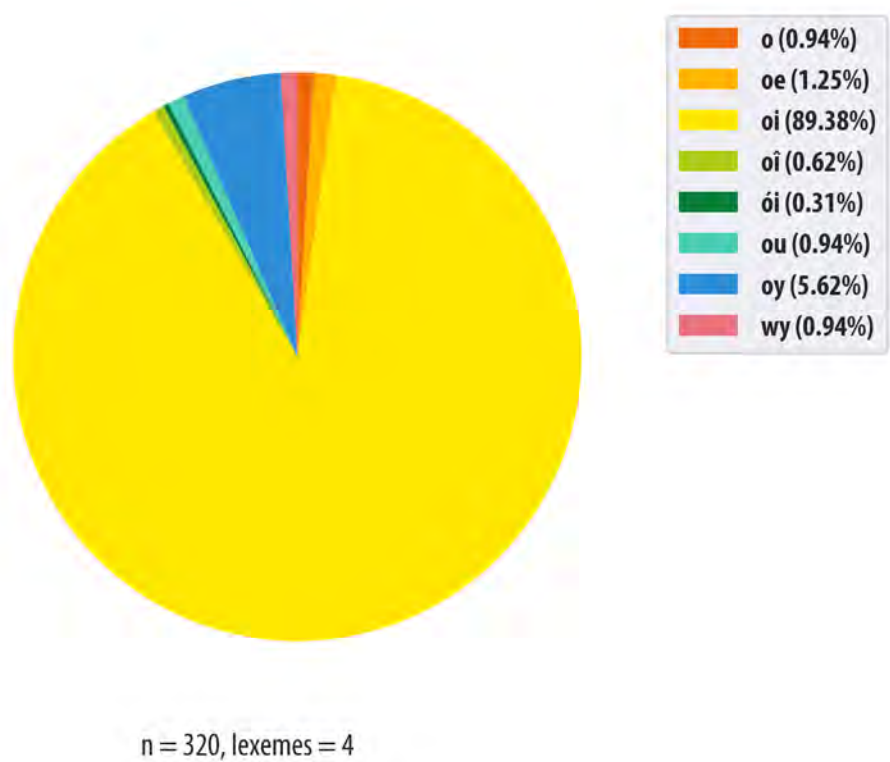


(d) Twitter (official) data

Figure C.15: Distribution of graphemes for phoneme /ai/ (**PRICE**) (cont.)

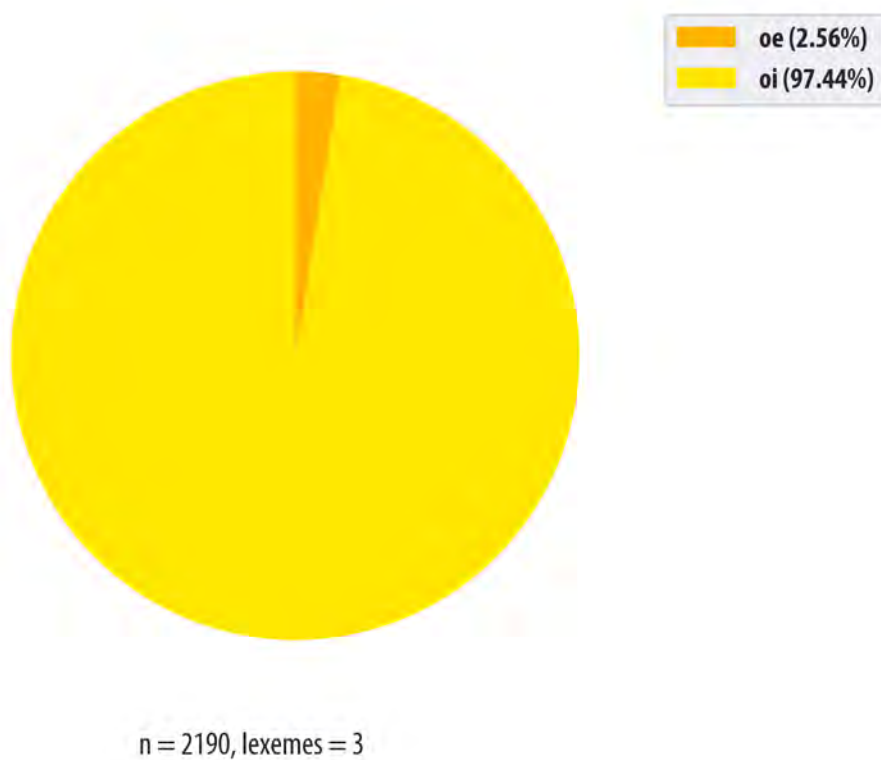


(a) GPC data

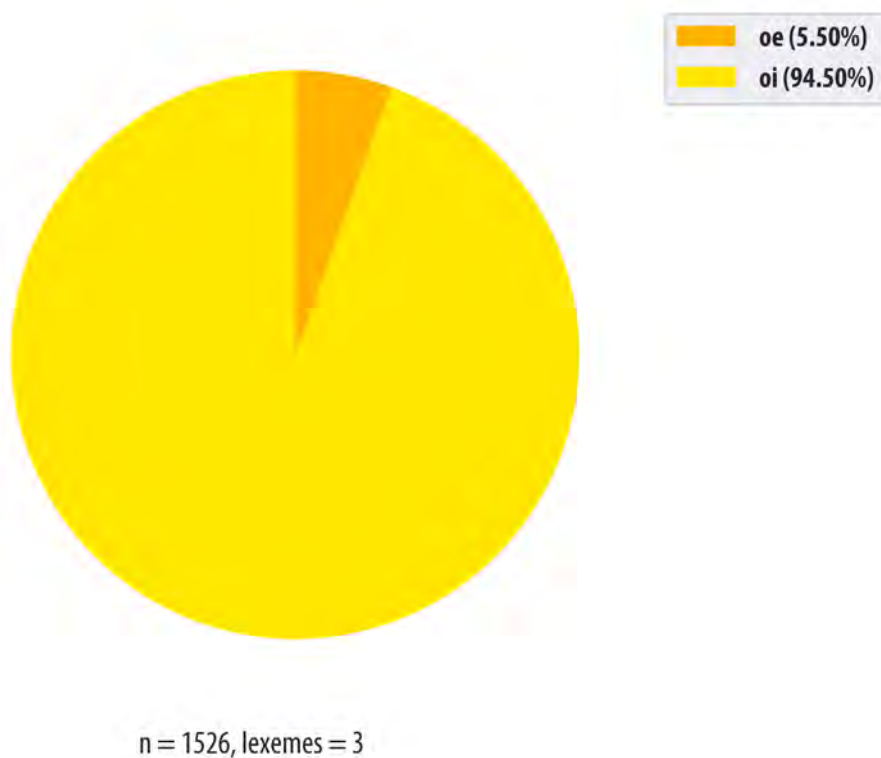


(b) Survey data

Figure C.16: Distribution of graphemes for phoneme /oi/ (**CHOICE**)

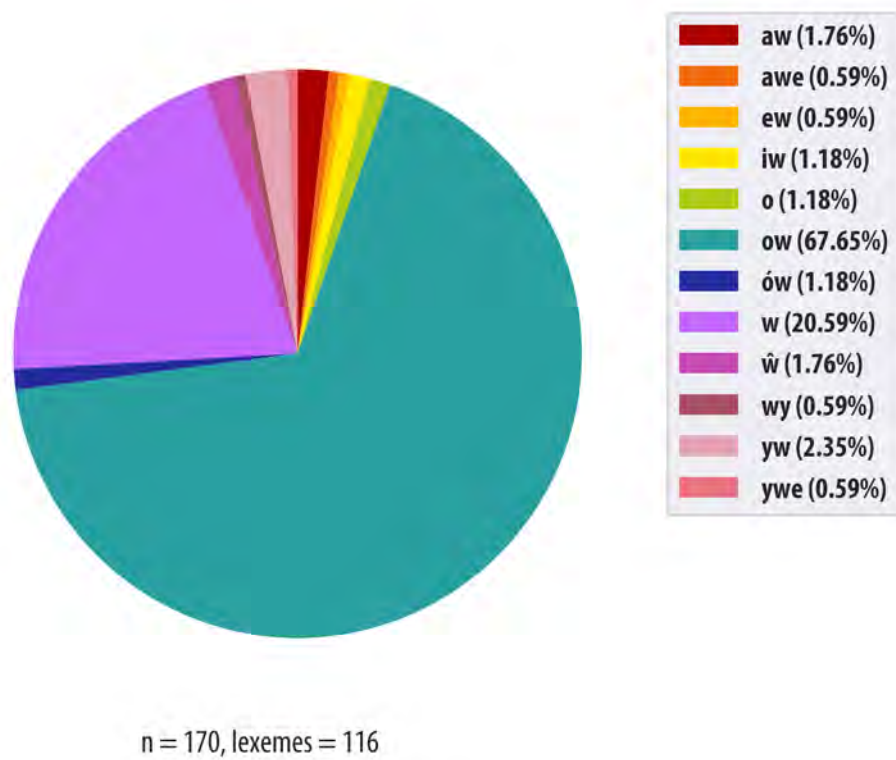


(c) Twitter (personal) data

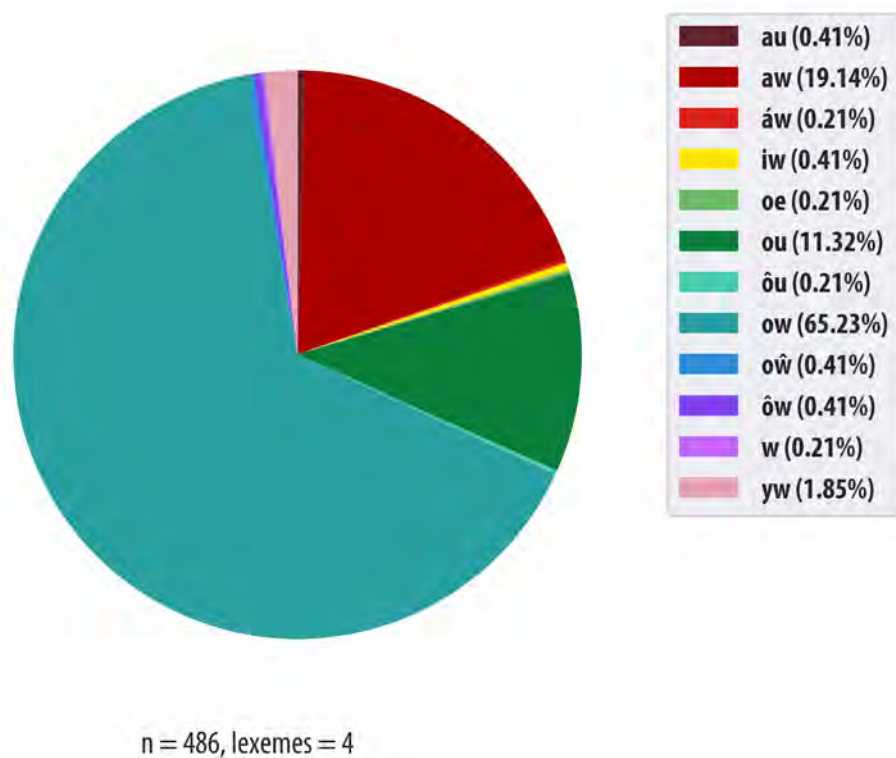


(d) Twitter (official) data

Figure C.16: Distribution of graphemes for phoneme /oi/ (**CHOICE**) (cont.)

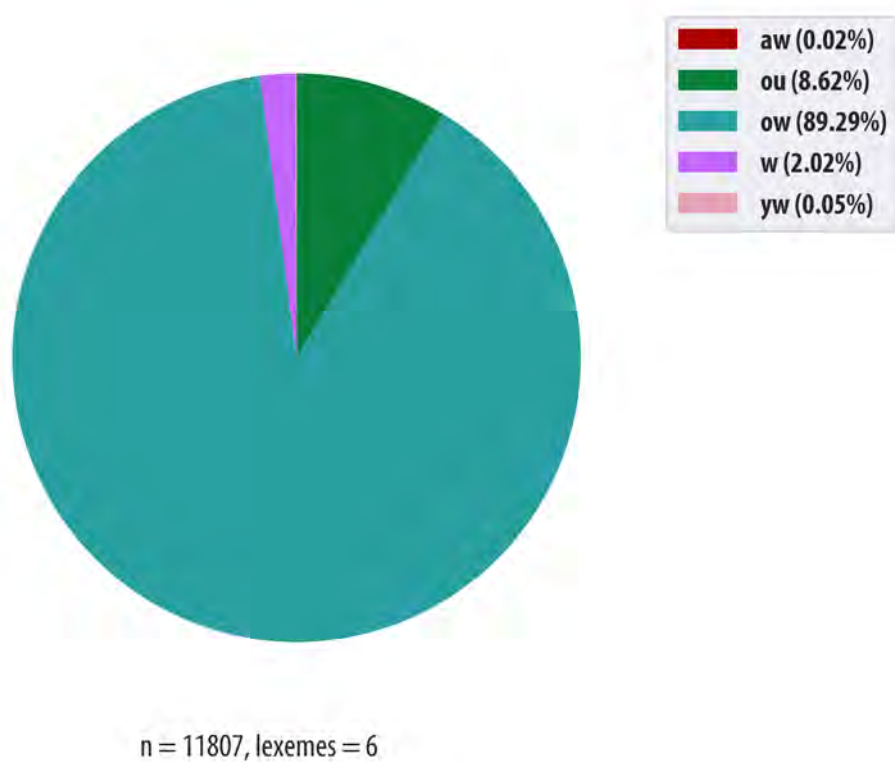


(a) GPC data

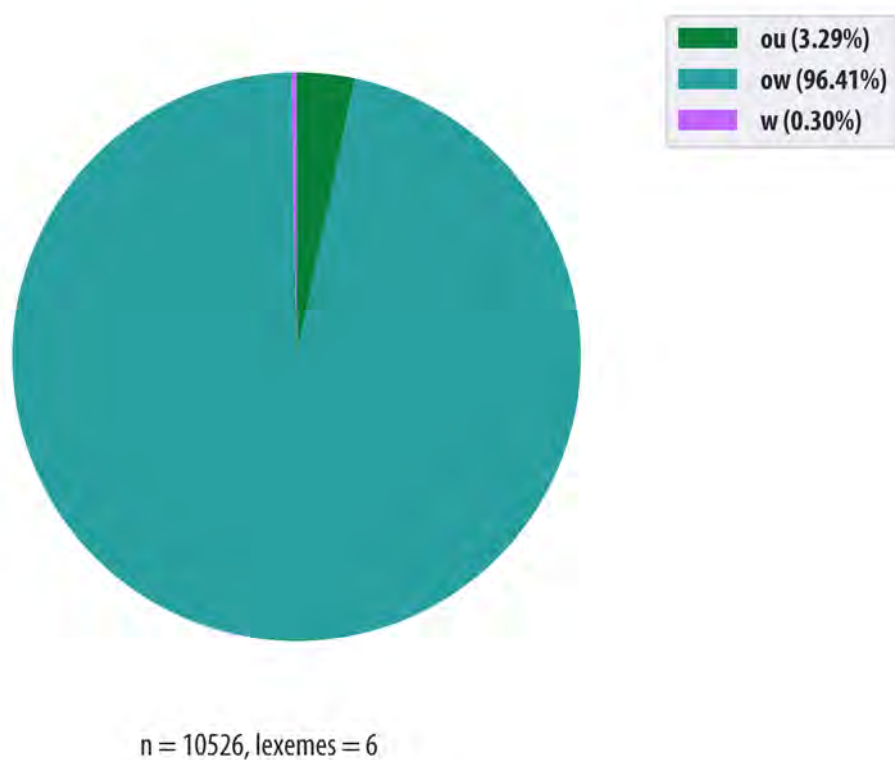


(b) Survey data

Figure C.17: Distribution of graphemes for phoneme /au/ (**mouth**)

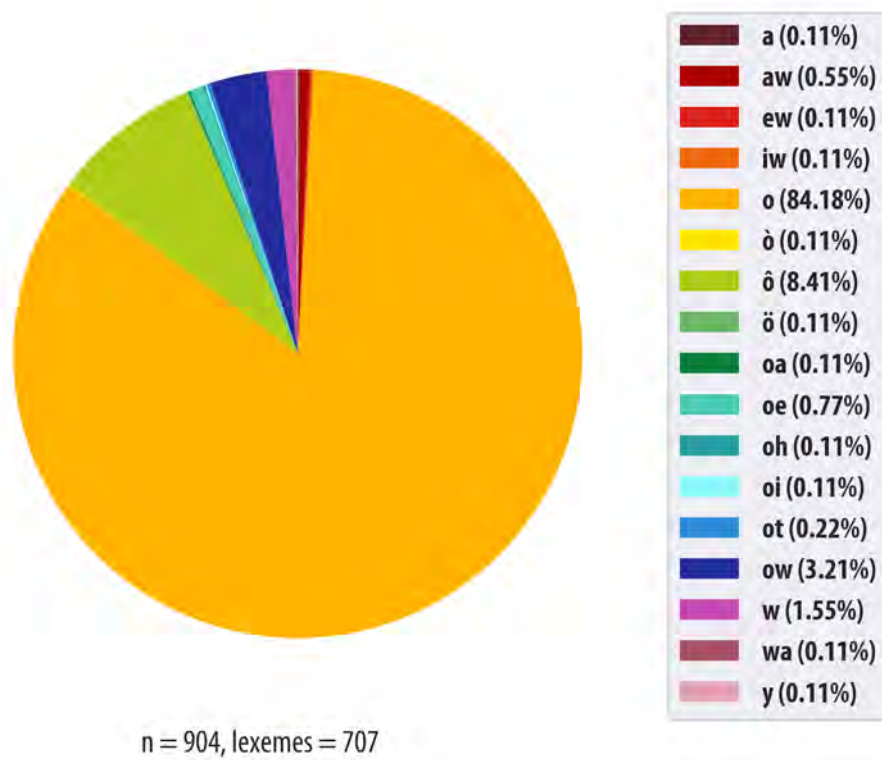


(c) Twitter (personal) data

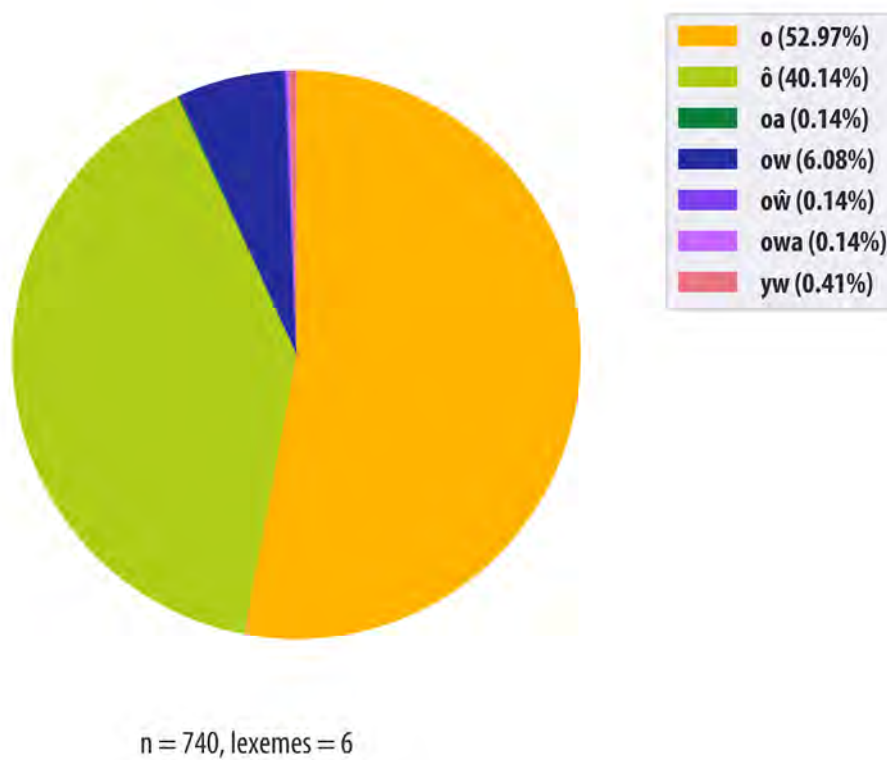


(d) Twitter (official) data

Figure C.17: Distribution of graphemes for phoneme /au/ (**mouth**) (cont.)

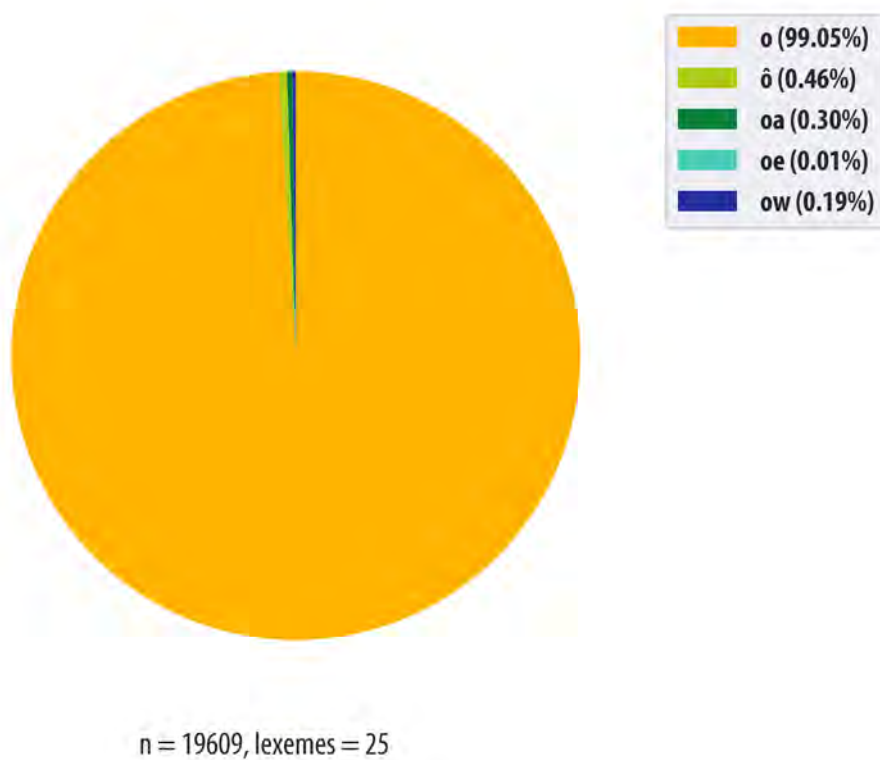


(a) GPC data

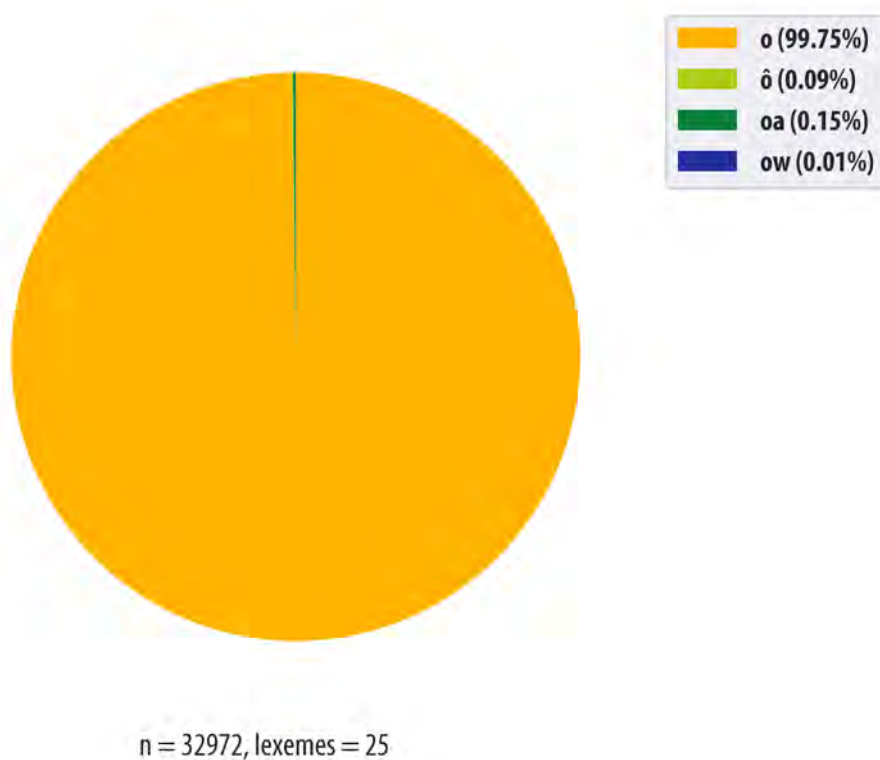


(b) Survey data

Figure C.18: Distribution of graphemes for phoneme /əʊ/ (**GOAT**~**KNOW**)

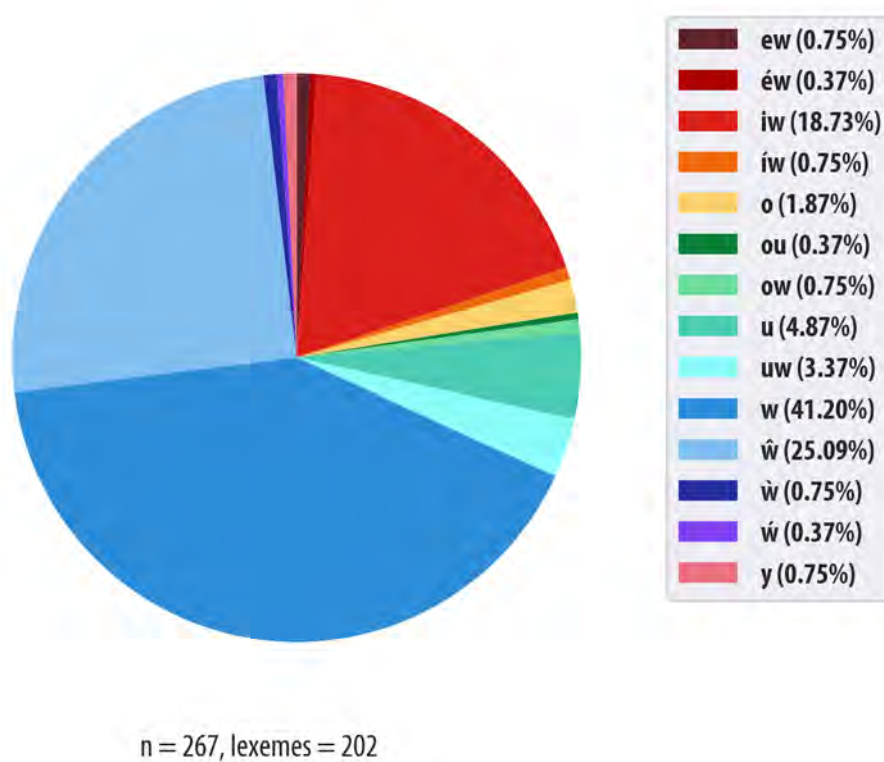


(c) Twitter (personal) data

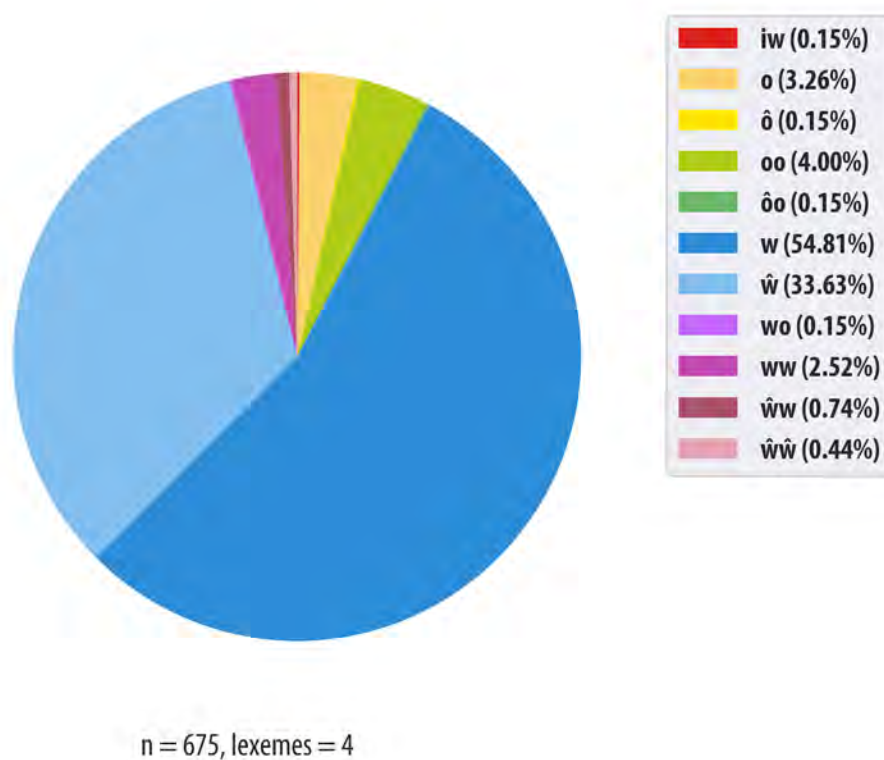


(d) Twitter (official) data

Figure C.18: Distribution of graphemes for phoneme /əʊ/ (**GOAT**~**KNOW**) (cont.)

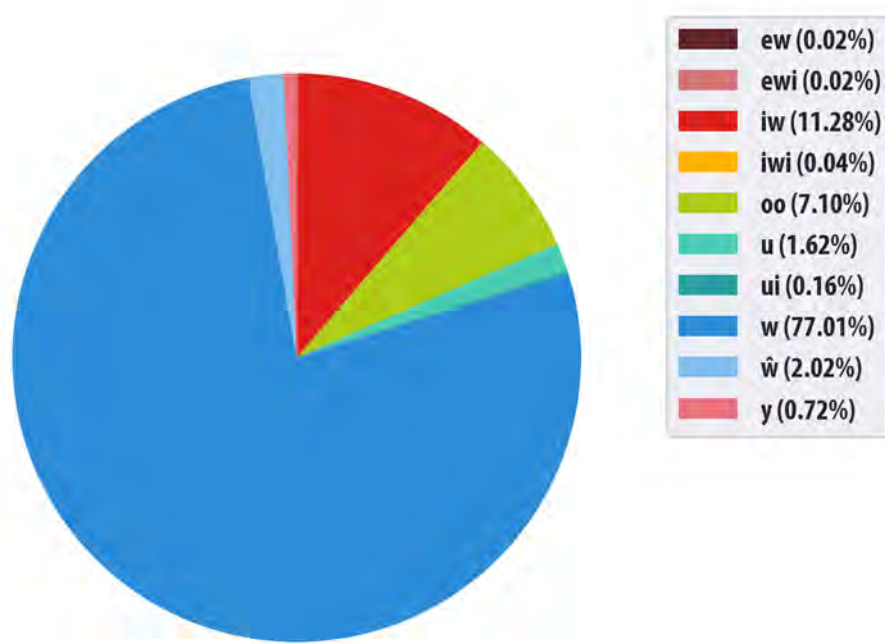


(a) GPC data



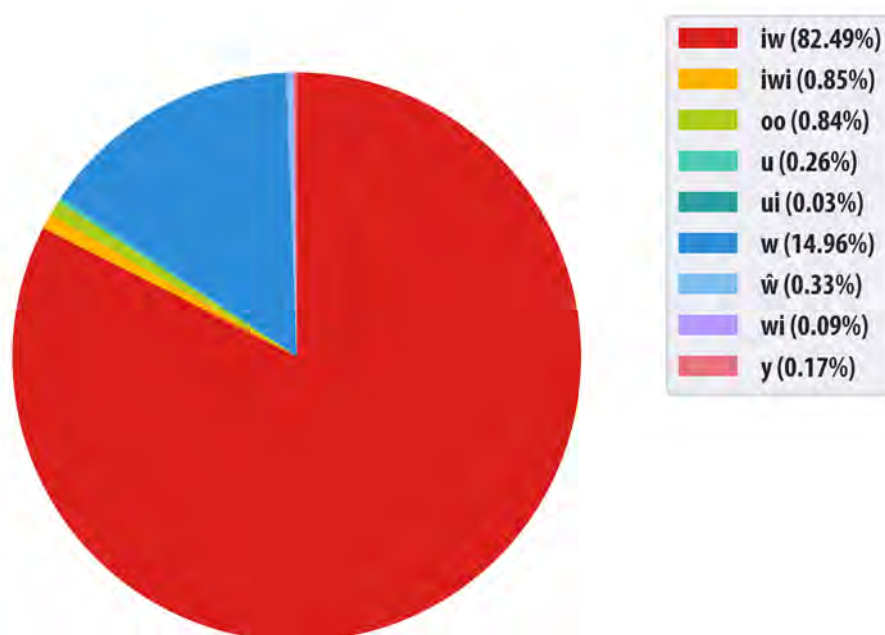
(b) Survey data

Figure C.19: Distribution of graphemes for phoneme /u/ (**GOOSE**~**FLEW**)



n = 4989, lexemes = 12

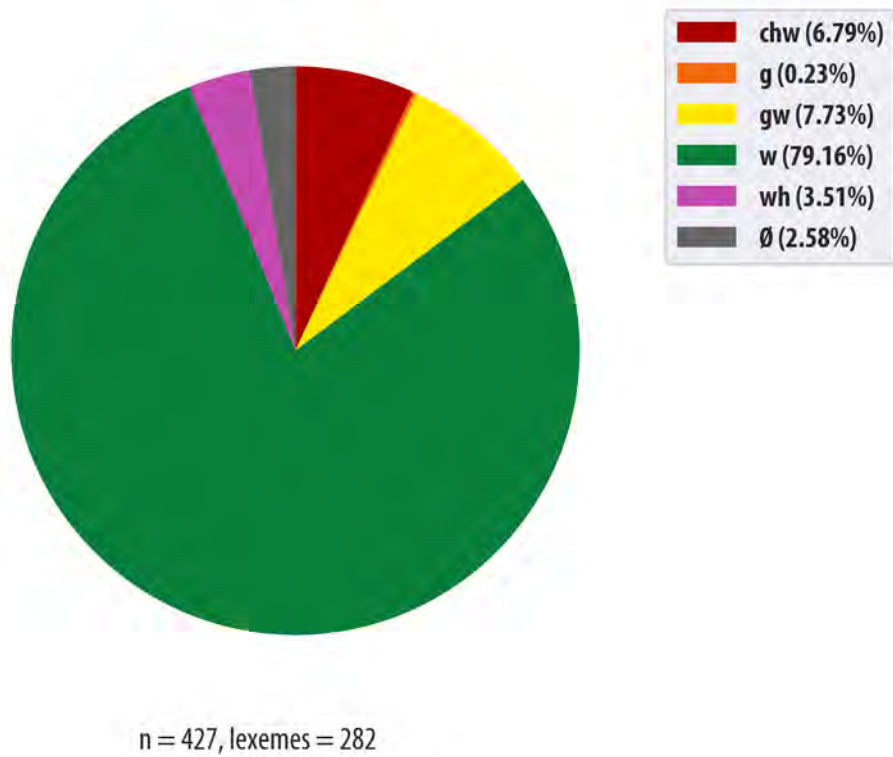
(c) Twitter (personal) data



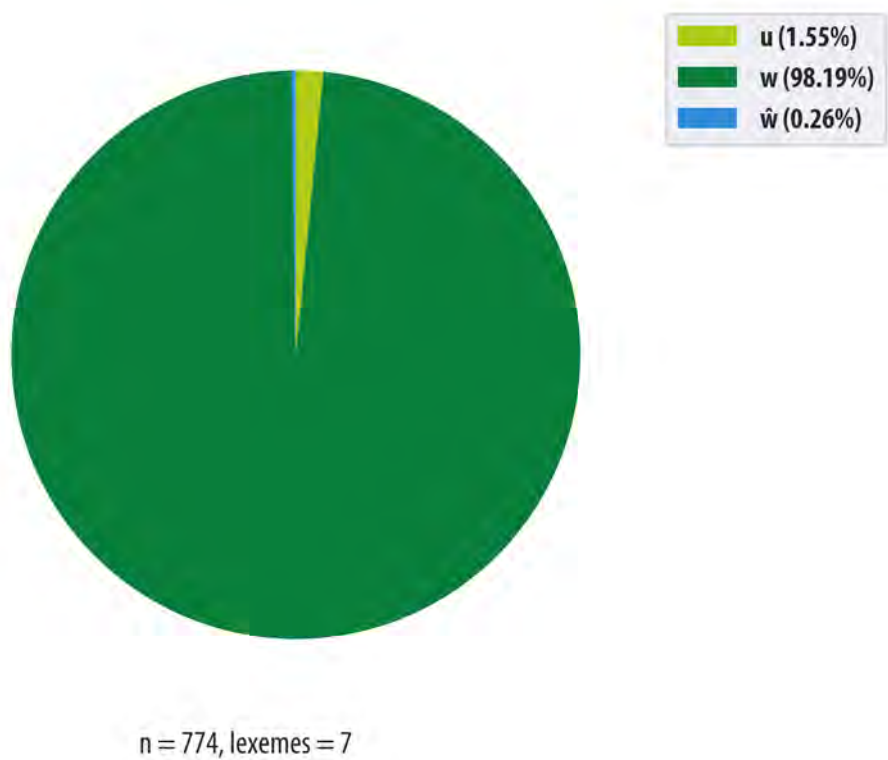
n = 7034, lexemes = 12

(d) Twitter (official) data

Figure C.19: Distribution of graphemes for phoneme /u/ (**GOOSE**~**FLEW**) (cont.)

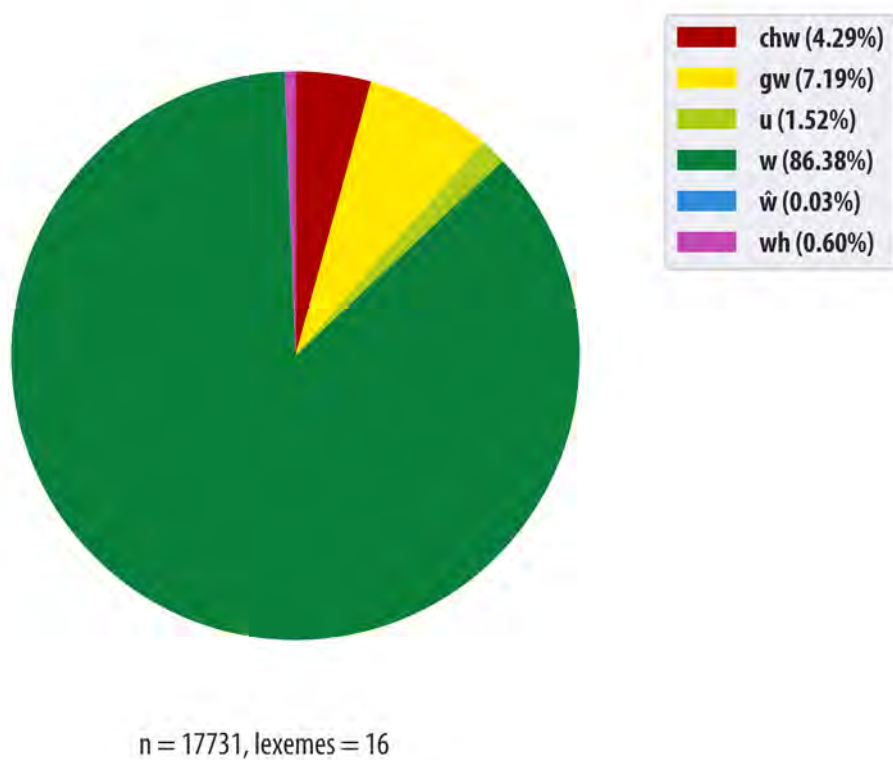


(a) GPC data

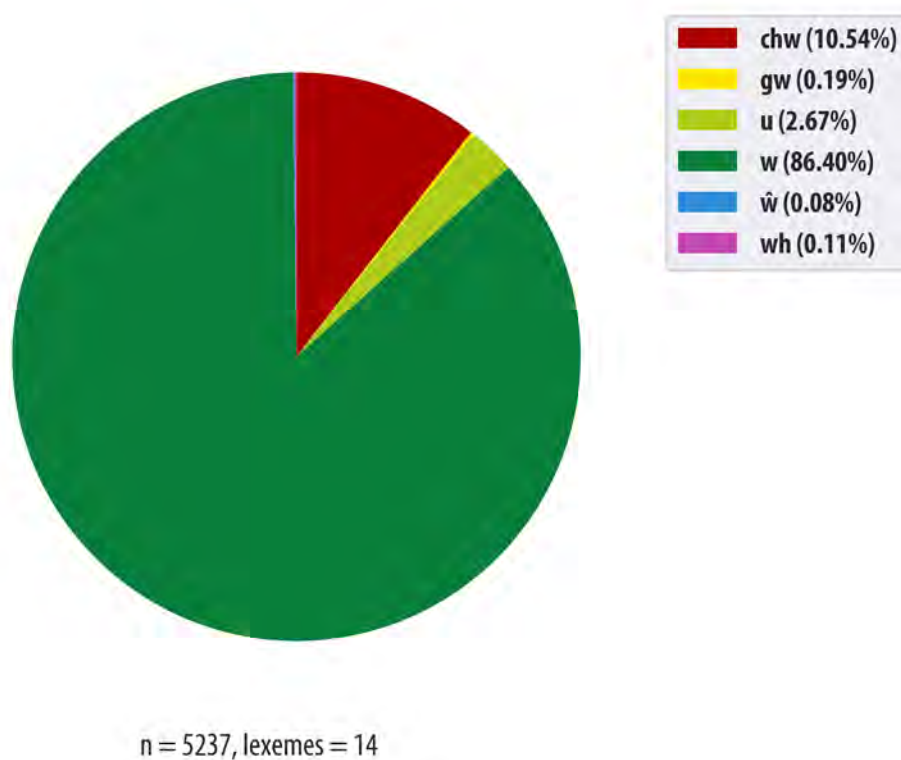


(b) Survey data

Figure C.20: Distribution of graphemes for phoneme /w/

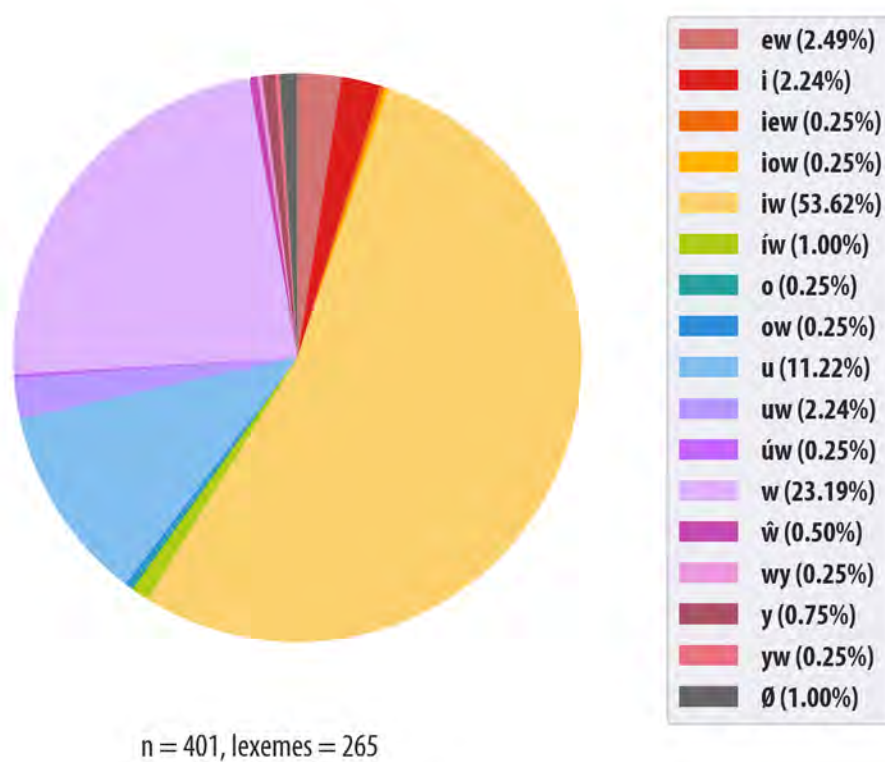


(c) Twitter (personal) data

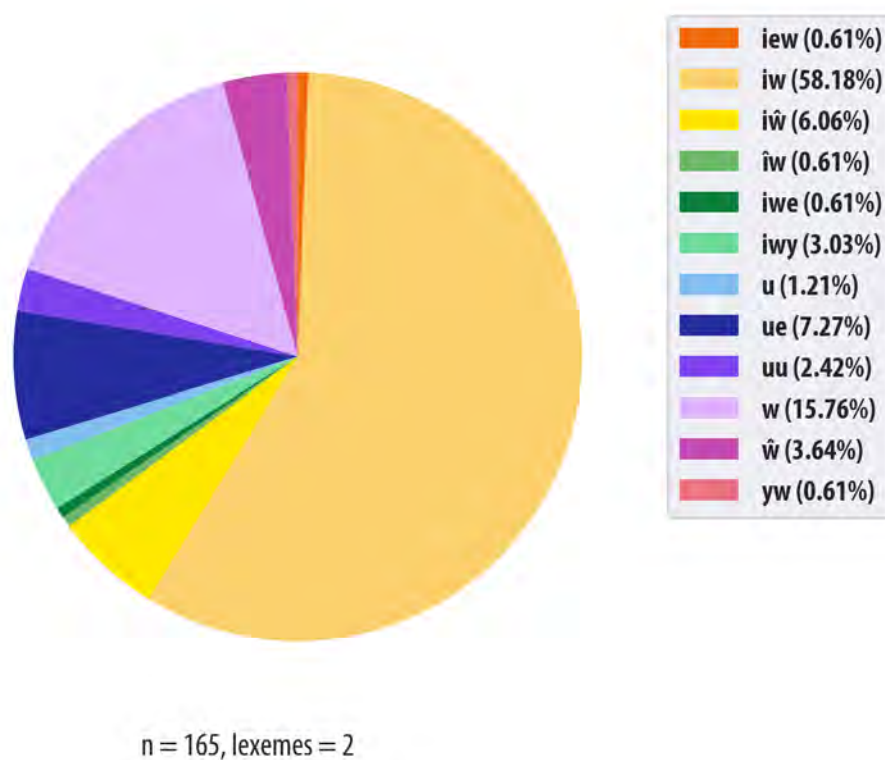


(d) Twitter (official) data

Figure C.20: Distribution of graphemes for phoneme /w/ (cont.)

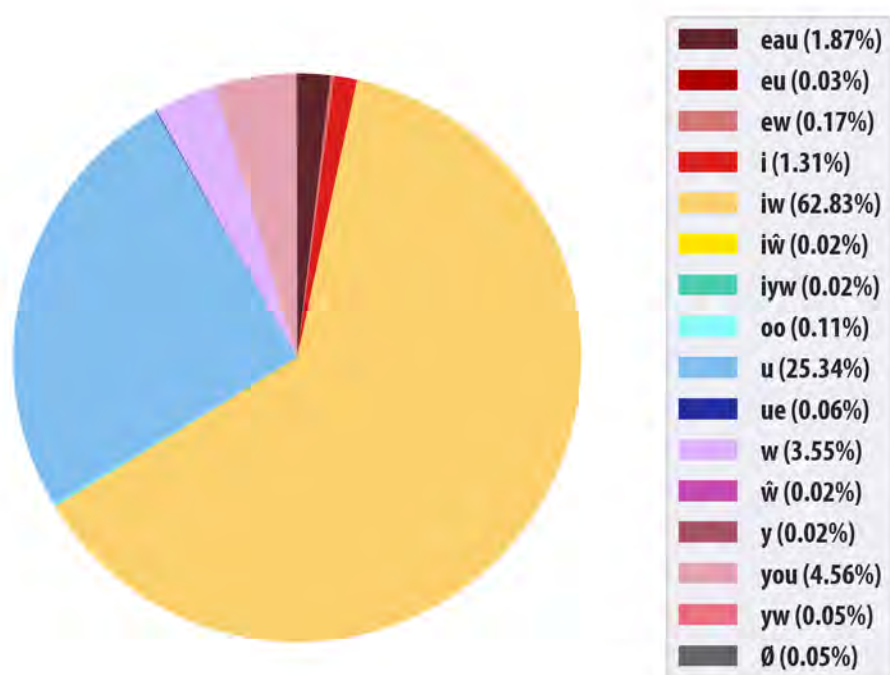


(a) GPC data



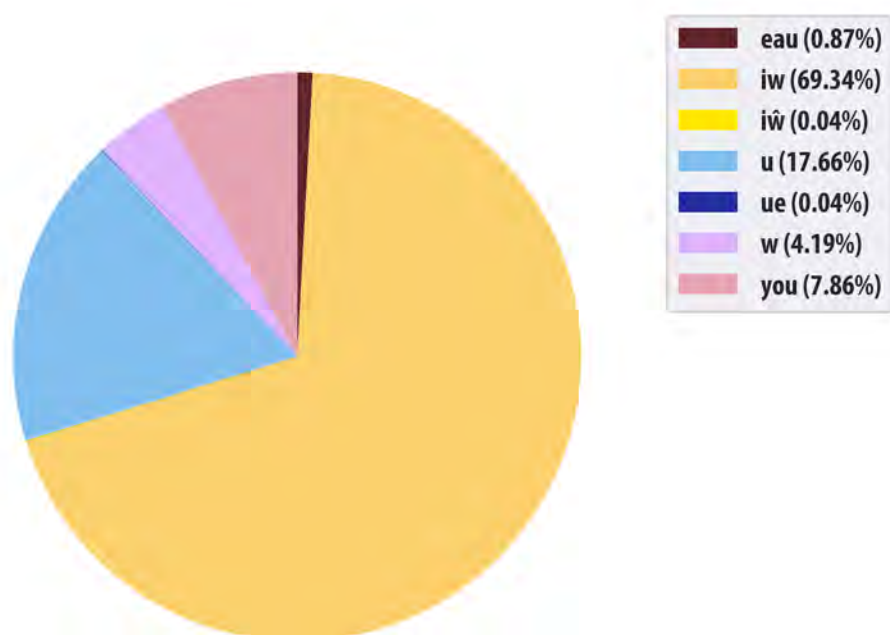
(b) Survey data

Figure C.21: Distribution of graphemes for phoneme /jɥ/ (TUNE)



n = 6314, lexemes = 25

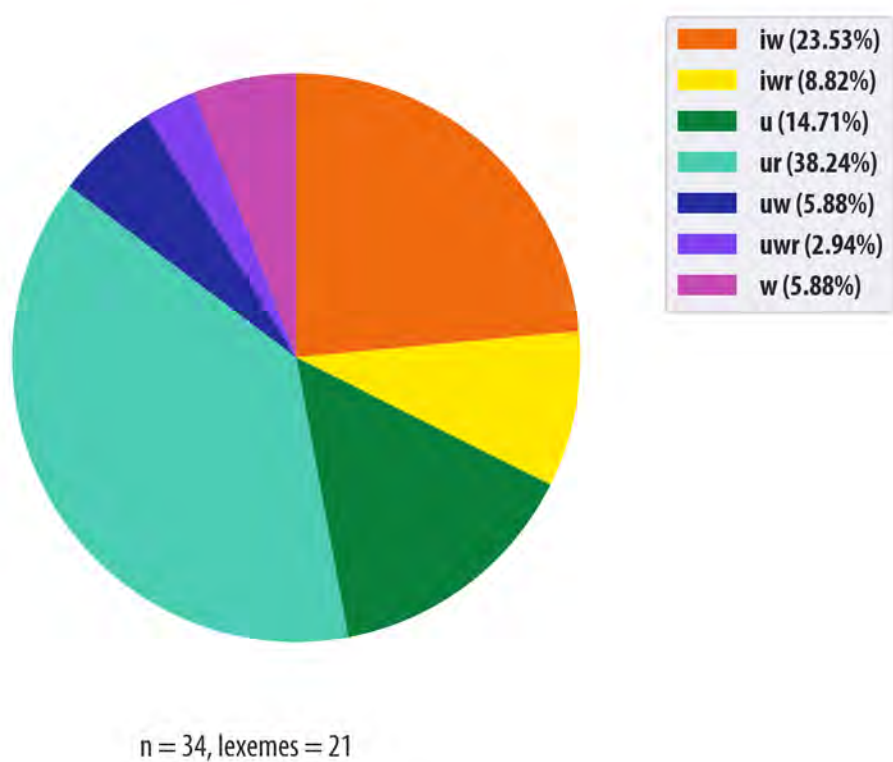
(c) Twitter (personal) data



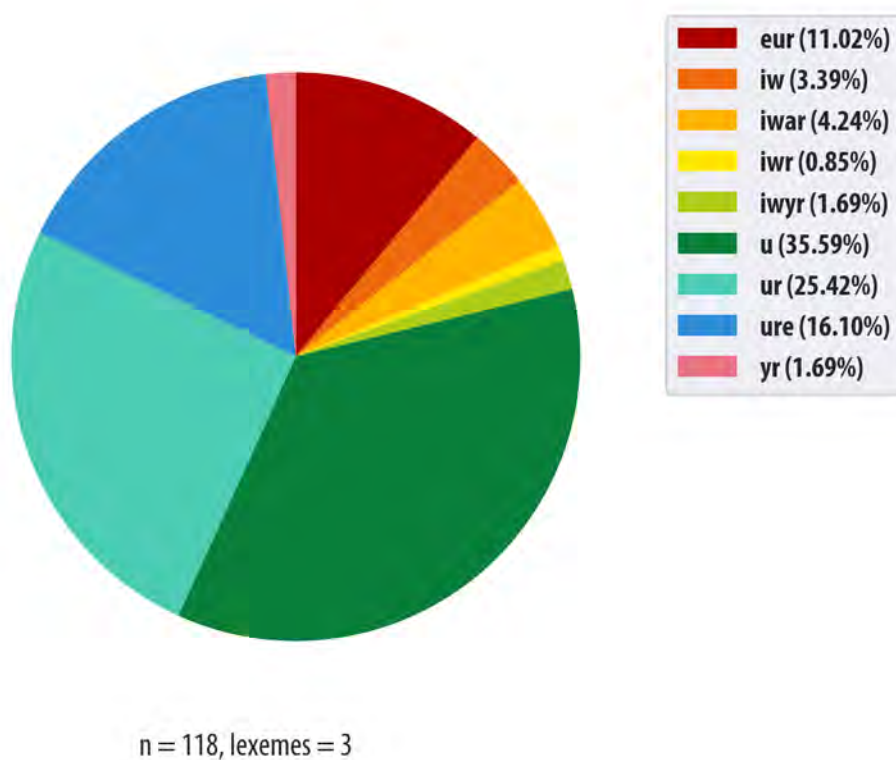
n = 4582, lexemes = 23

(d) Twitter (official) data

Figure C.21: Distribution of graphemes for phoneme /jɥu/ (TUNE) (cont.)

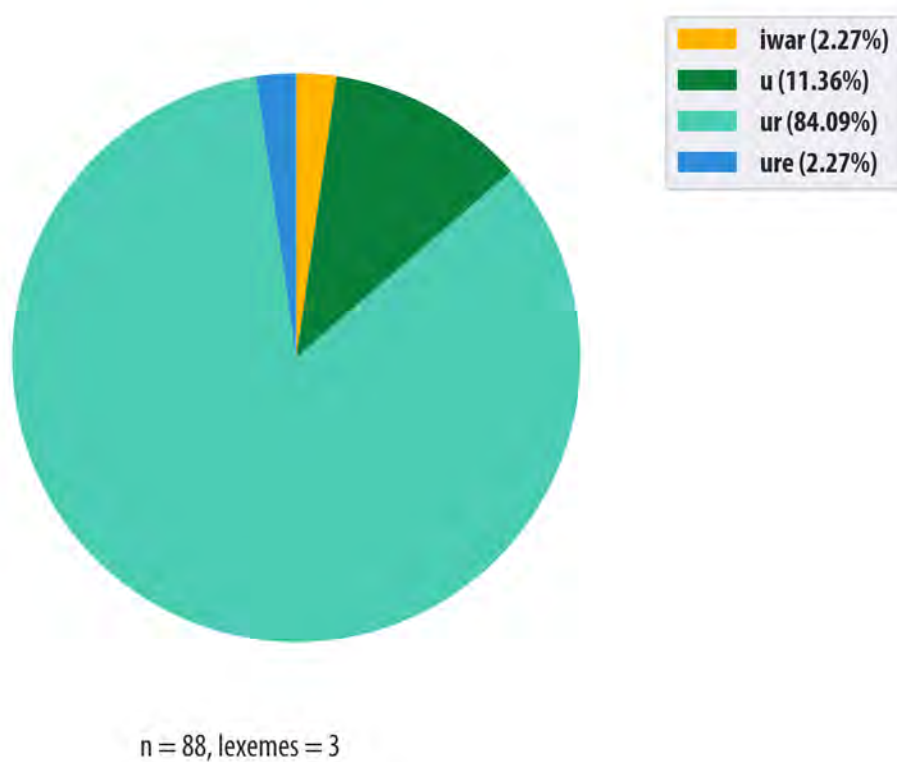


(a) GPC data



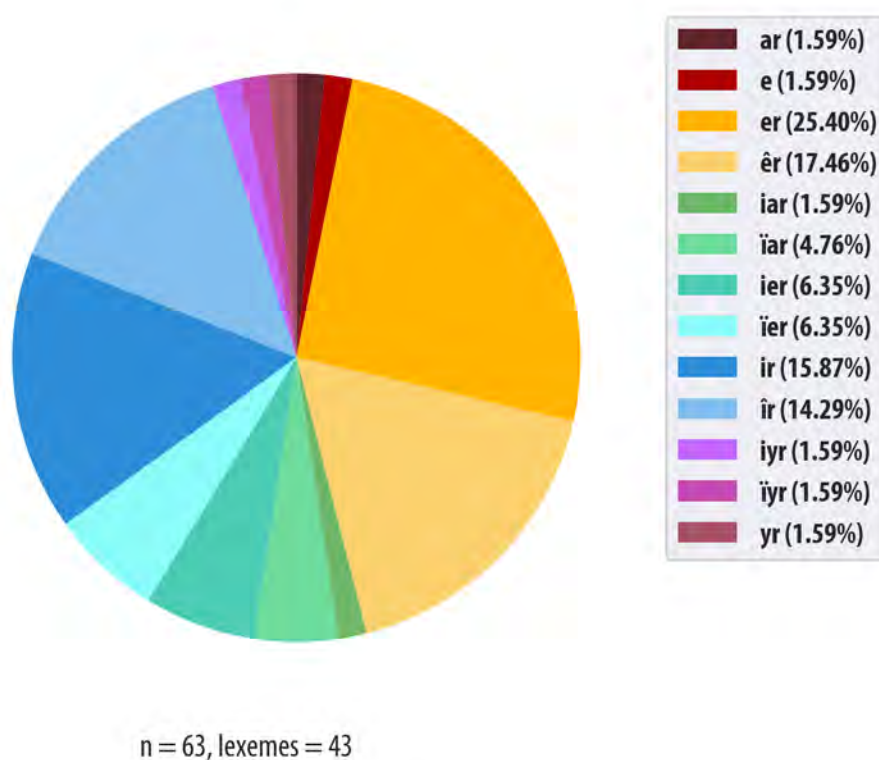
(b) : Twitter (personal) data

Figure C.22 Distribution of graphemes for phoneme /jʊə/ (CURE)

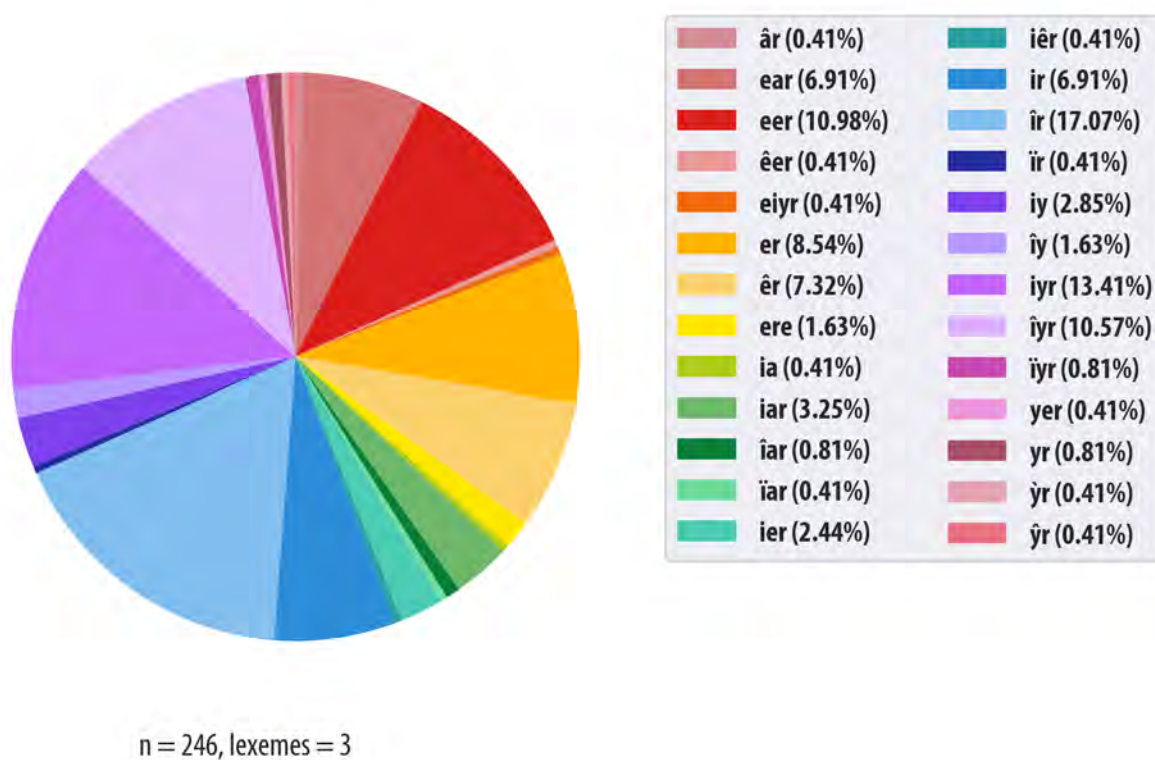


(c) Twitter (official) data

Figure C.22: Distribution of graphemes for phoneme /jʊə/ (**CURE**) (cont.)

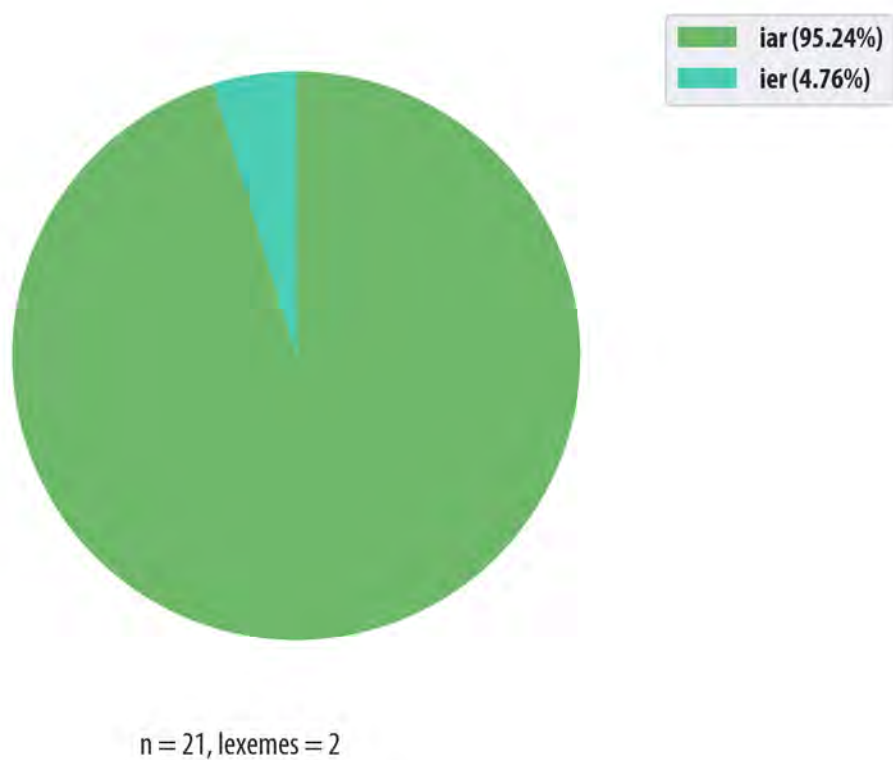


(a) GPC data

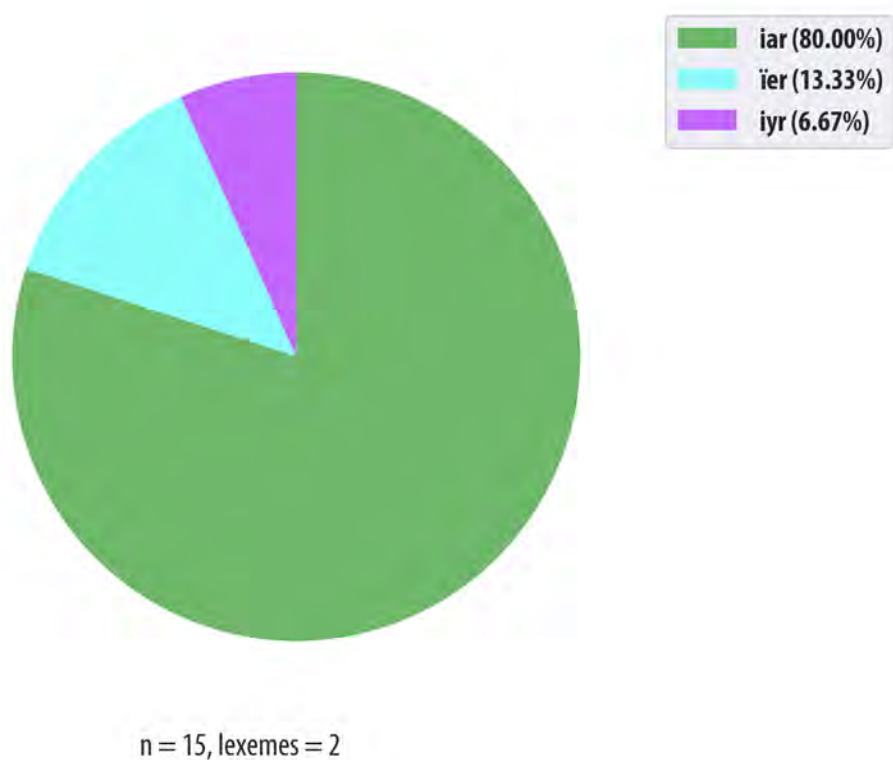


(b) Survey data

Figure C.23: Distribution of graphemes for phoneme /ɪə/ (NEAR~BEER)

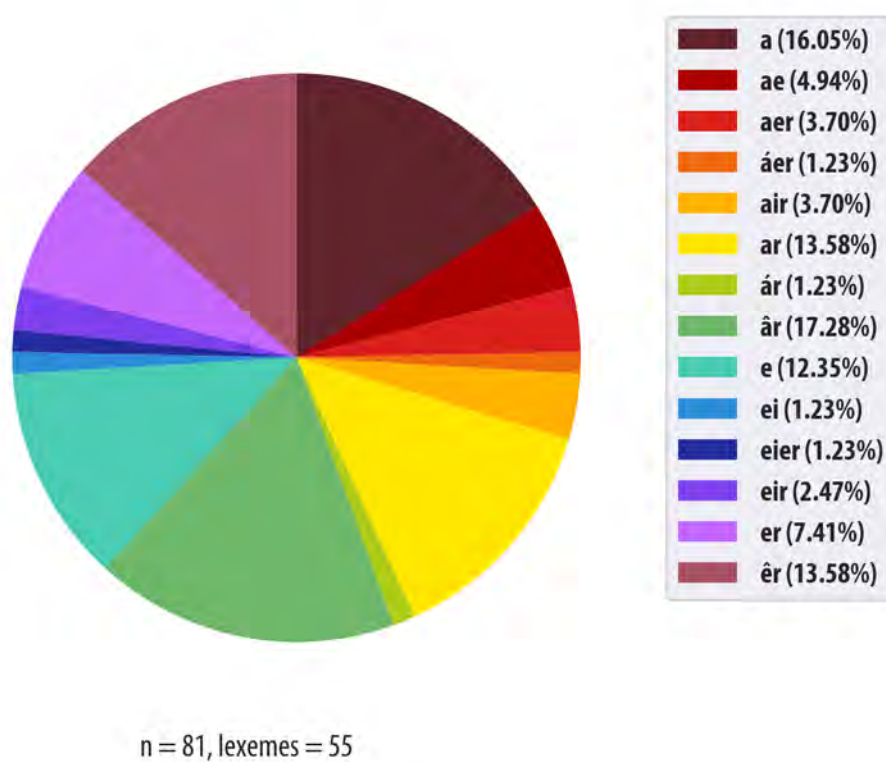


(c) Twitter (personal) data

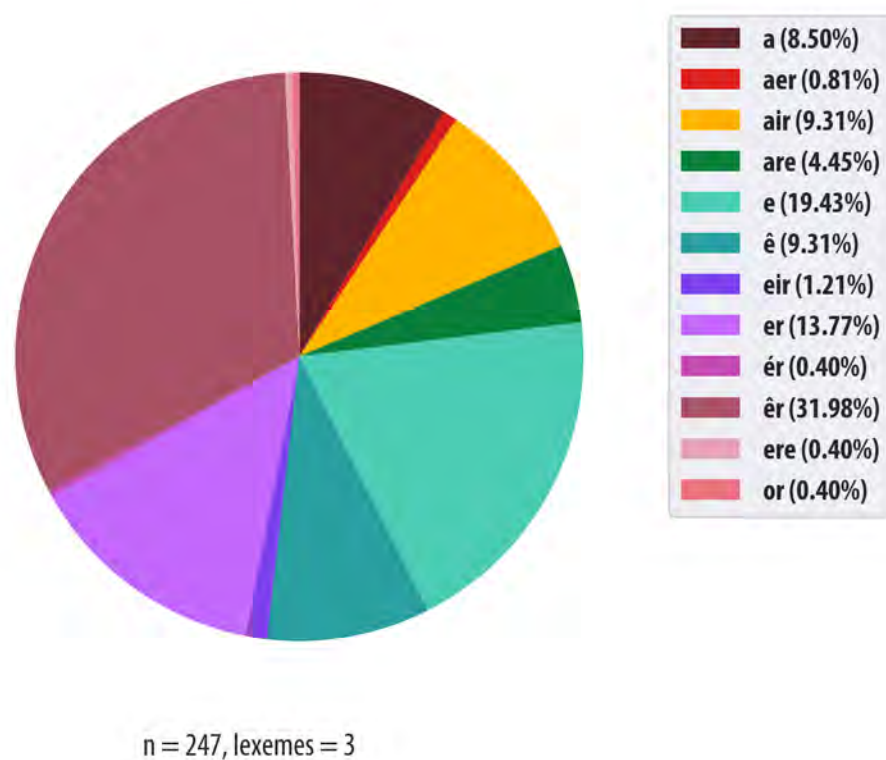


(d) Twitter (official) data

Figure C.23: Distribution of graphemes for phoneme /ɪə/ (**NEAR**~**BEER**) (cont.)

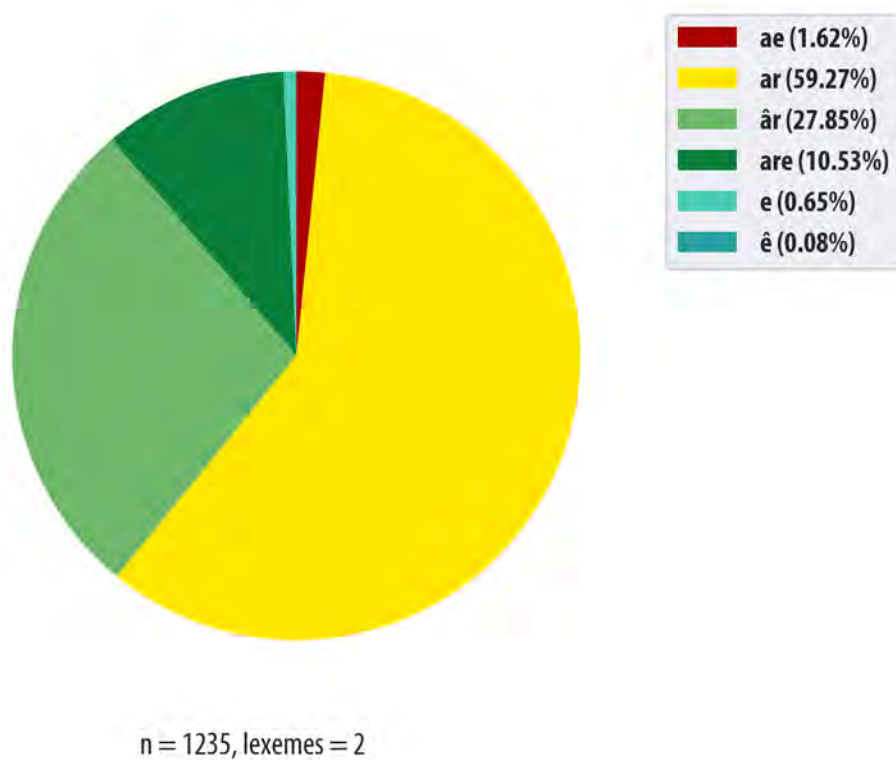


(a) GPC data

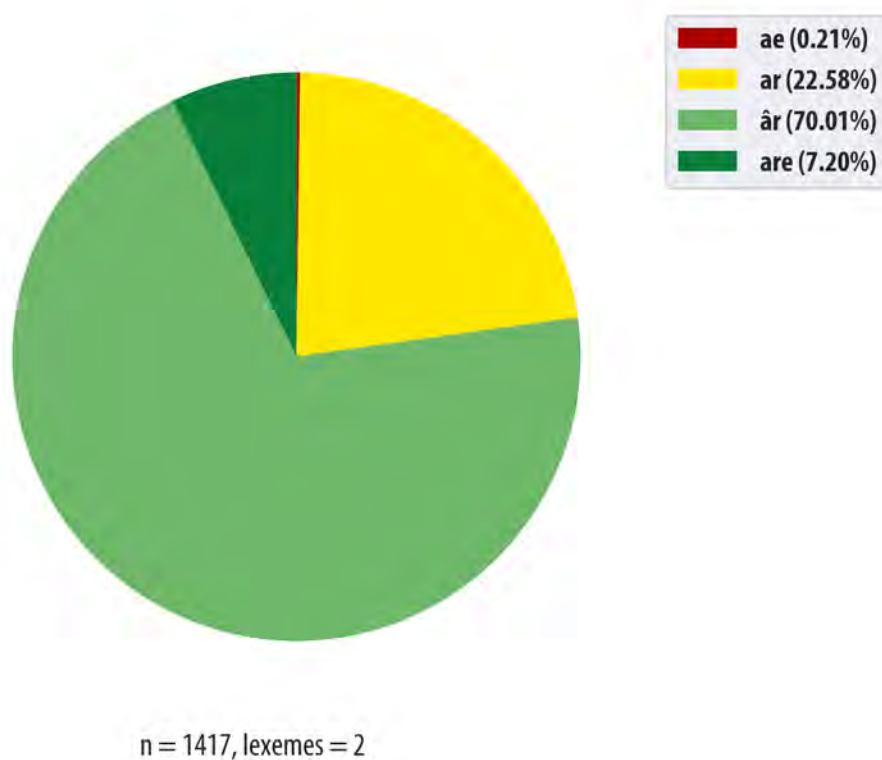


(b) Survey data

Figure C.24: Distribution of graphemes for phoneme /εθ/ (**SQUARE~LAIR**)

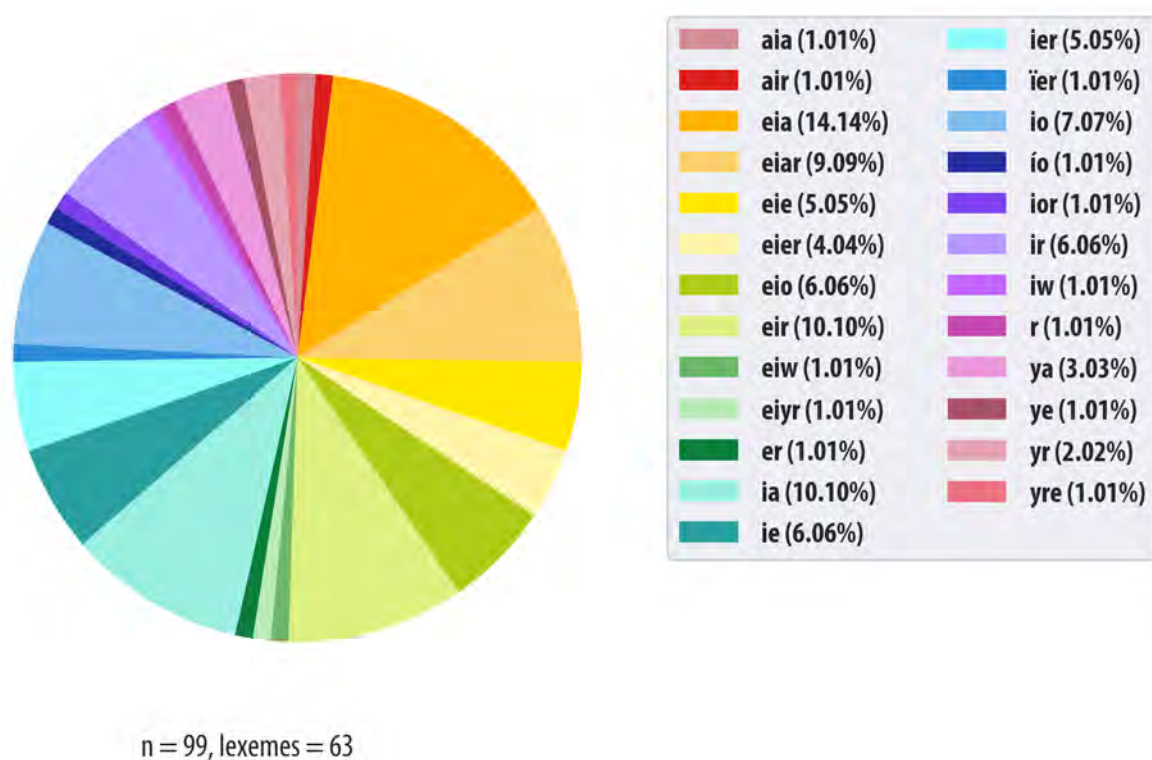


(c) Twitter (personal) data

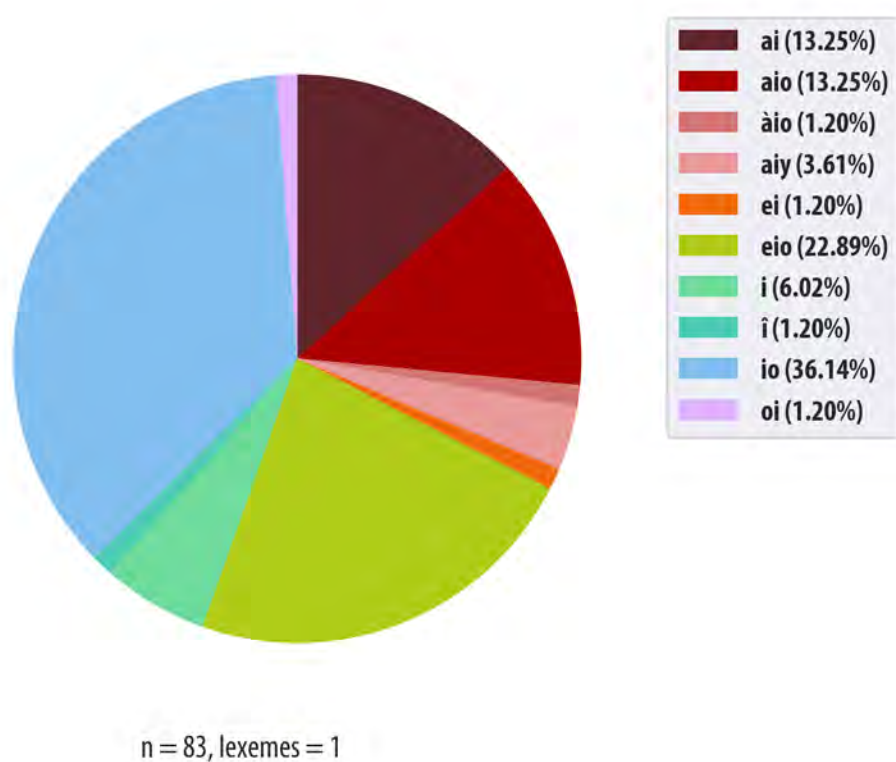


(d) Twitter (official) data

Figure C.24: Distribution of graphemes for phoneme /εθ/ (**SQUARE~LAIR**) (cont.)

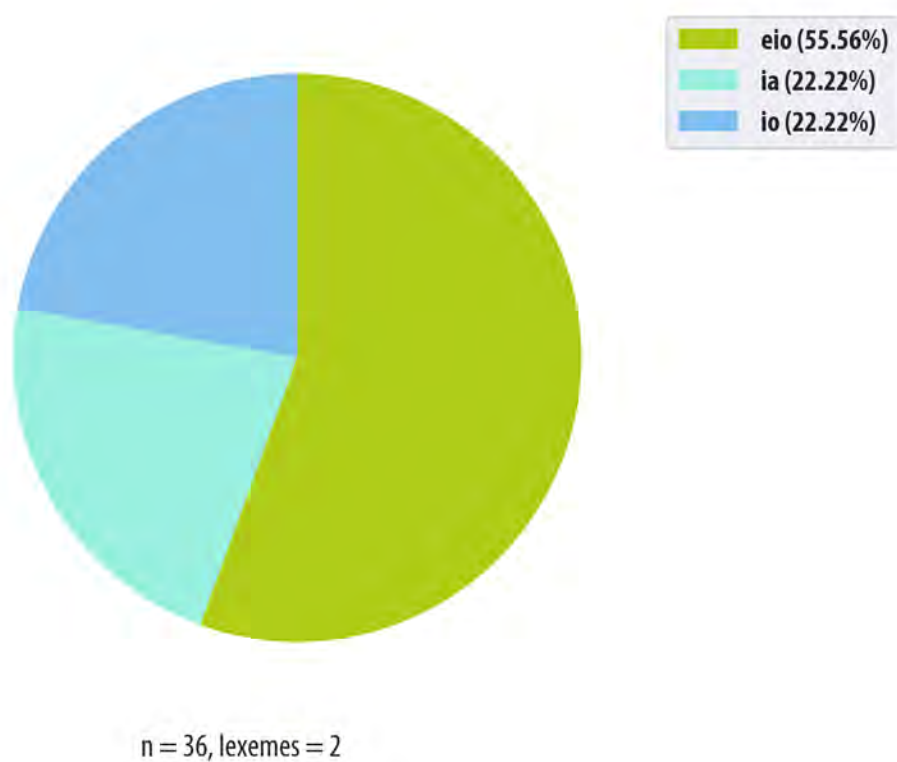


(a) GPC data

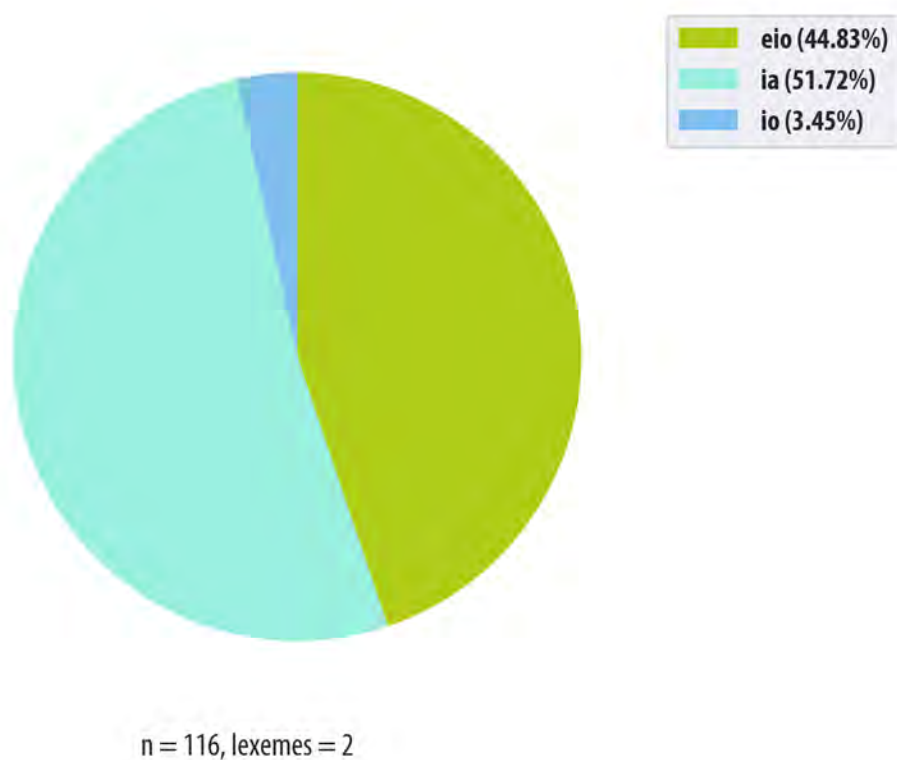


(b) Survey data

Figure C.25: Distribution of graphemes for phoneme /aə/ (**FIRE**)

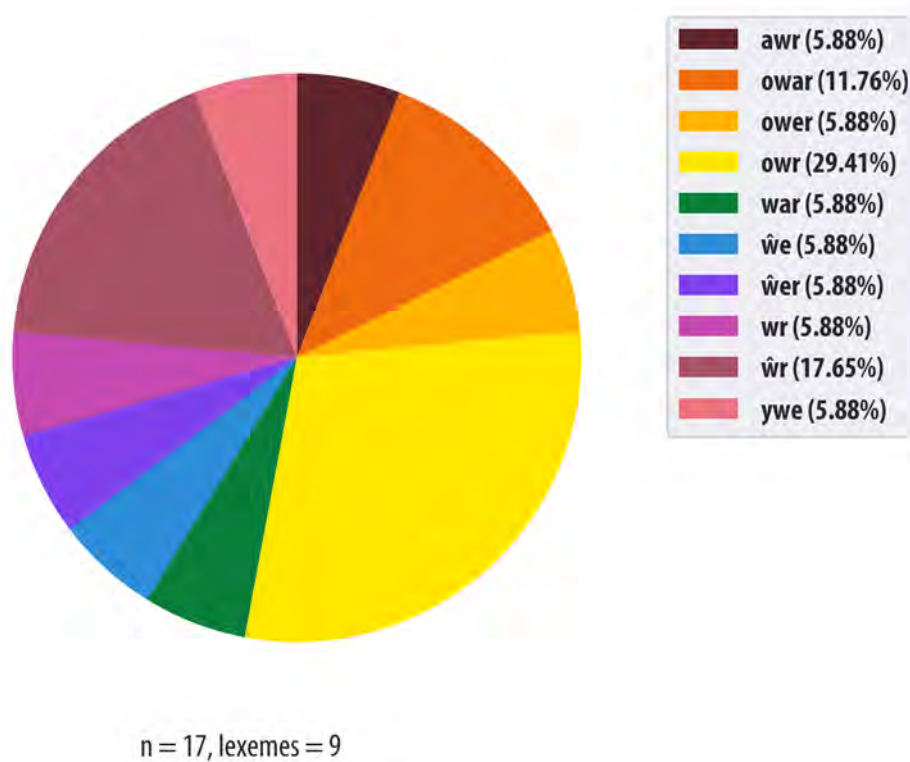


(c) Twitter (personal) data

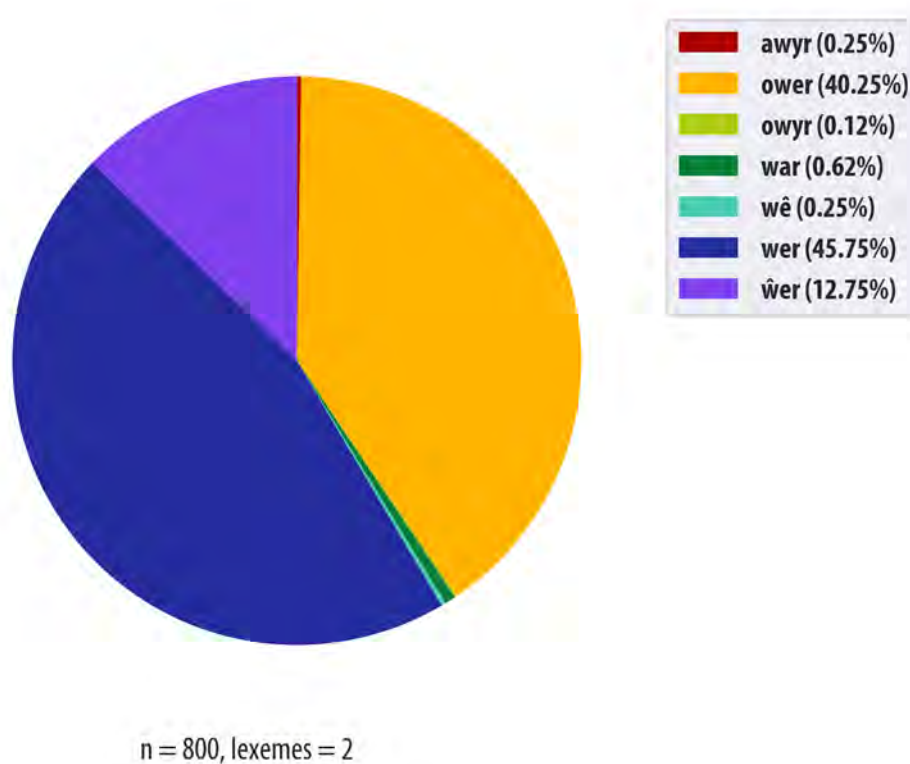


(d) Twitter (official) data

Figure C.25: Distribution of graphemes for phoneme /aə/ (**FIRE**) (cont.)

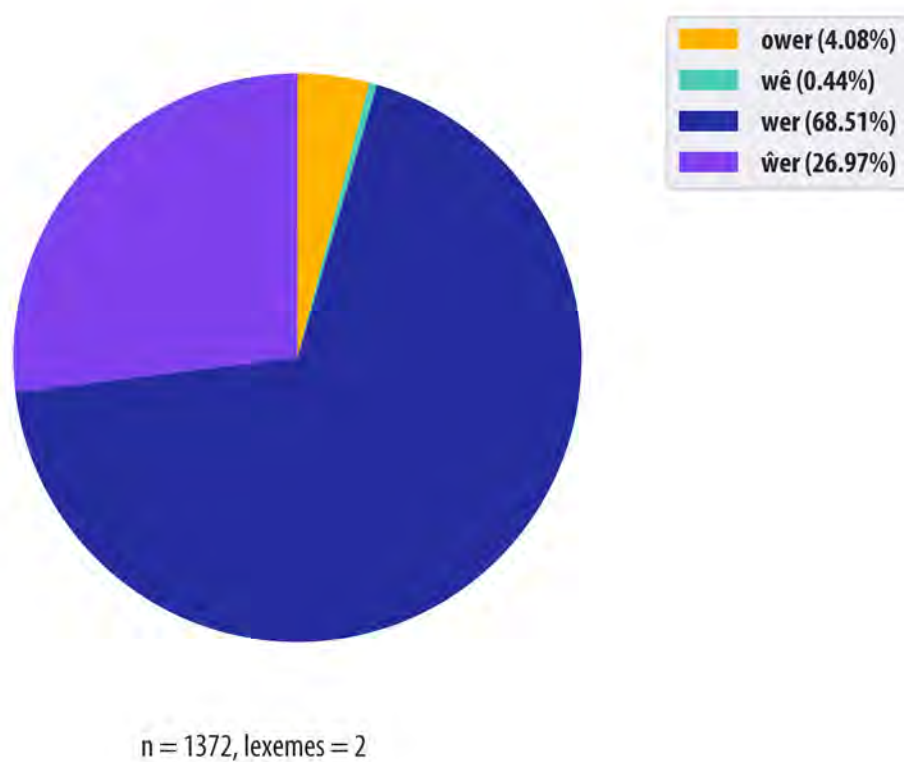


(a) GPC data



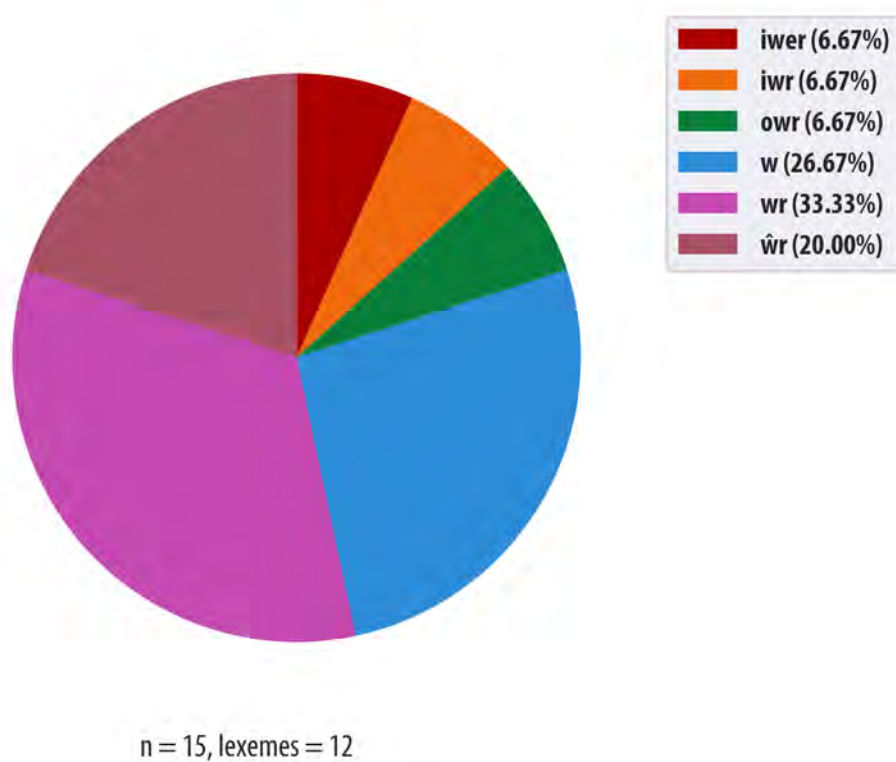
(b) Twitter (personal) data

Figure C.26: Distribution of graphemes for phoneme /ɑə/ (**TOWER**)

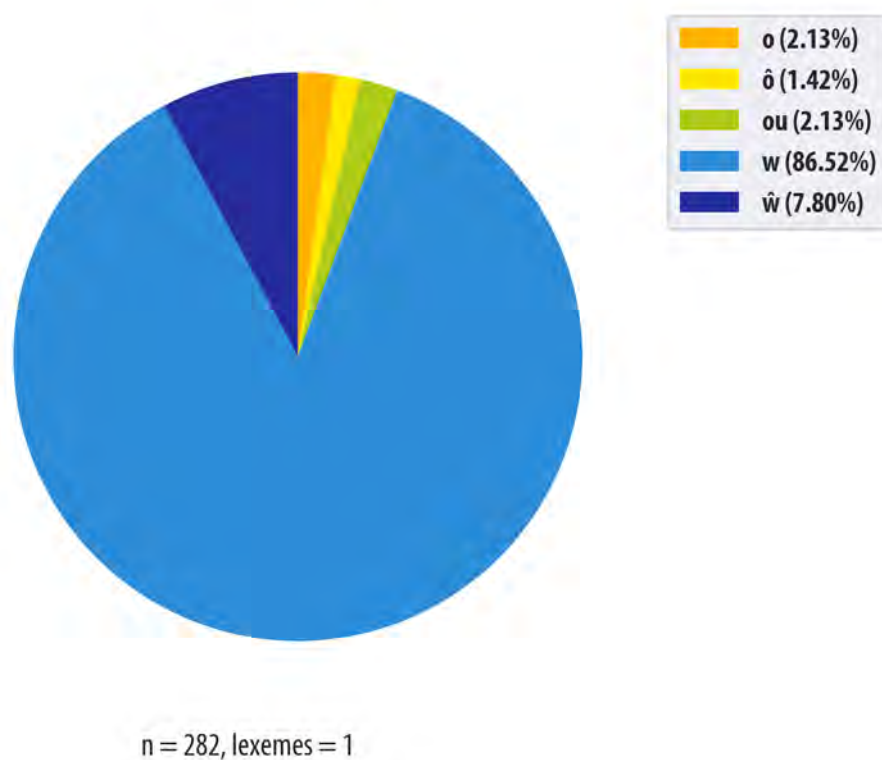


(c) Twitter (official) data

Figure C.26: Distribution of graphemes for phoneme /ɑə/ (**TOWER**) (cont.)

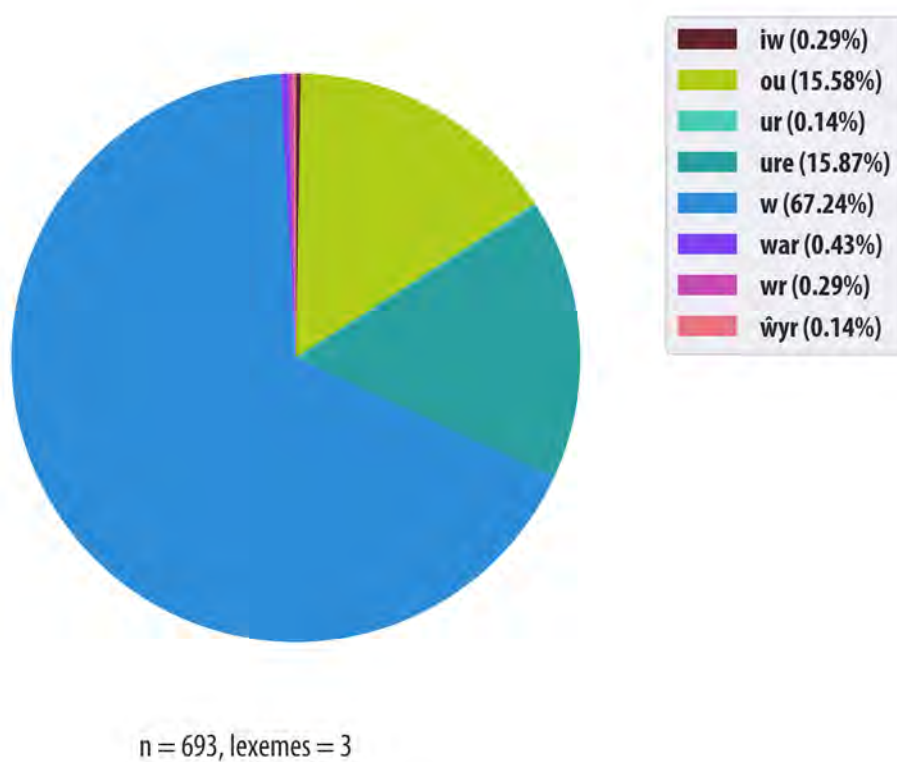


(a) GPC data

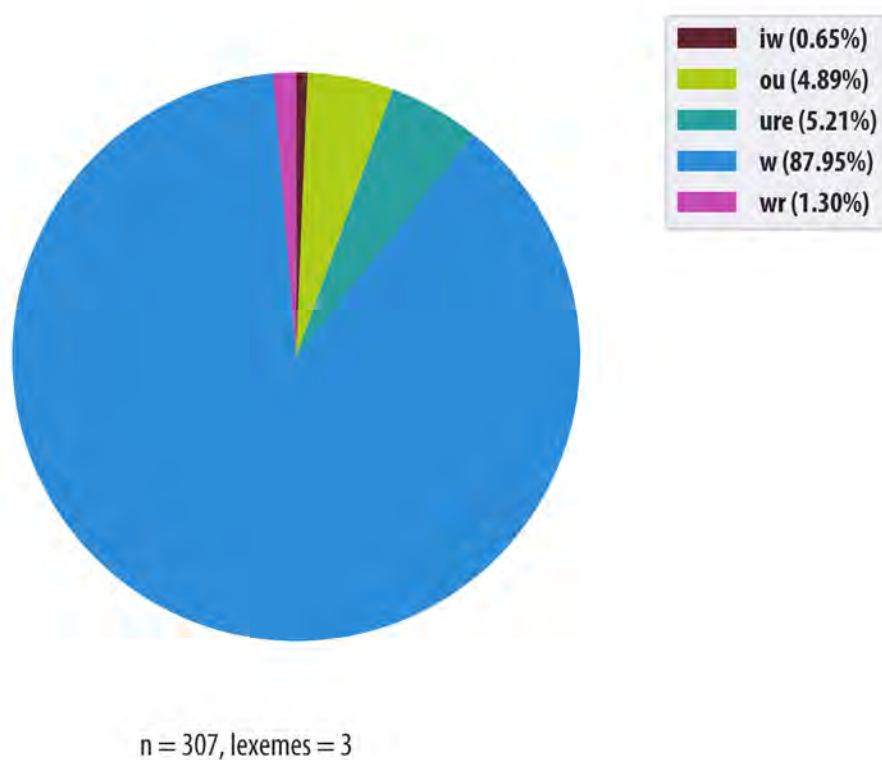


(b) Survey data

Figure C.27: Distribution of graphemes for phoneme /ʊə/ (**POOR**~**BREWER**)

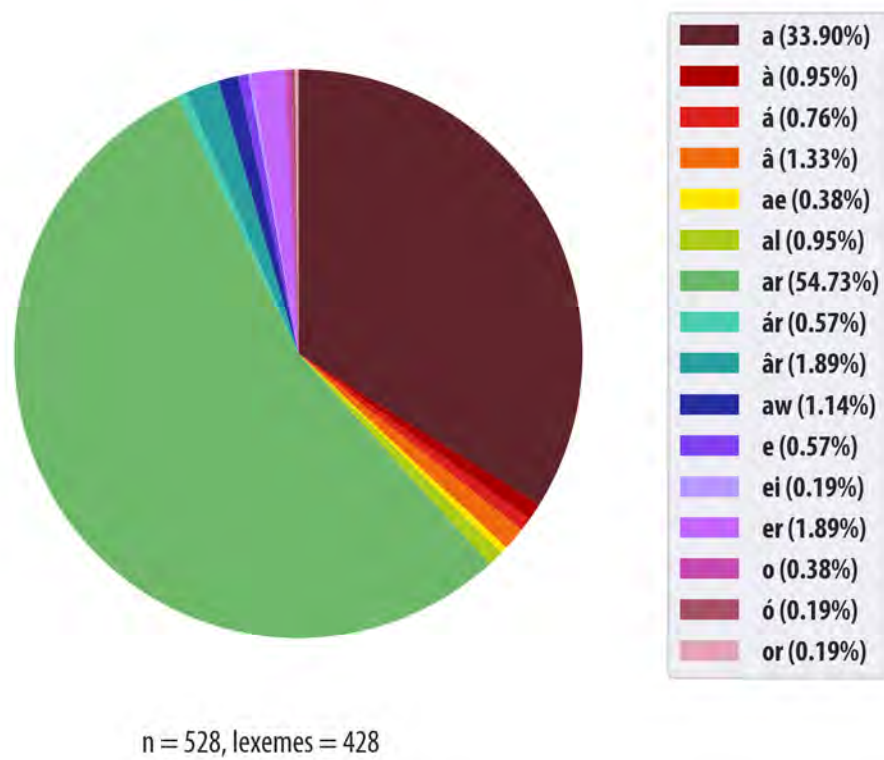


(c) Twitter (personal) data

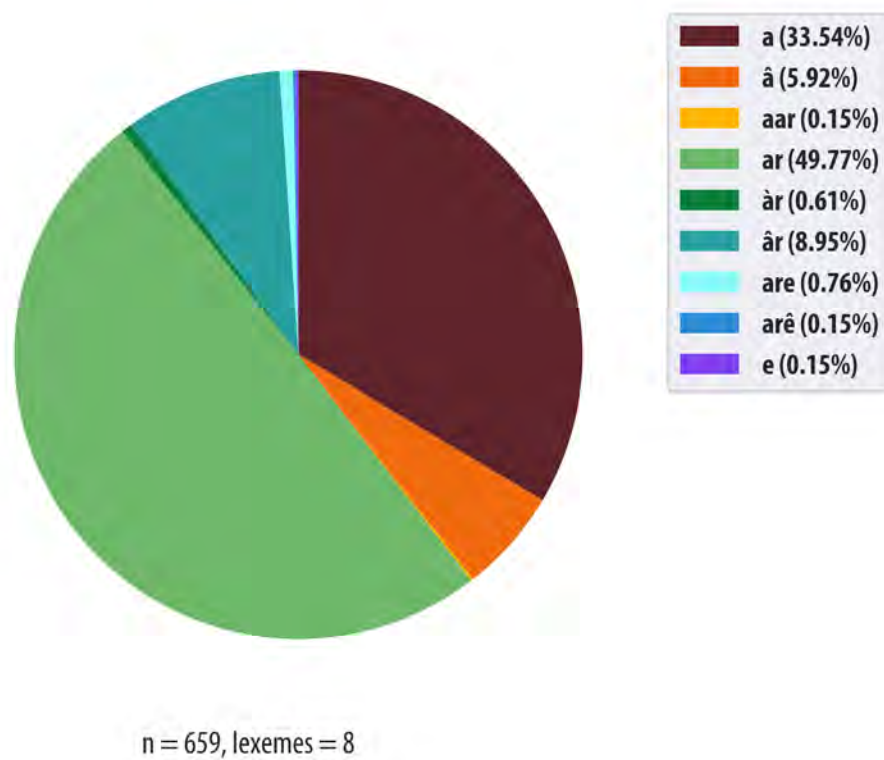


(d) Twitter (official) data

Figure C.27: Distribution of graphemes for phoneme /ʊə/ (**POOR~BREWER**) (cont.)

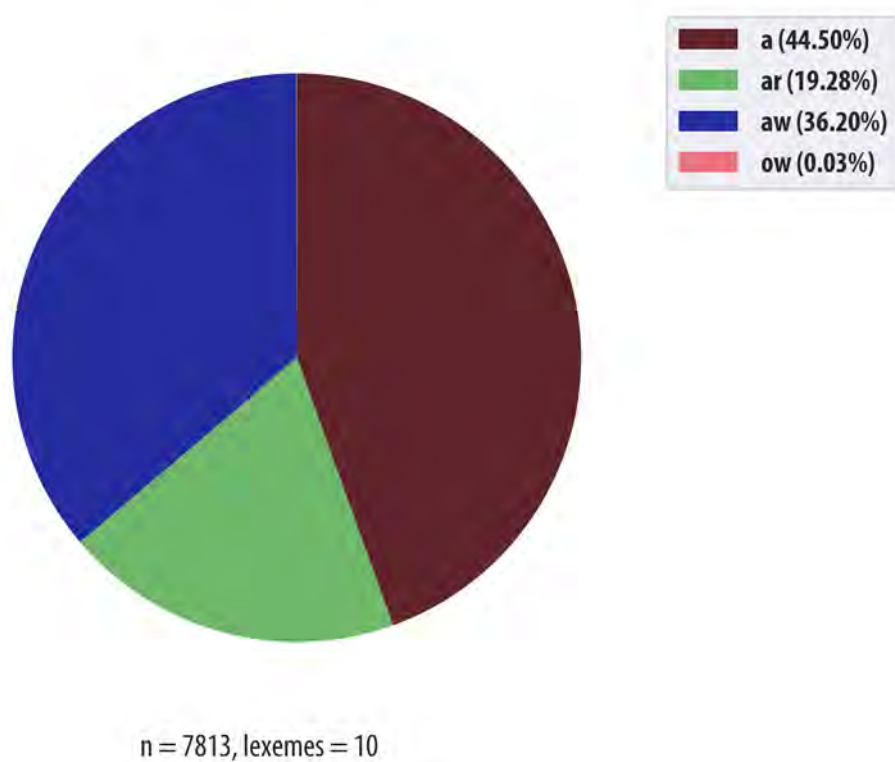


(a) GPC data

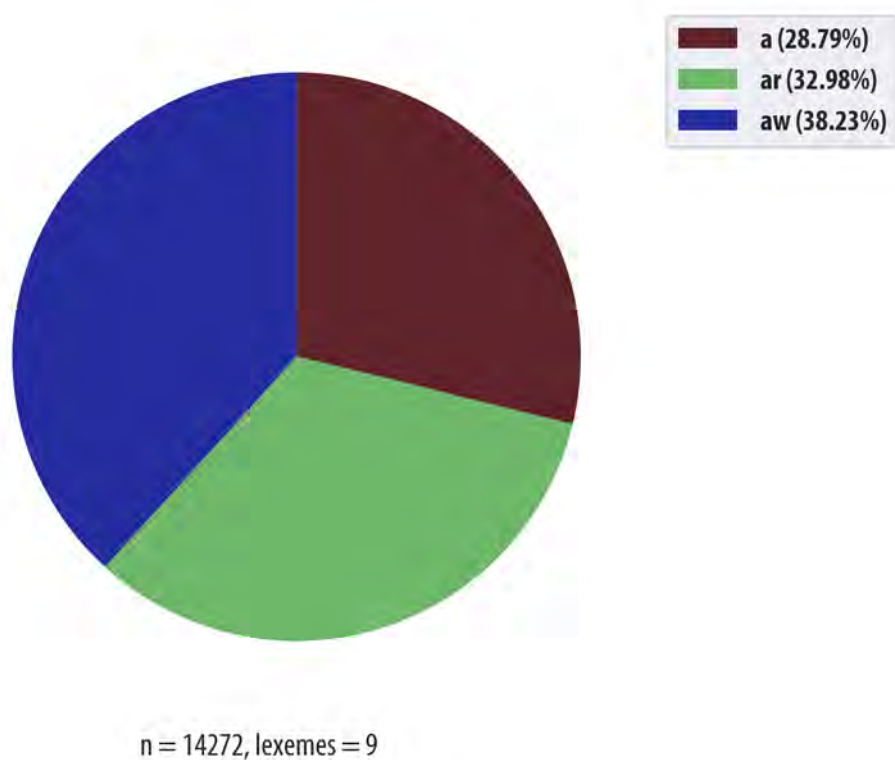


(b) Survey data

Figure C.28: Distribution of graphemes for phoneme /ɑ:/ (**BATH**~**PALM**~**START**)

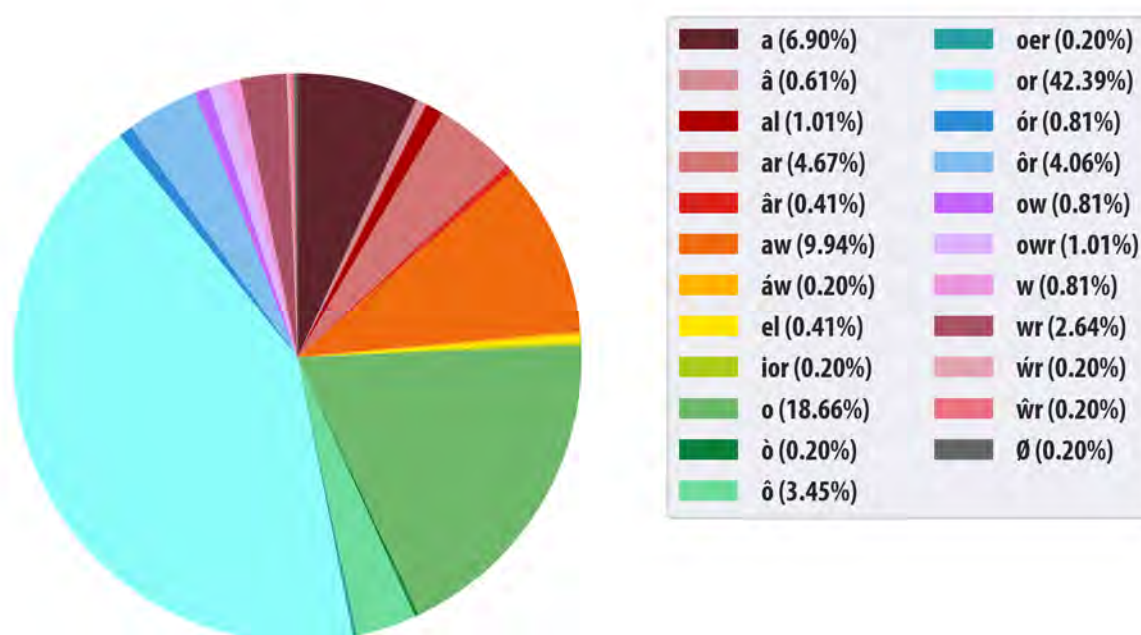


(c) Twitter (personal) data



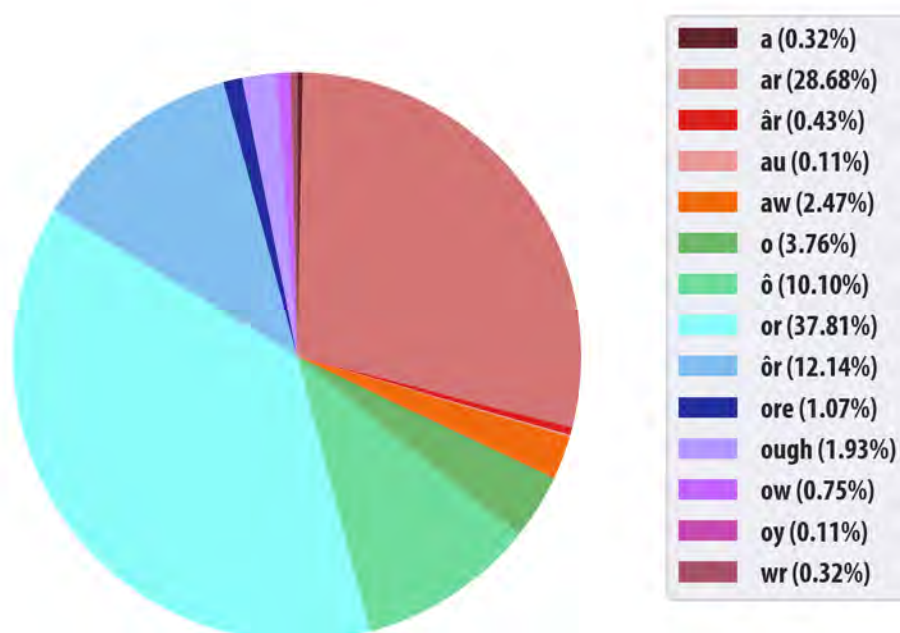
(d) Twitter (official) data

Figure C.28: Distribution of graphemes for phoneme /ɑ:/ (**BATH**~**PALM**~**START**) (cont.)



n = 493, lexemes = 358

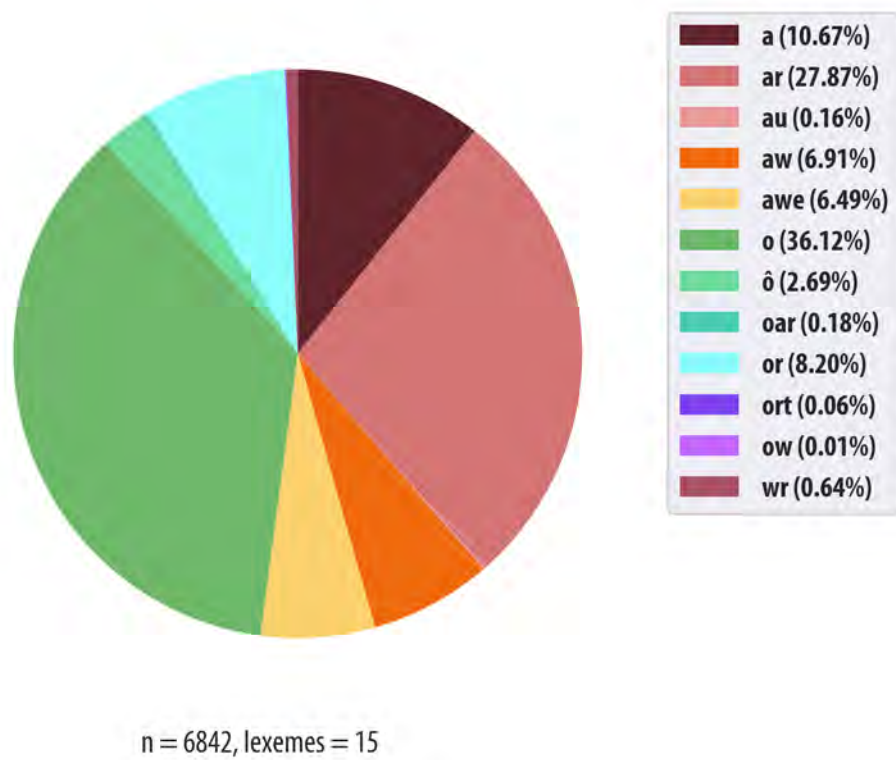
(a) GPC data



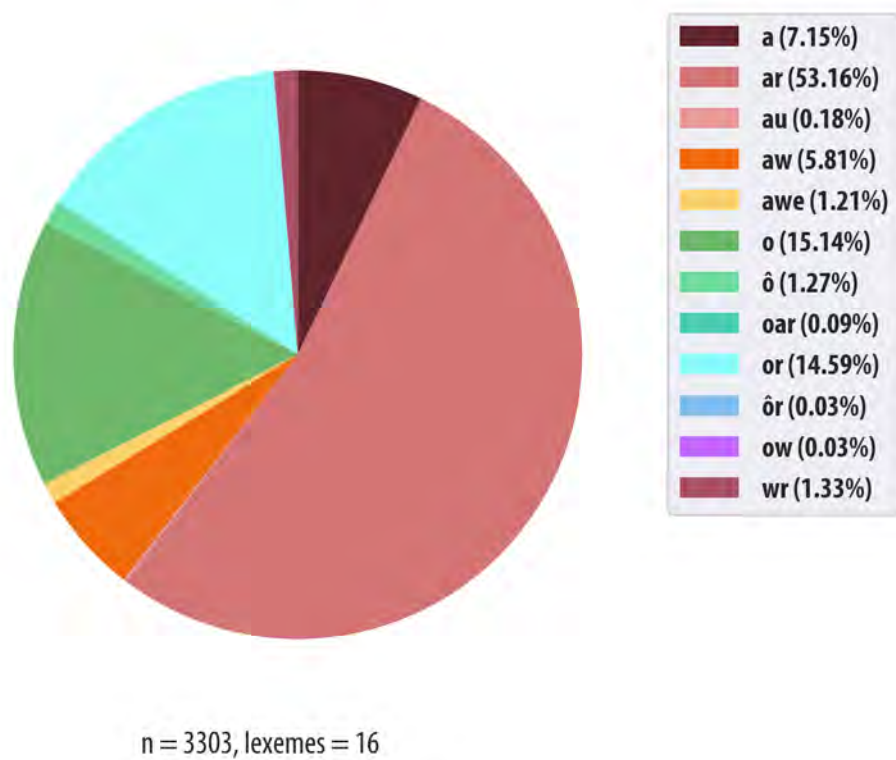
n = 931, lexemes = 8

(b) Survey data

Figure C.29: Distribution of graphemes for phoneme /o:/
(THOUGHT~NORTH~FORCE)

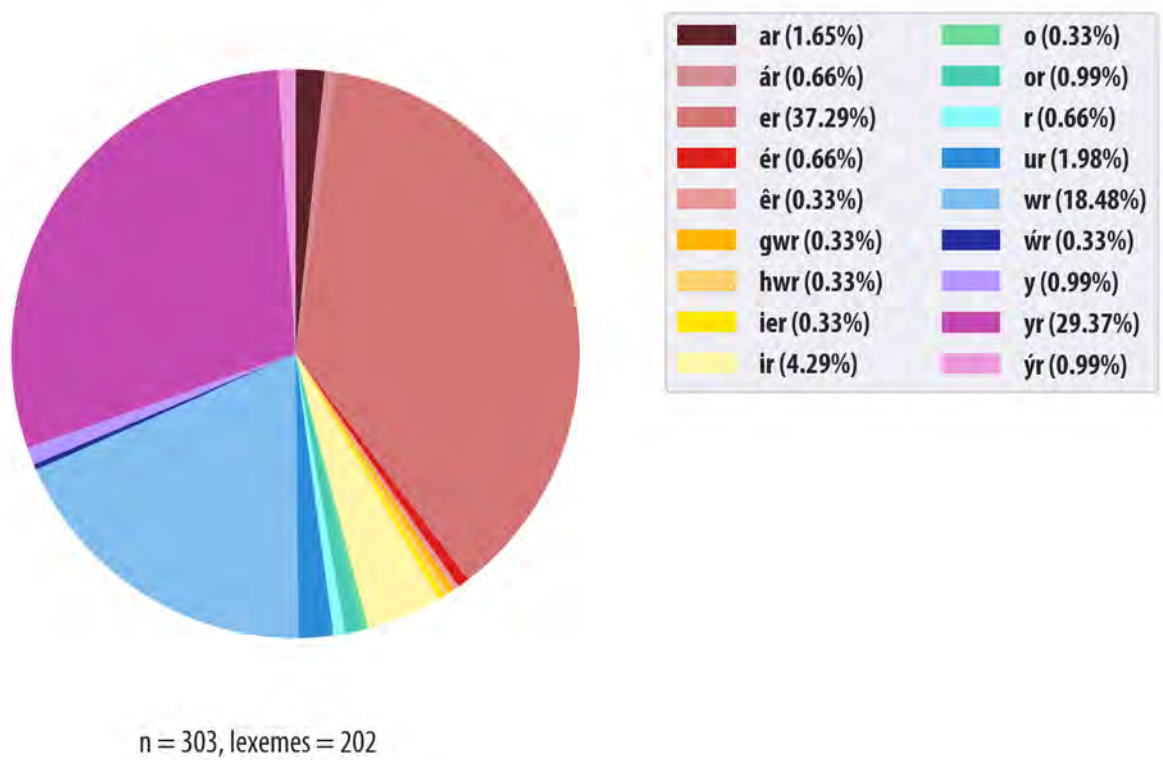


(c) Twitter (personal) data

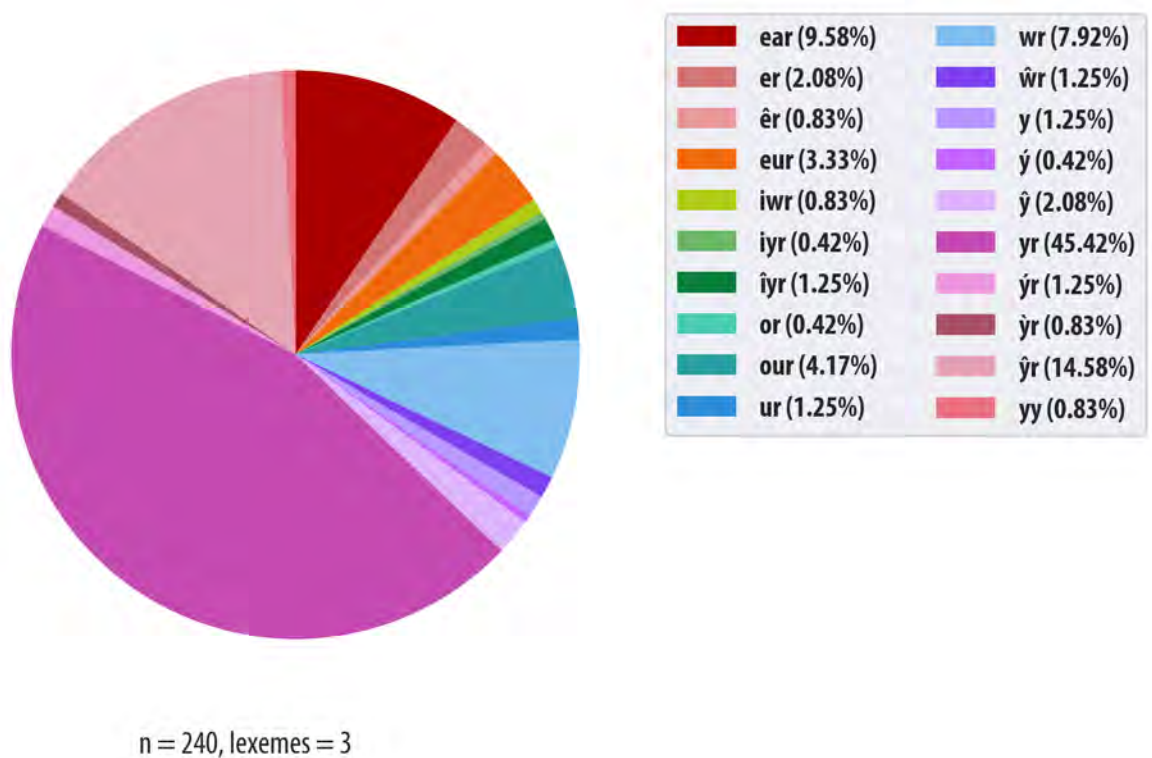


(d) Twitter (official) data

Figure C.29: Distribution of graphemes for phoneme /o:/
(**THOUGHT**~**NORTH**~**FORCE**) (cont.)

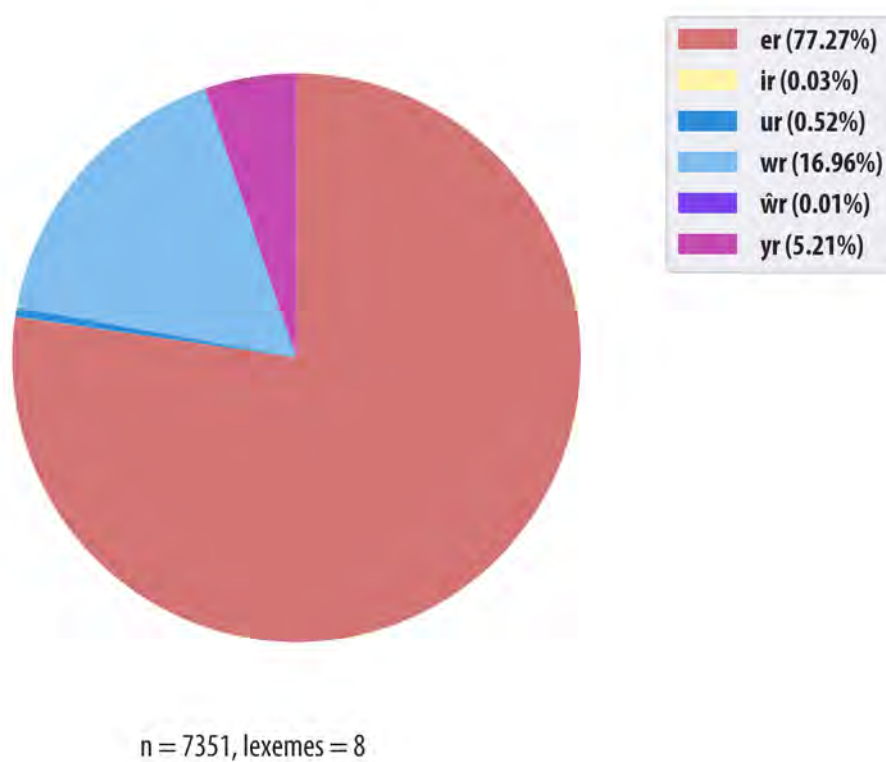


(a) GPC data

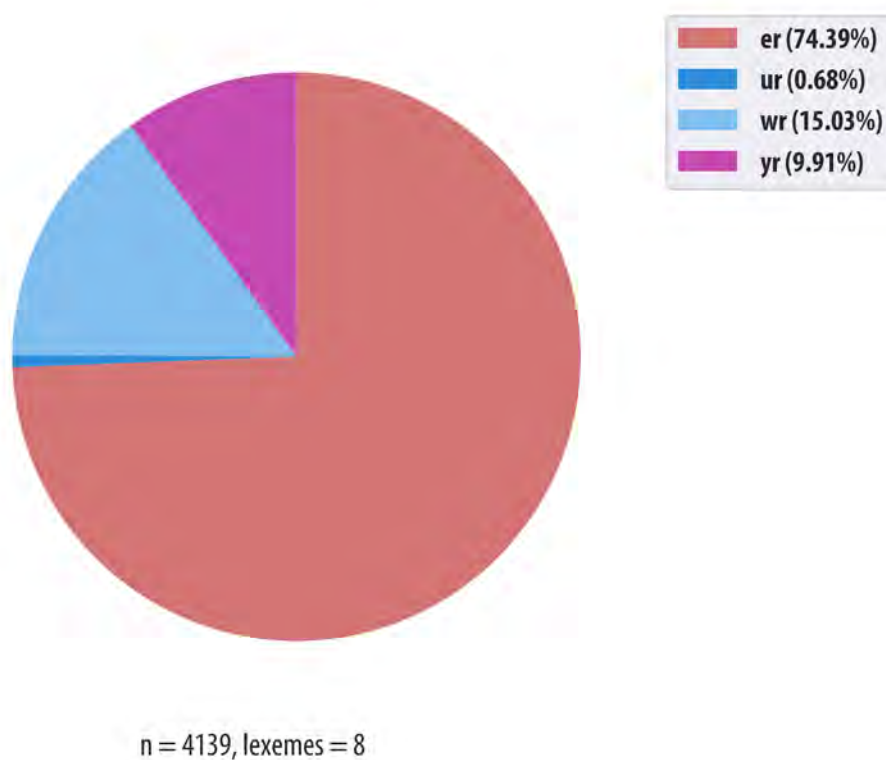


(b) Survey data

Figure C.30: Distribution of graphemes for phoneme /3:/ (**NURSE**)

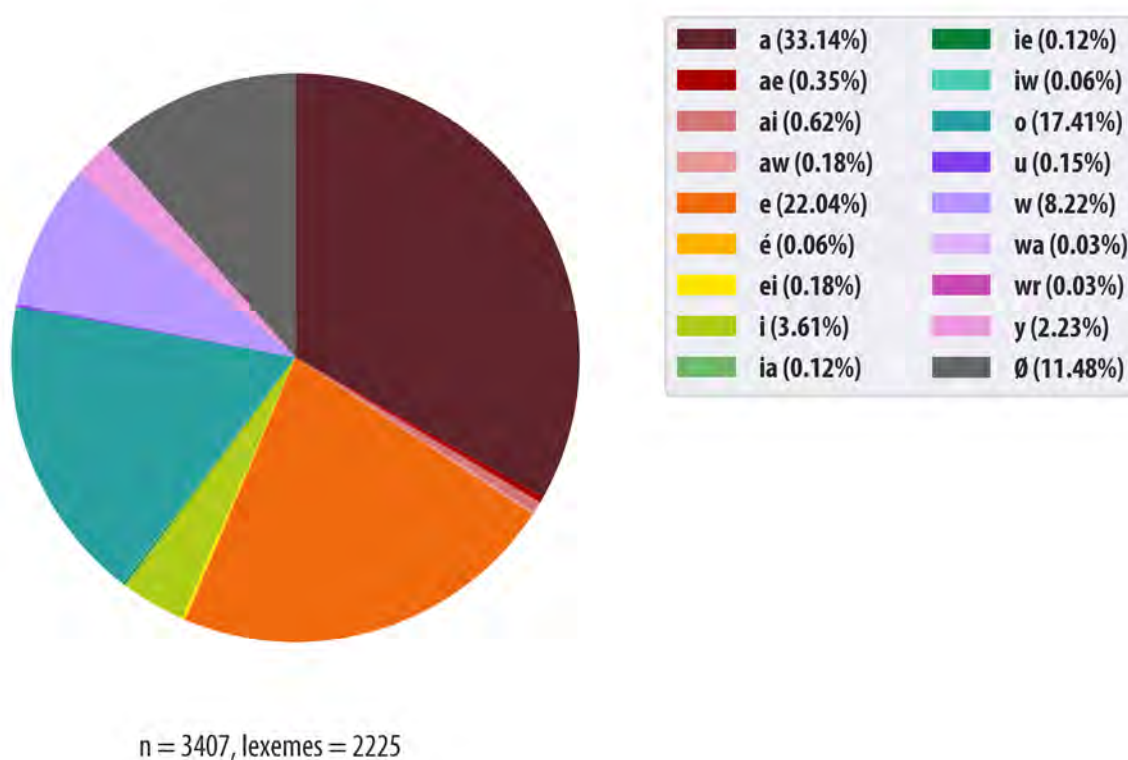


(c) Twitter (personal) data

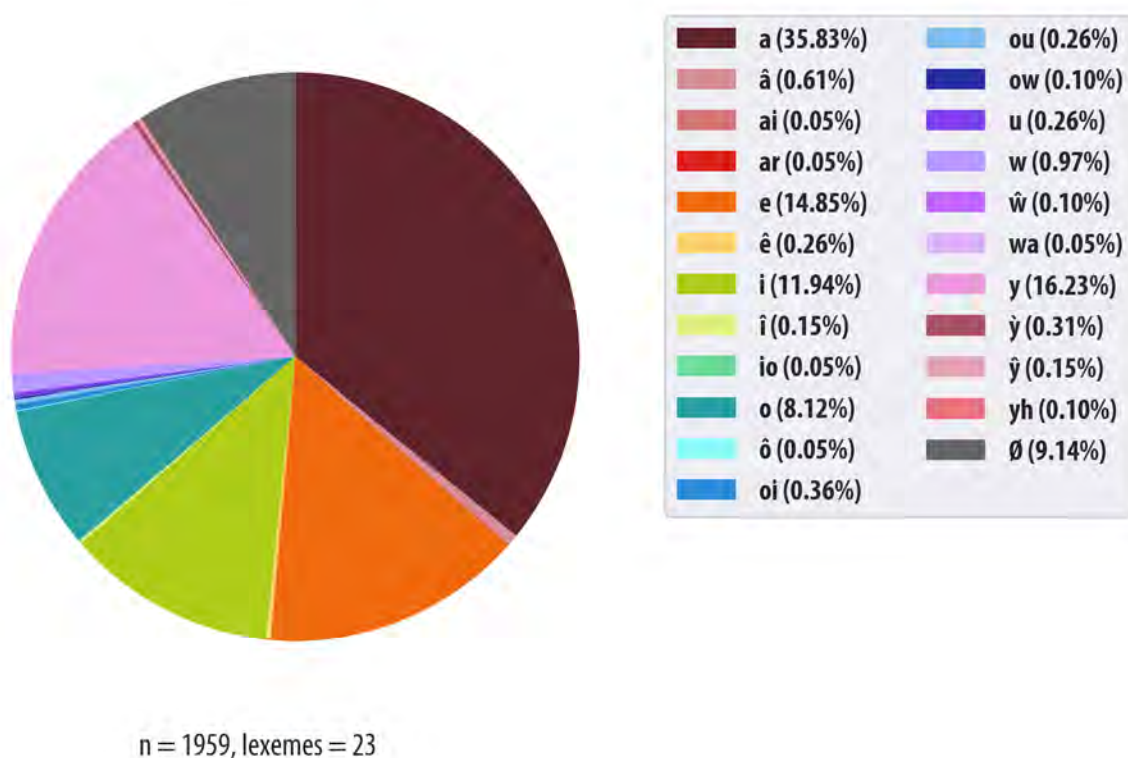


(d) Twitter (official) data

Figure C.30: Distribution of graphemes for phoneme /ɜ:/ (**NURSE**) (cont.)

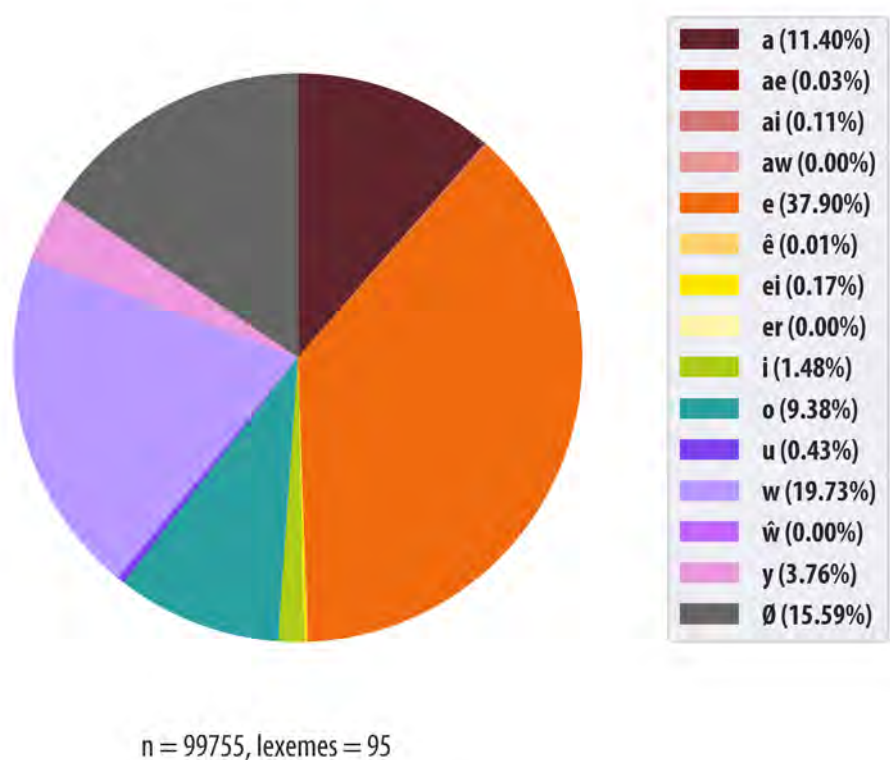


(a) GPC data

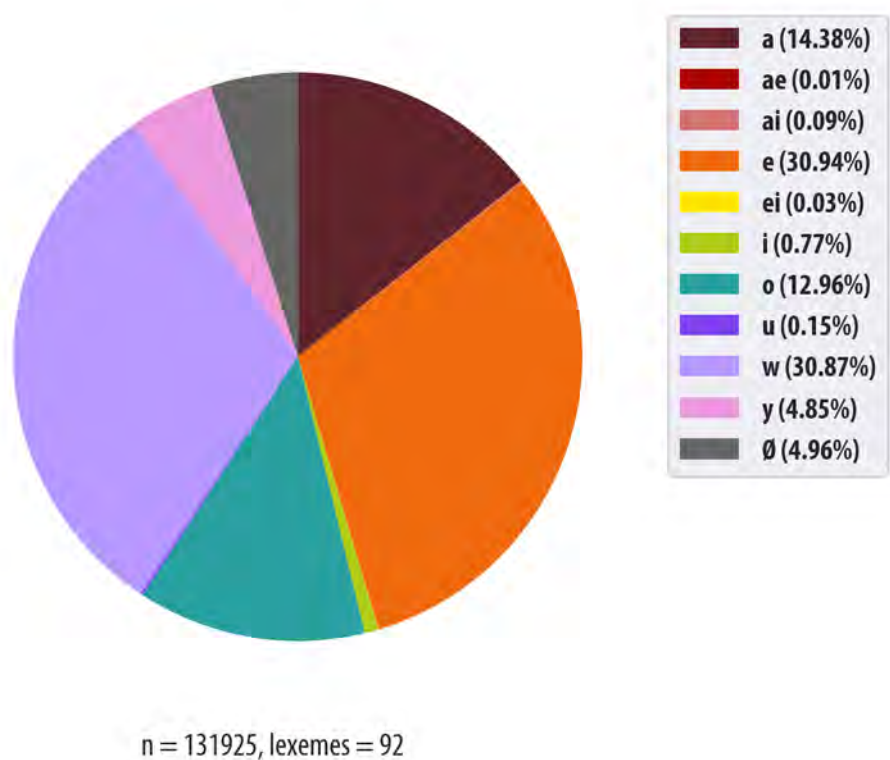


(b) Survey data

Figure C.31: Distribution of graphemes for phoneme /ə/ (COMMA)

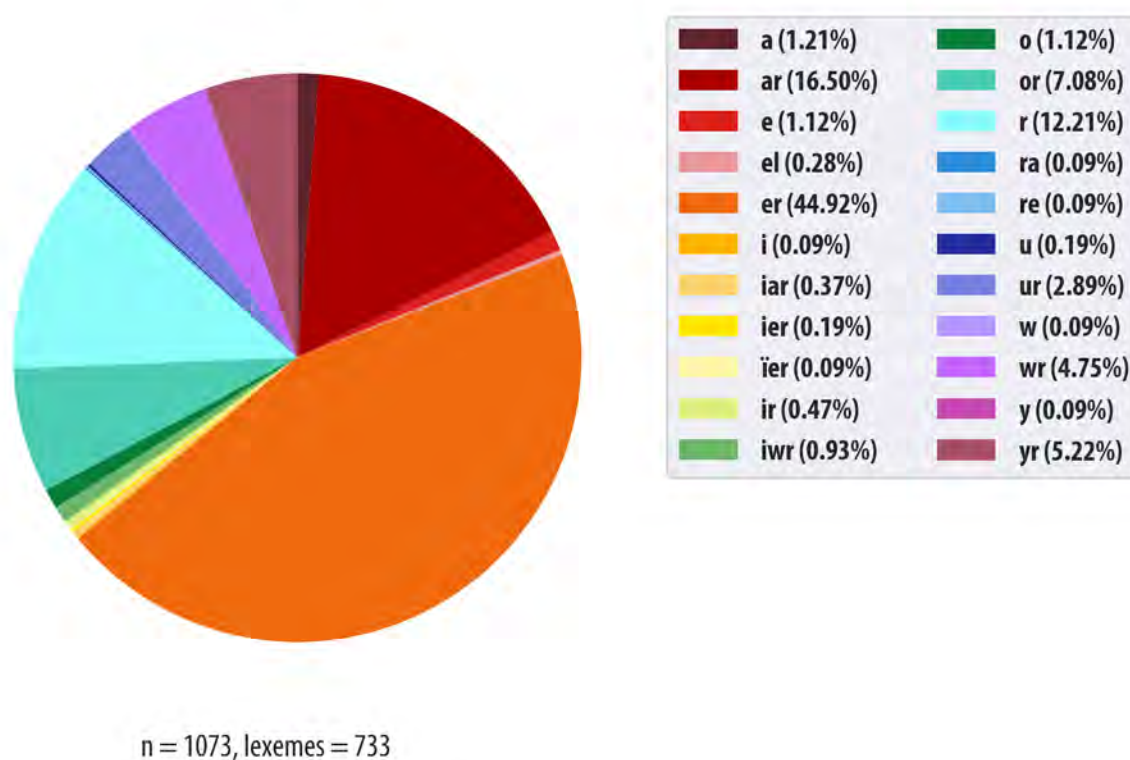


(c) Twitter (personal) data

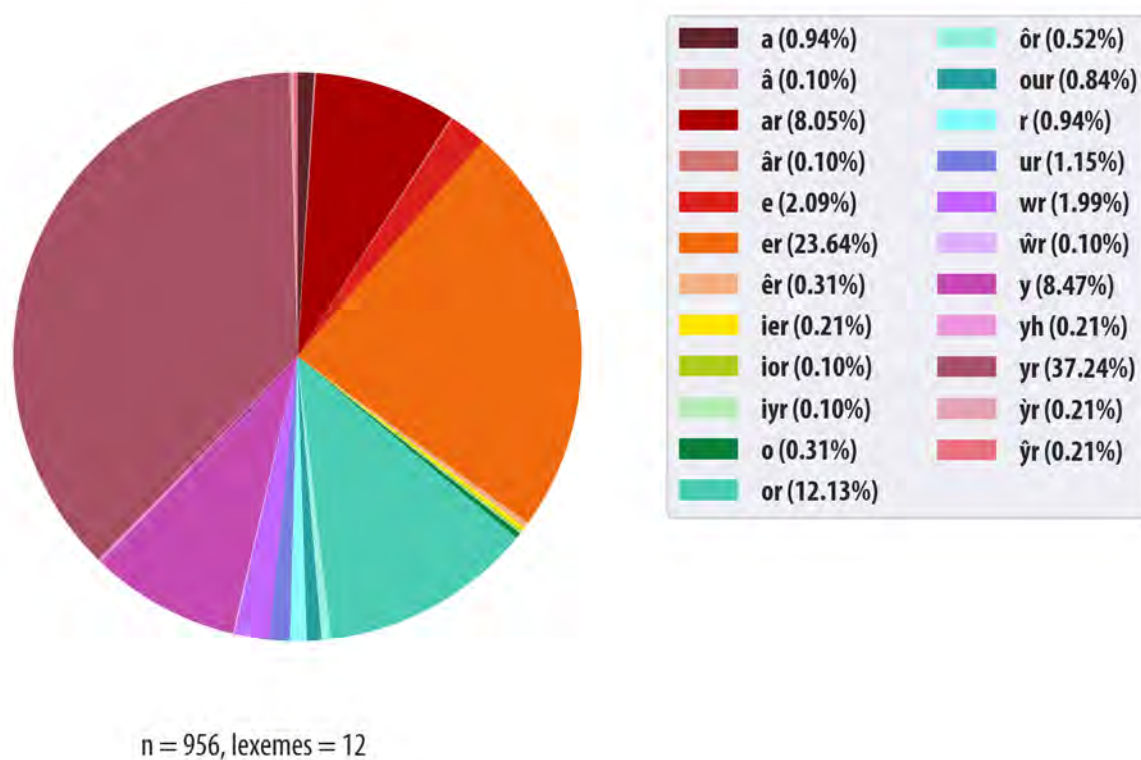


(d) Twitter (official) data

Figure C.31: Distribution of graphemes for phoneme /ə/ (COMMA) (cont.)

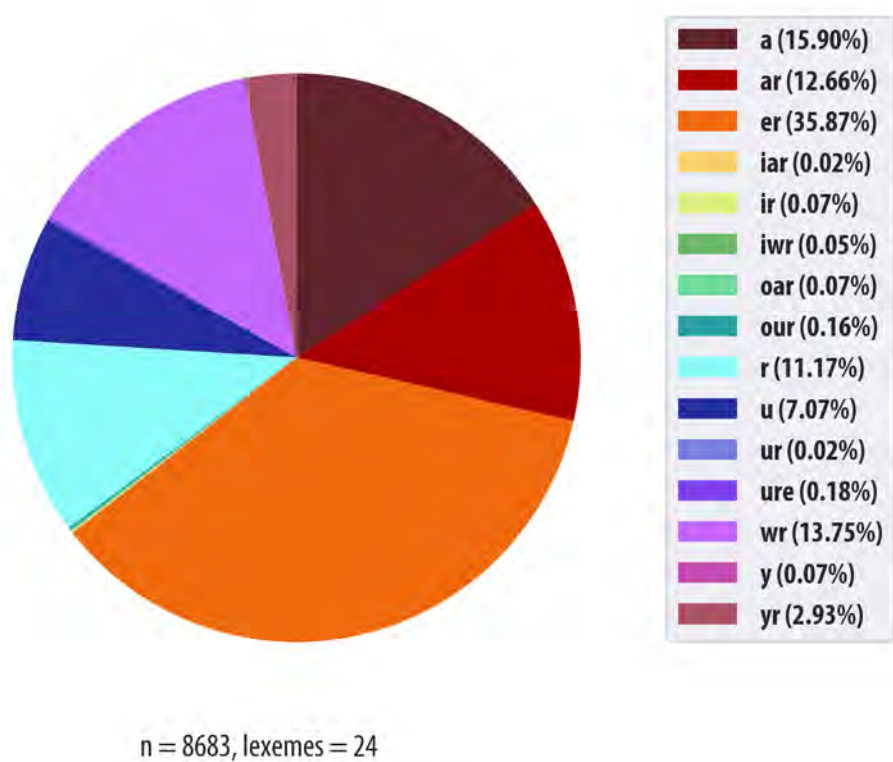


(a) GPC data

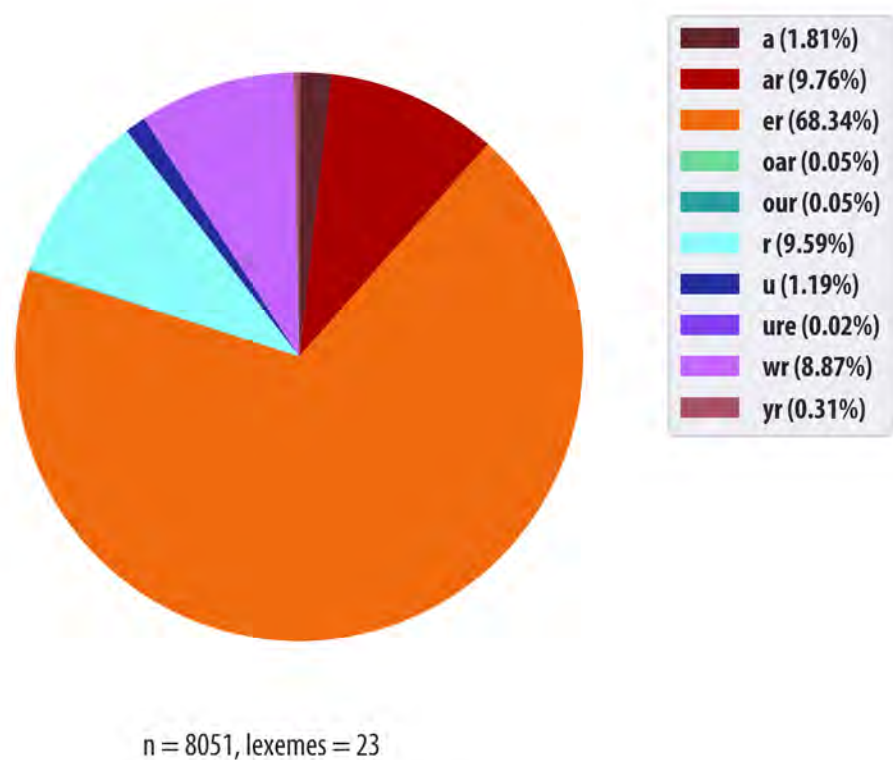


(b) Survey data

Figure C.32: Distribution of graphemes for phoneme /ə/ (**LETTER**)

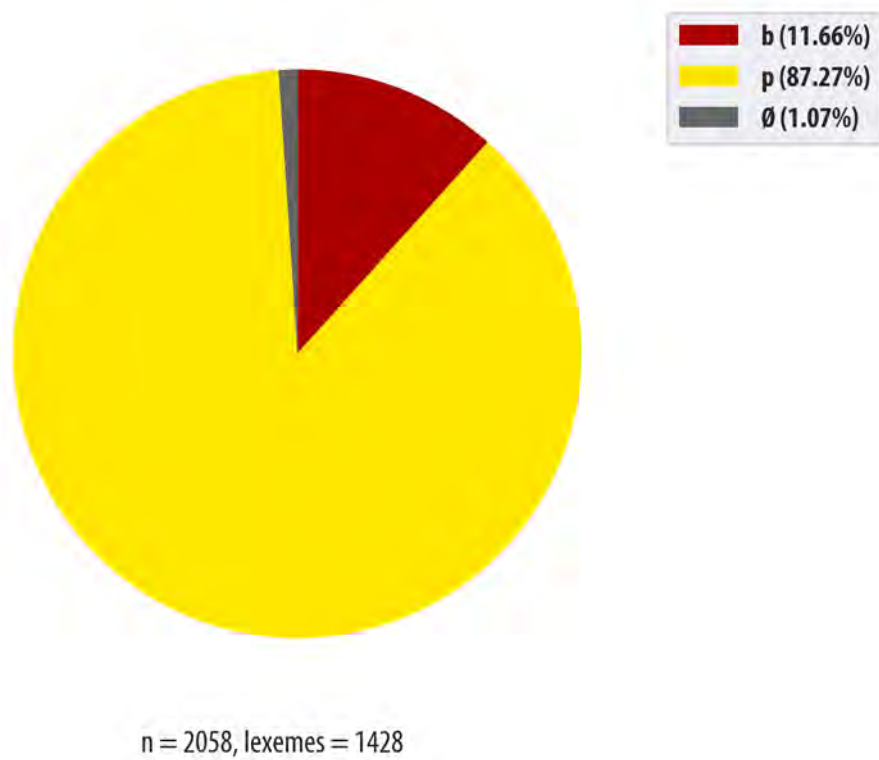


(c) Twitter (personal) data

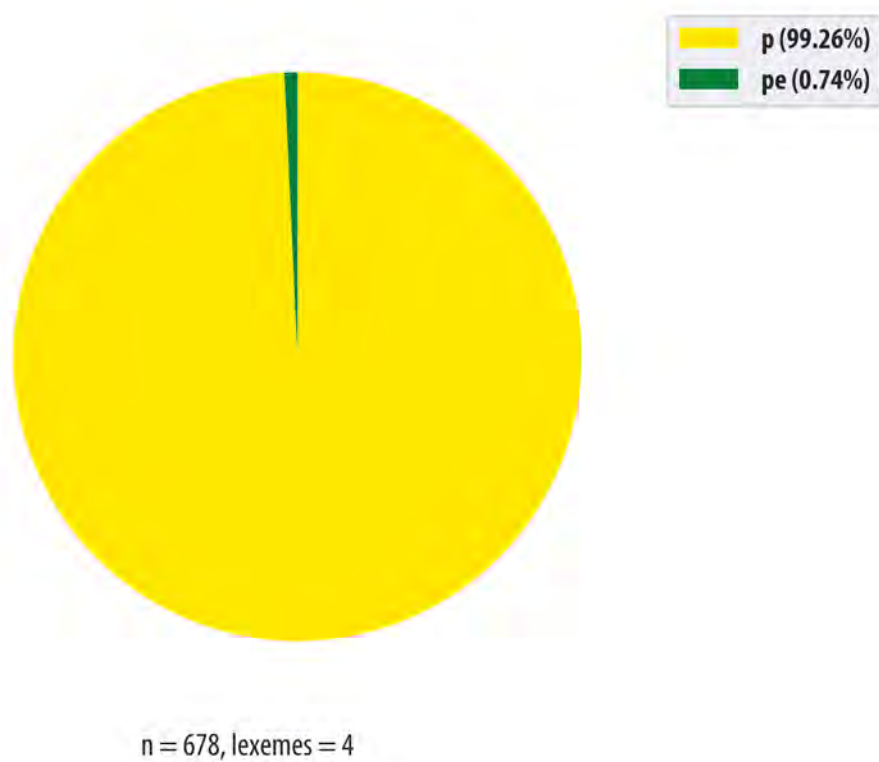


(d) Twitter (official) data

Figure C.32: Distribution of graphemes for phoneme /ə/ (**LETTER**) (cont.)

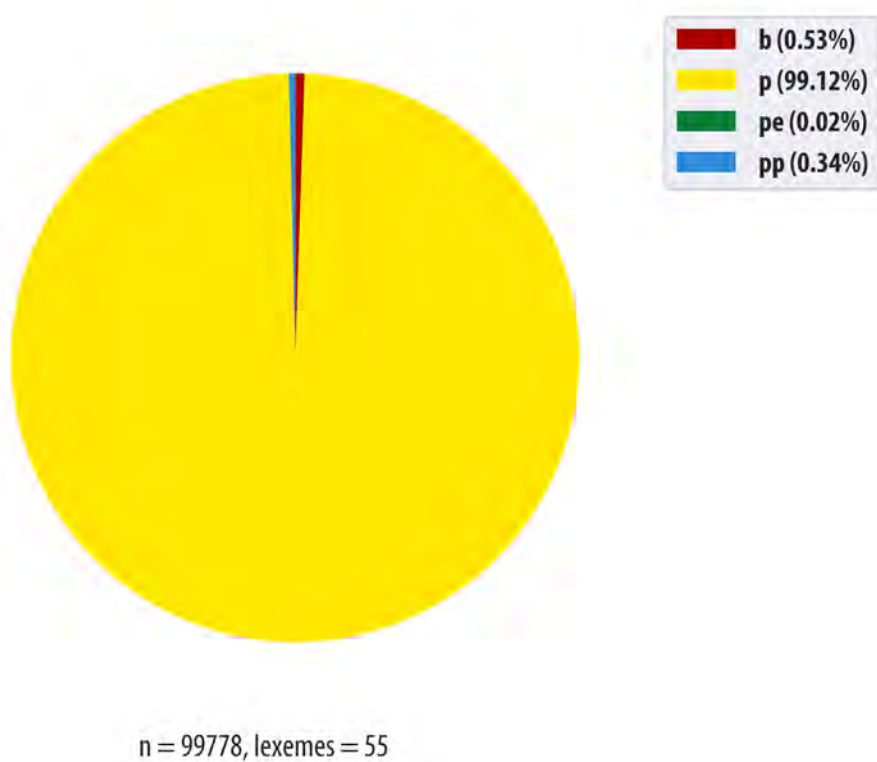


(a) GPC data

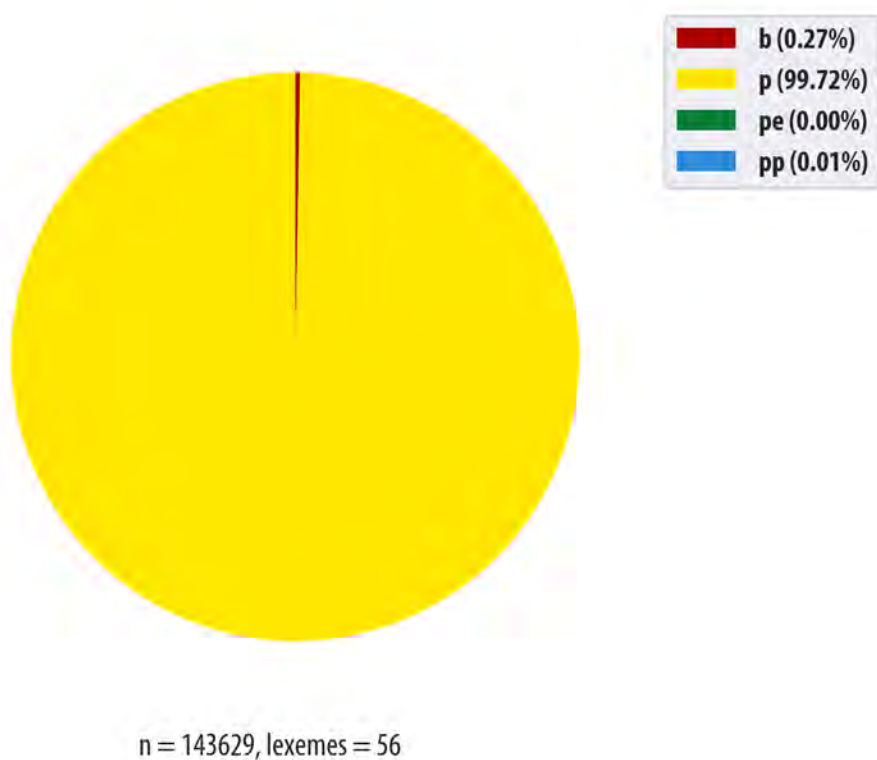


(b) Survey data

Figure C.33: Distribution of graphemes for phoneme /p/

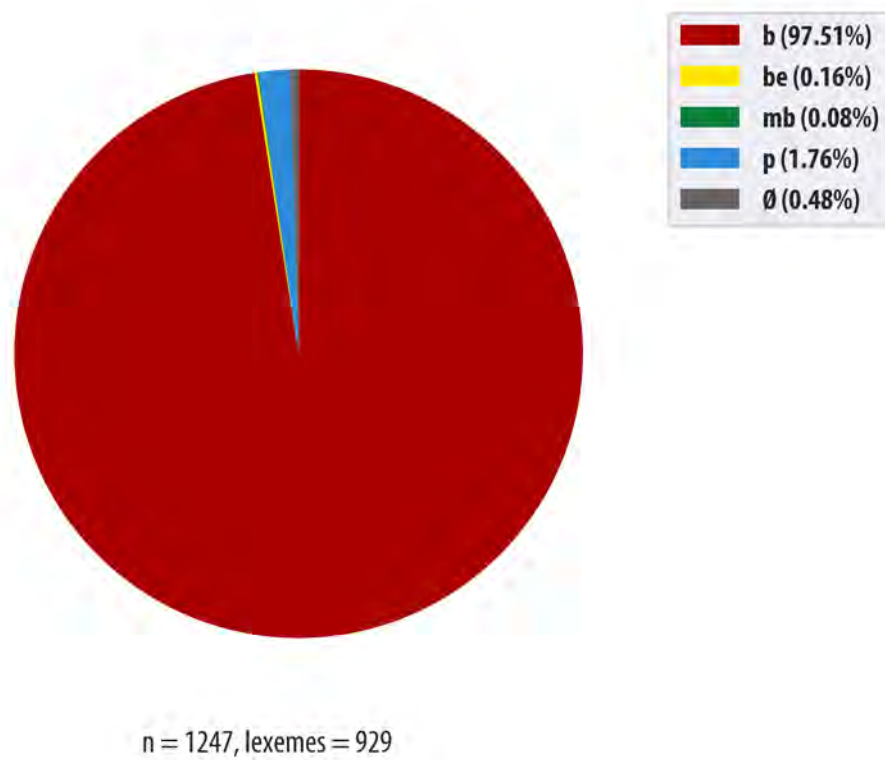


(c) Twitter (personal) data

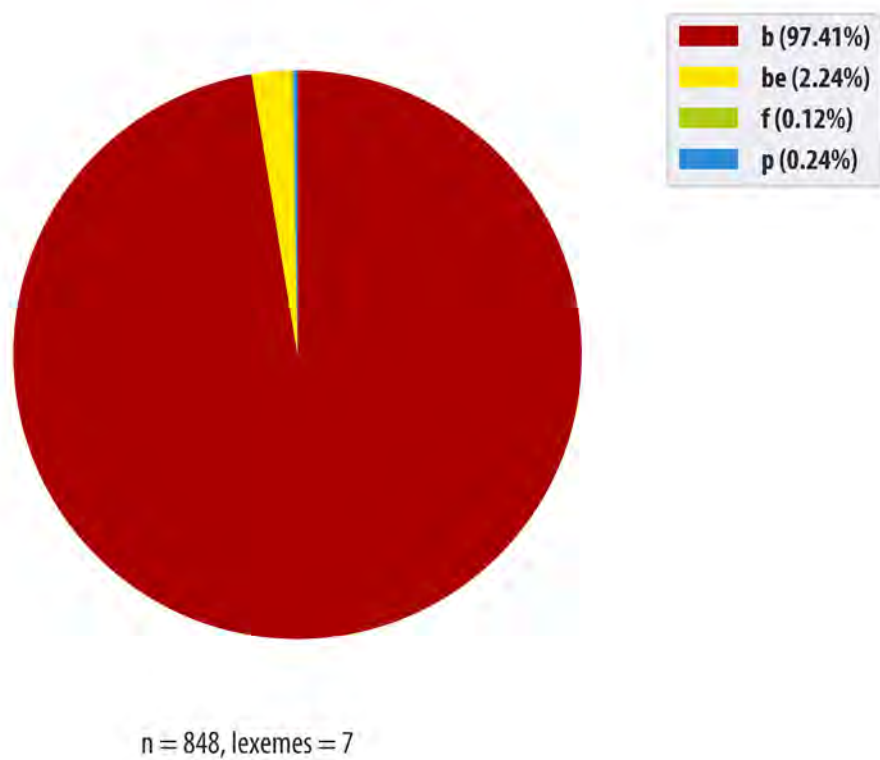


(d) Twitter (official) data

Figure C.33: Distribution of graphemes for phoneme /p/ (cont.)

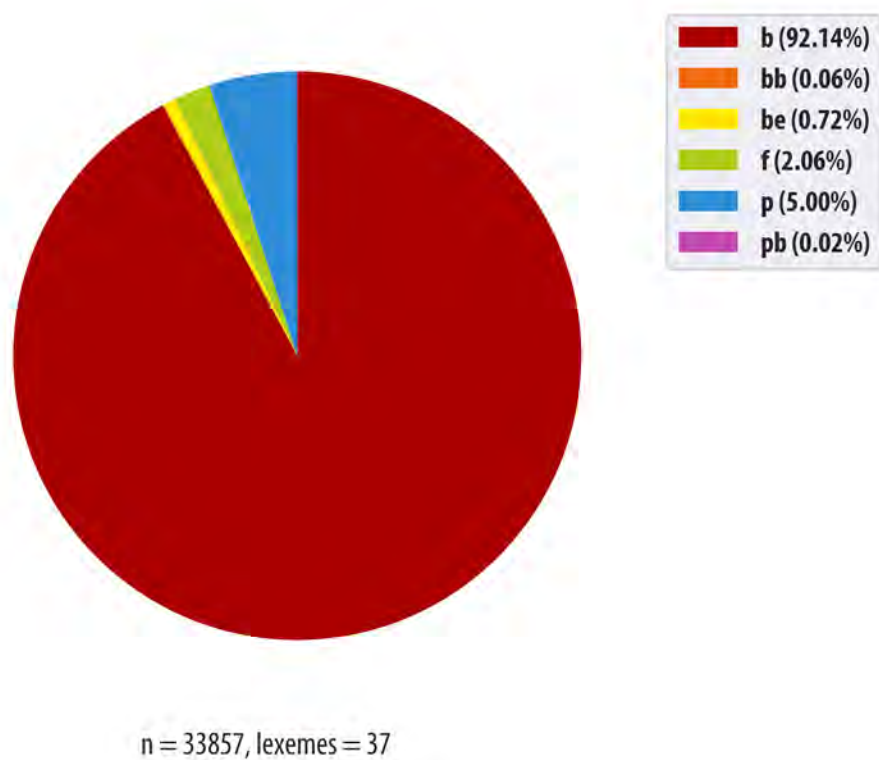


(a) GPC data



(b) Survey data

Figure C.34: Distribution of graphemes for phoneme /b/

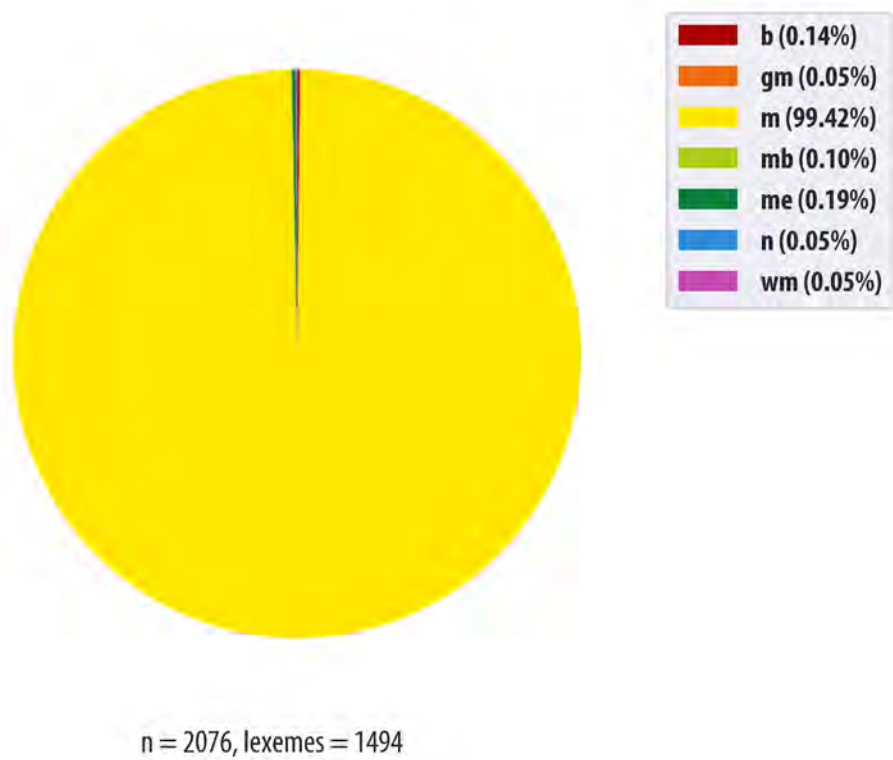


(c) Twitter (personal) data

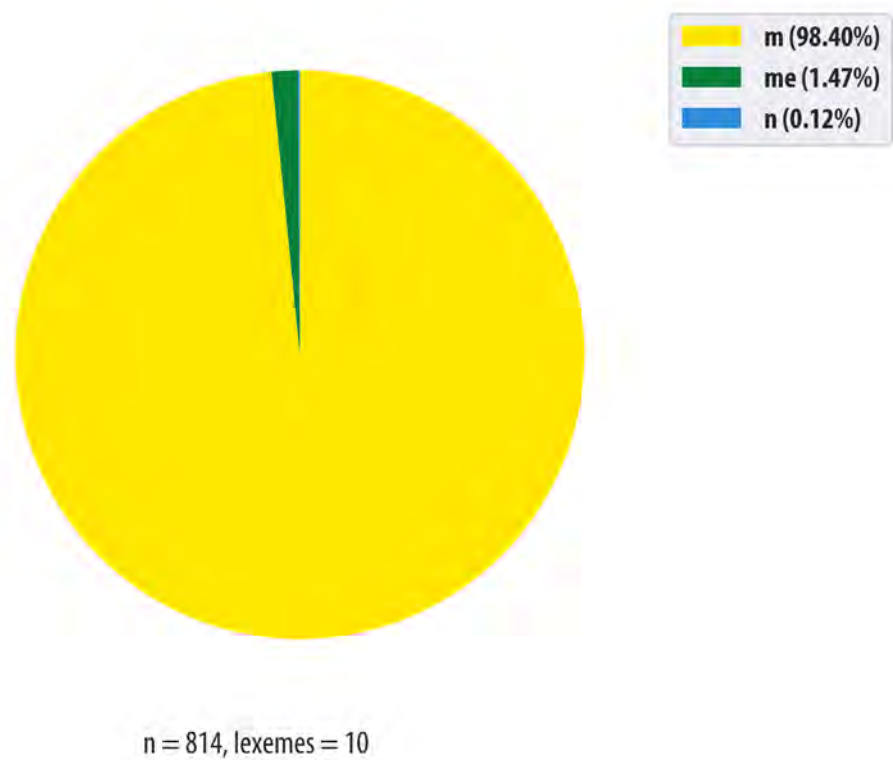


(d) Twitter (official) data

Figure C.34: Distribution of graphemes for phoneme /b/ (cont.)

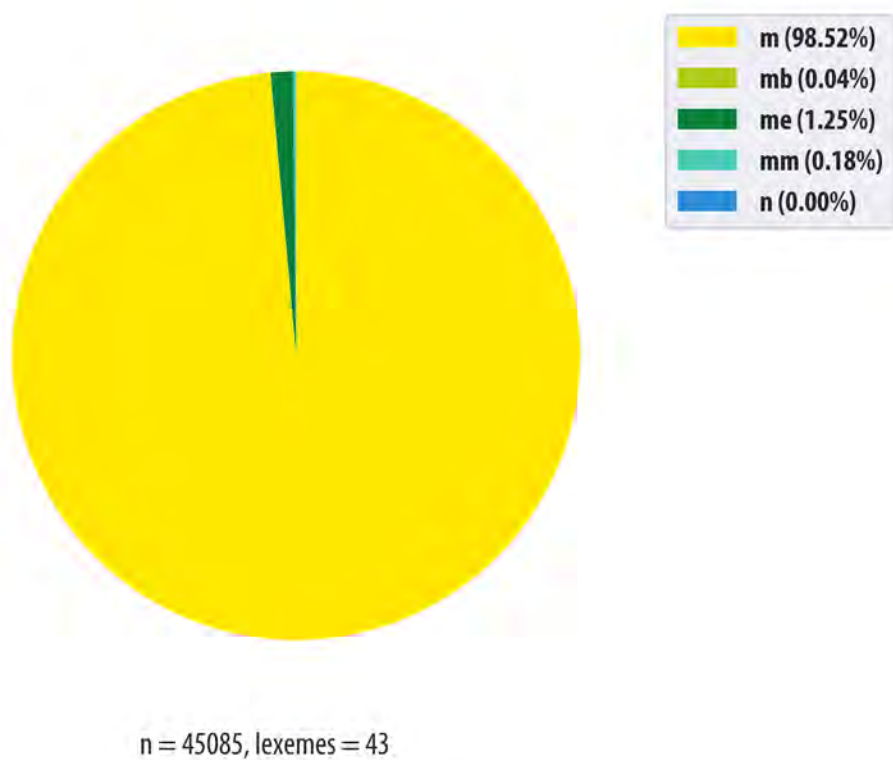


(a) GPC data

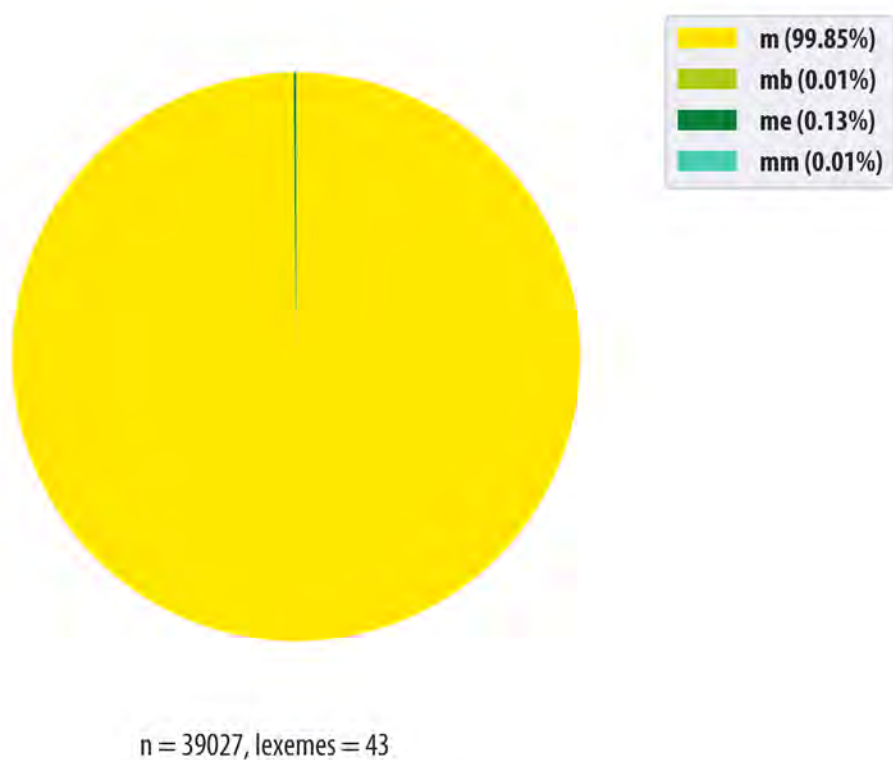


(b) Survey data

Figure C.35: Distribution of graphemes for phoneme /m/

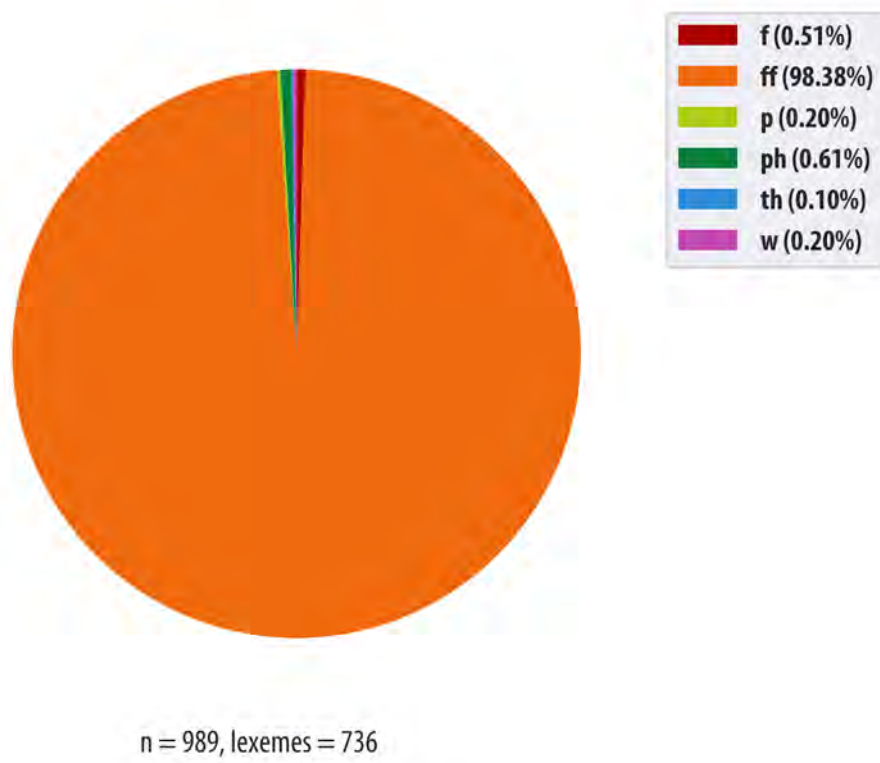


(c) Twitter (personal) data

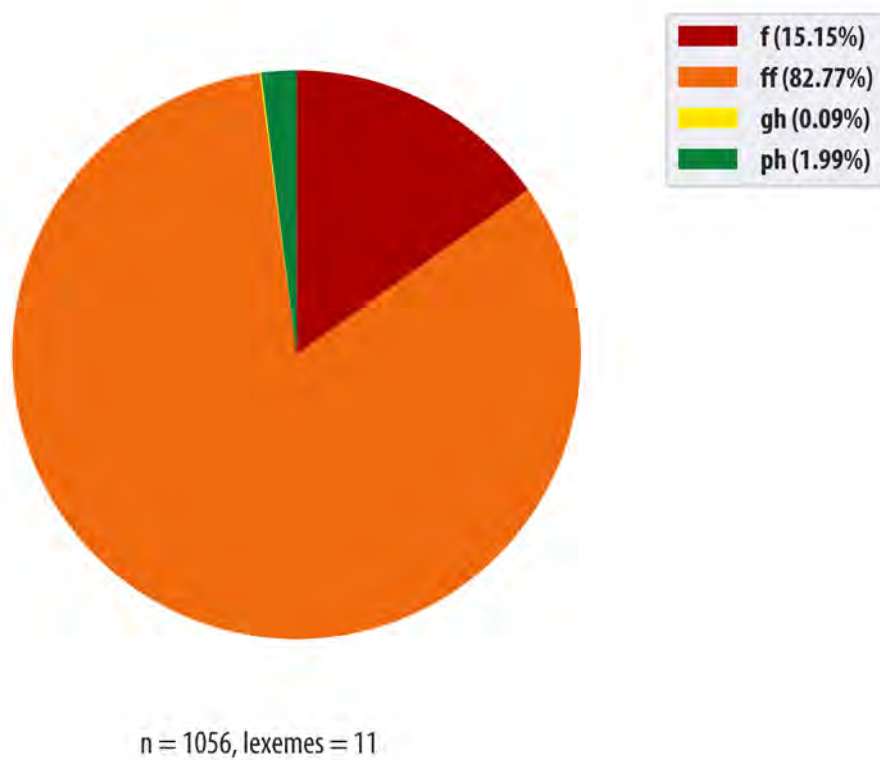


(d) Twitter (official) data

Figure C.35: Distribution of graphemes for phoneme /m/ (cont.)

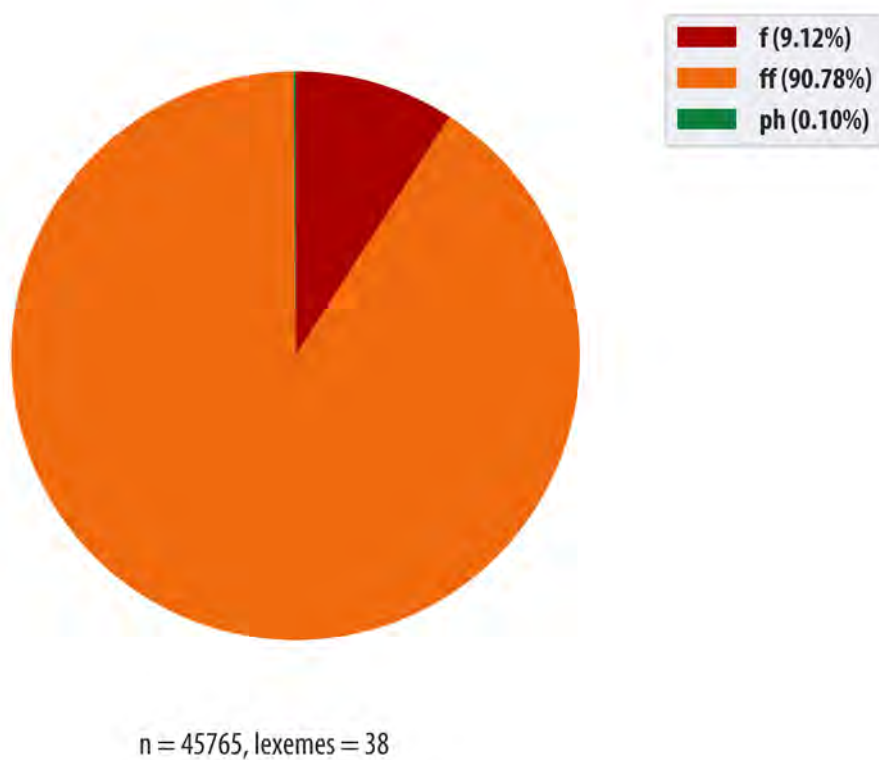


(a) GPC data

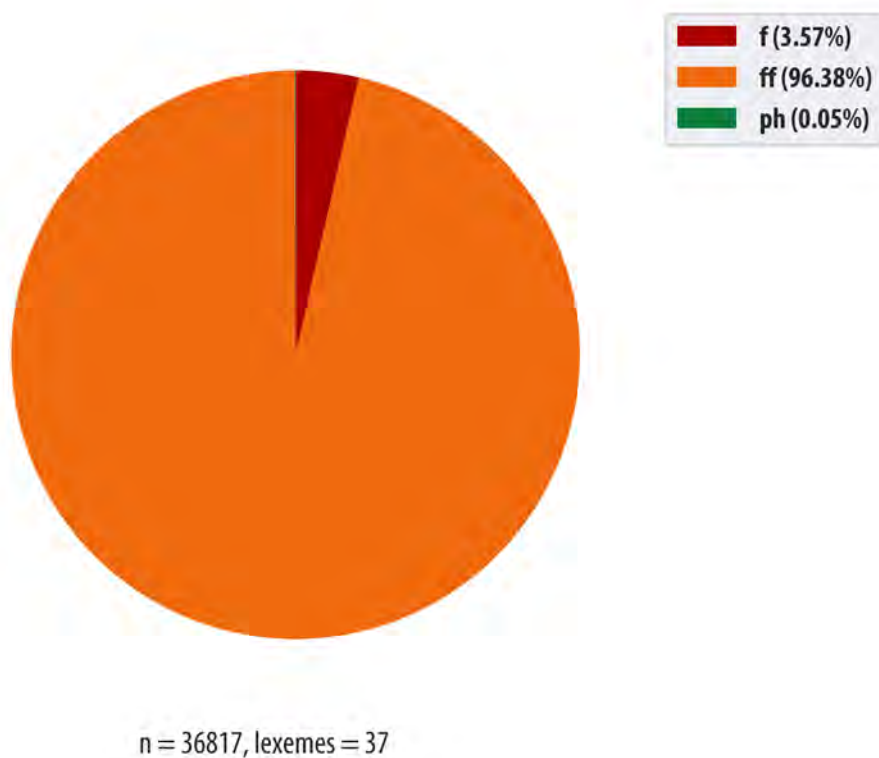


(b) Survey data

Figure C.36: Distribution of graphemes for phoneme /f/

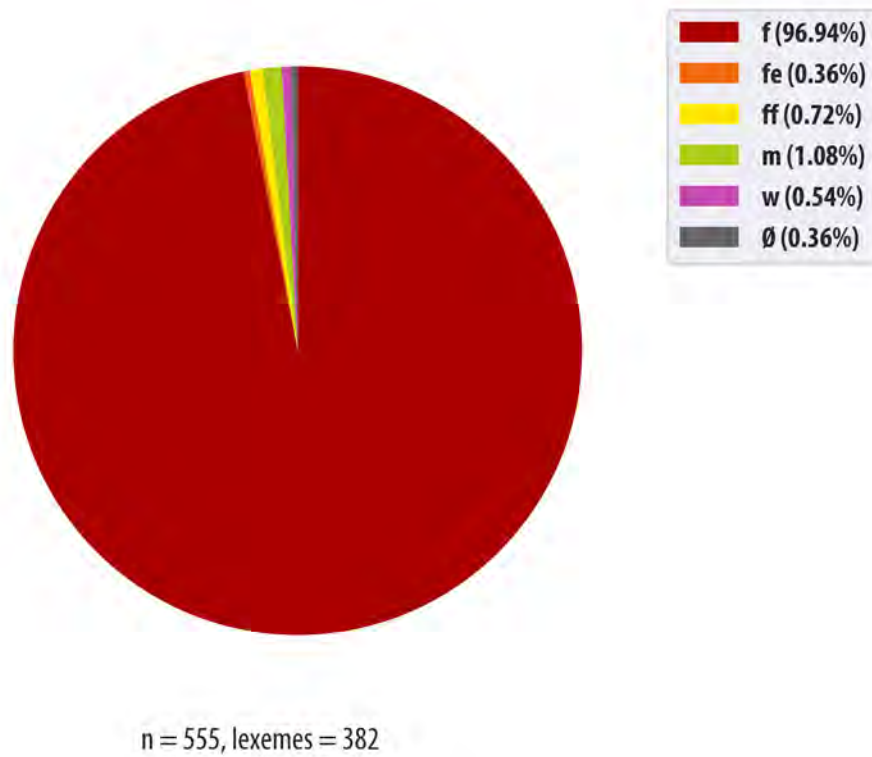


(c) Twitter (personal) data

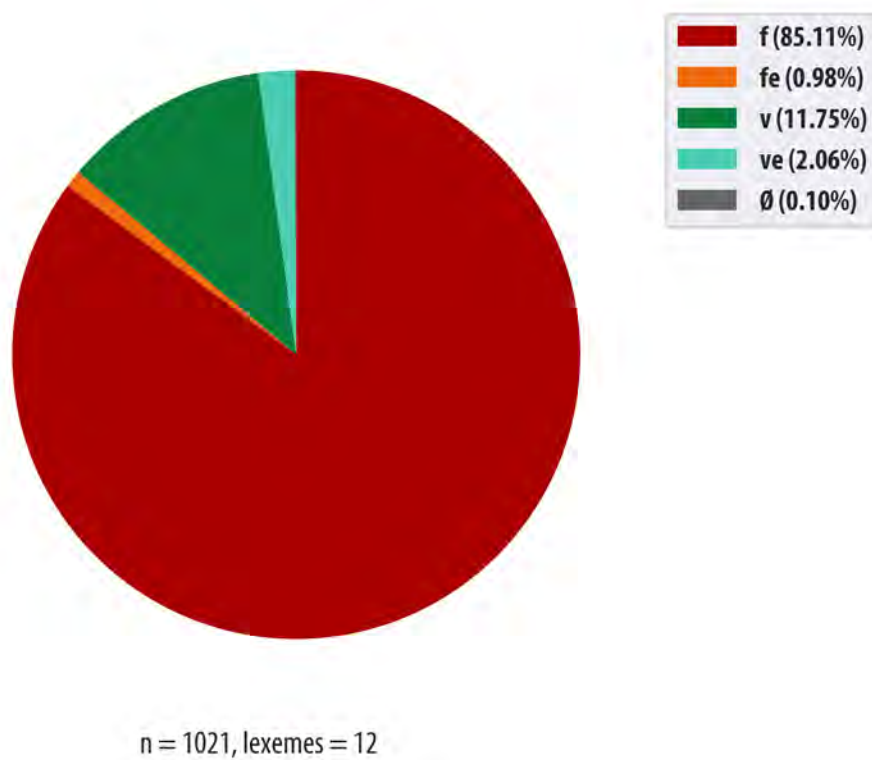


(d) Twitter (official) data

Figure C.36: Distribution of graphemes for phoneme /f/ (cont.)

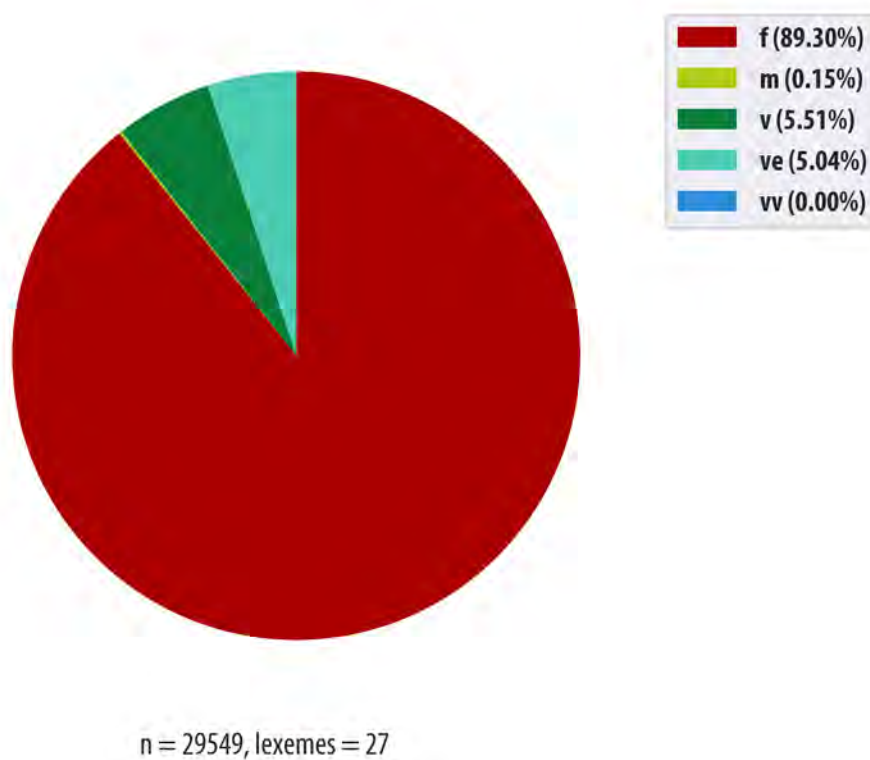


(a) GPC data

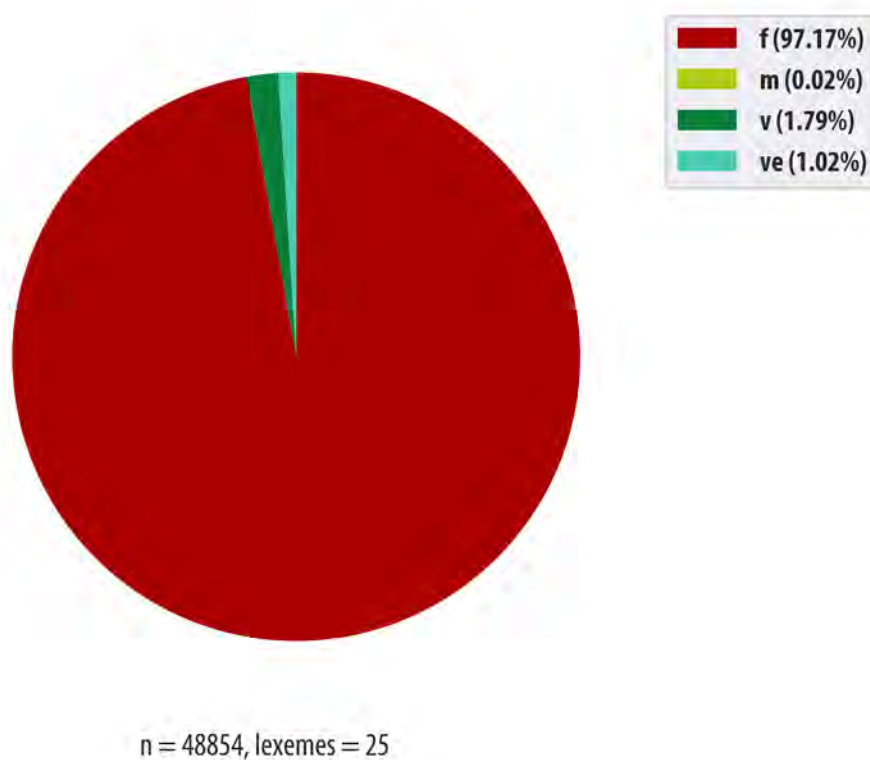


(b) Survey data

Figure C.37: Distribution of graphemes for phoneme /v/

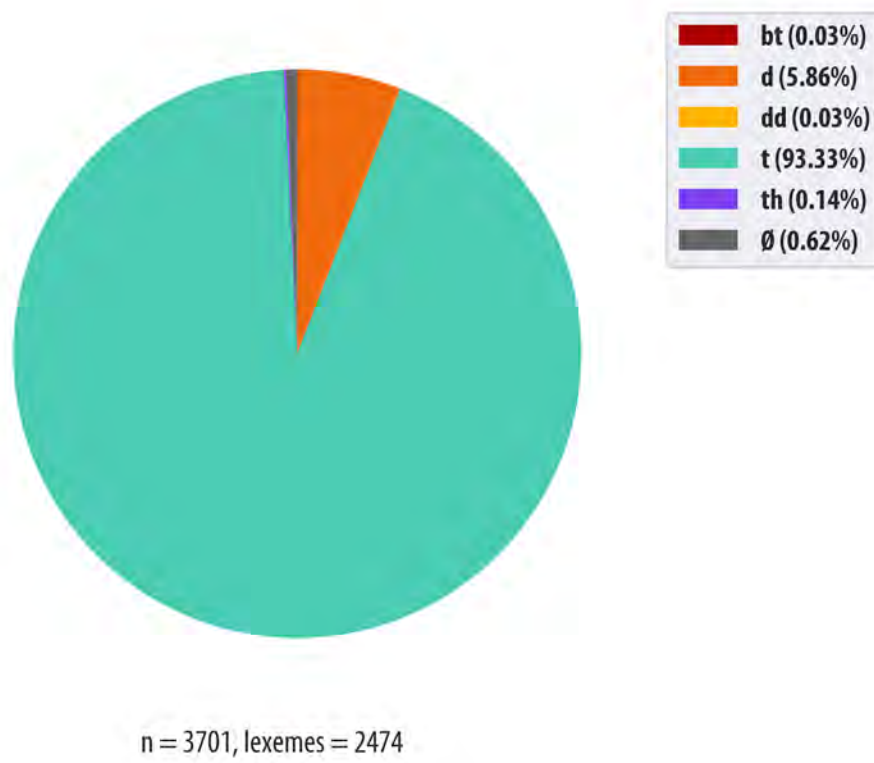


(c) Twitter (personal) data

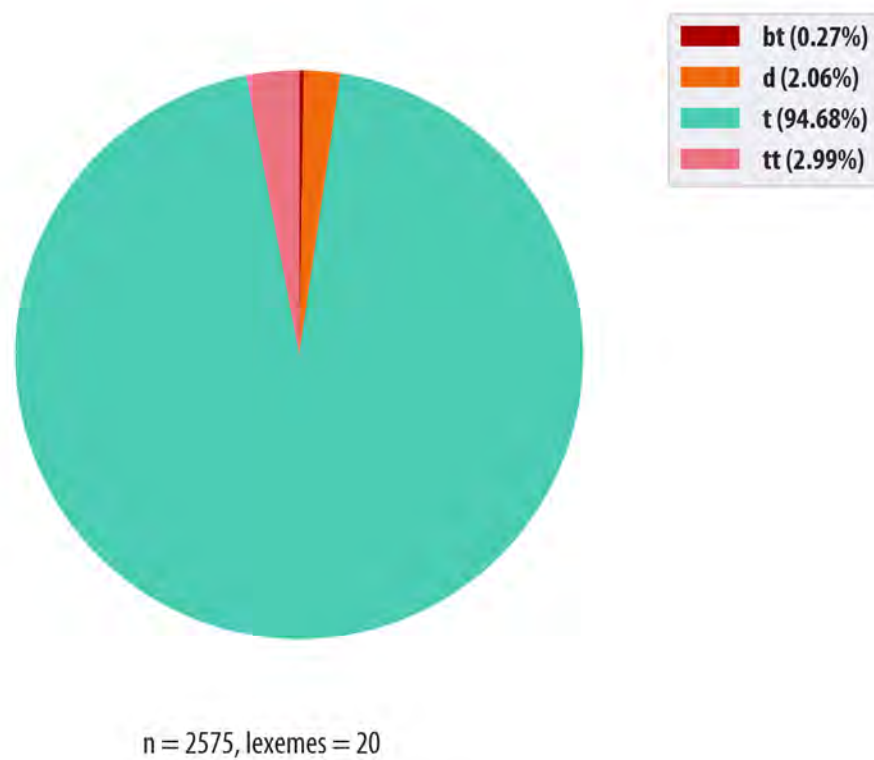


(d) Twitter (official) data

Figure C.37: Distribution of graphemes for phoneme /v/ (cont.)

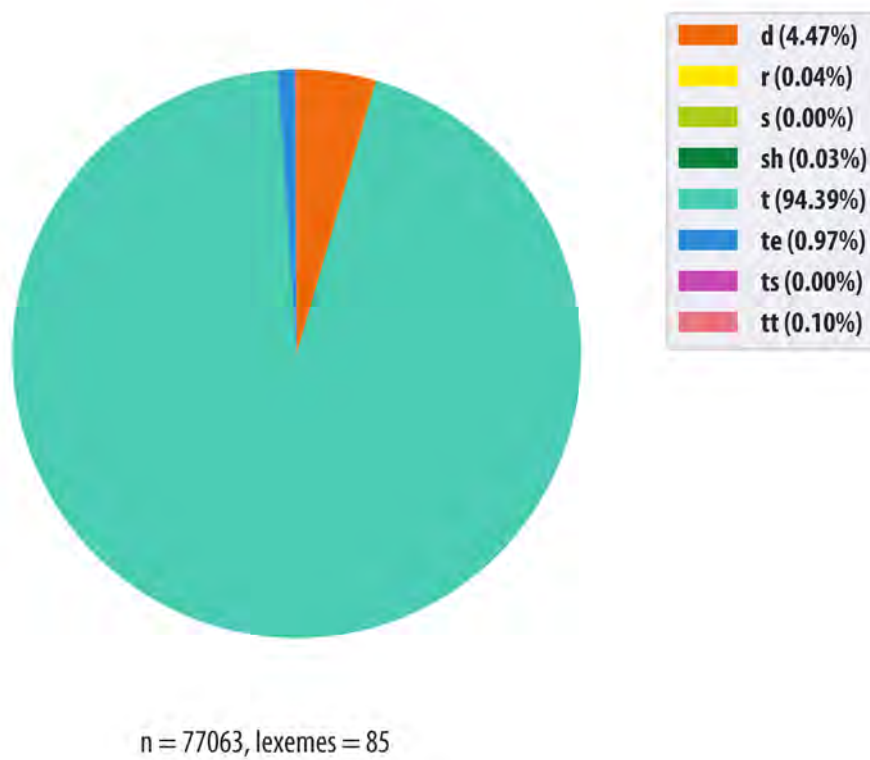


(a) GPC data

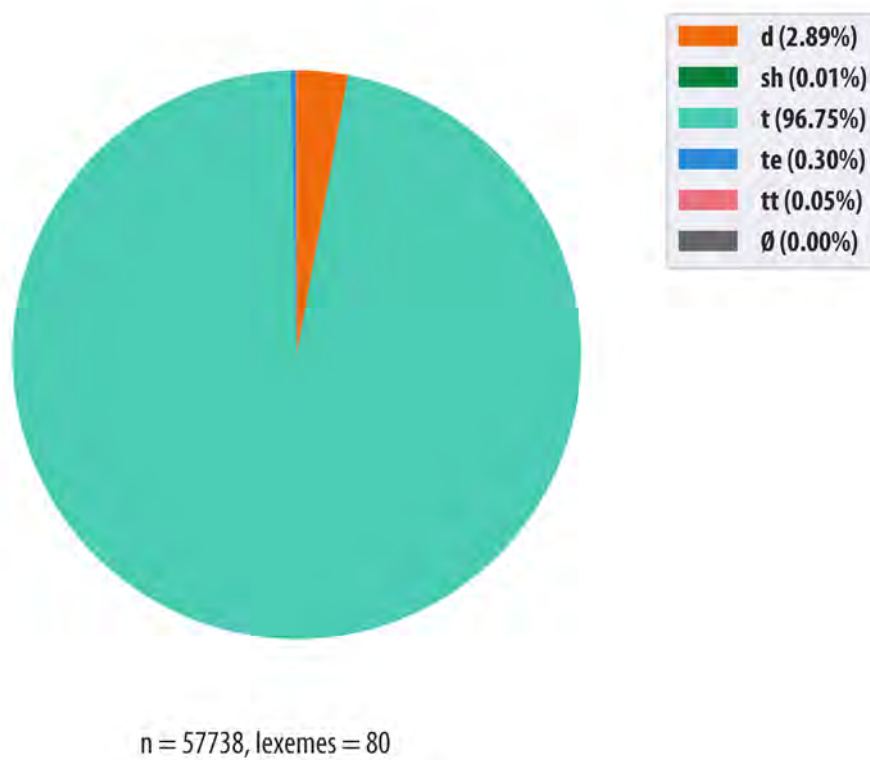


(b) Survey data

Figure C.38: Distribution of graphemes for phoneme /t/

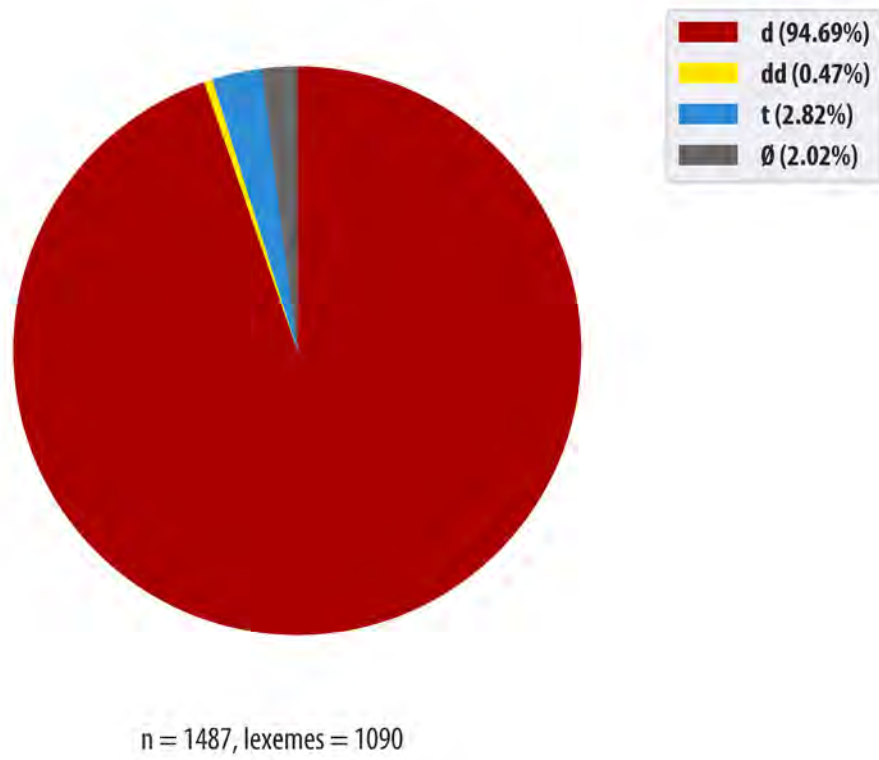


(c) Twitter (personal) data

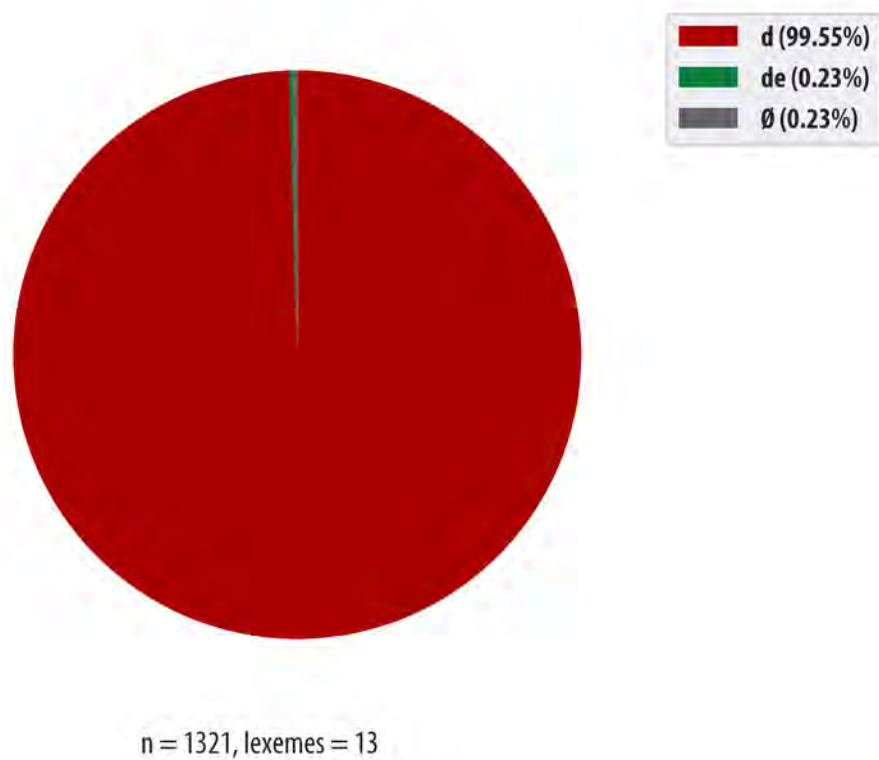


(d) Twitter (official) data

Figure C.38: Distribution of graphemes for phoneme /t/ (cont.)

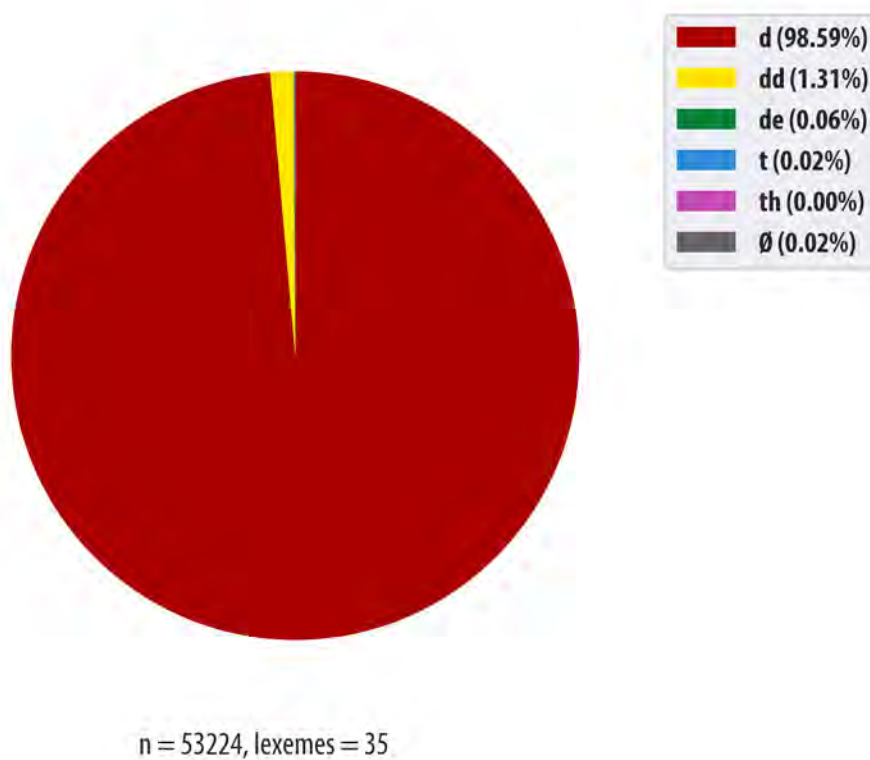


(a) GPC data

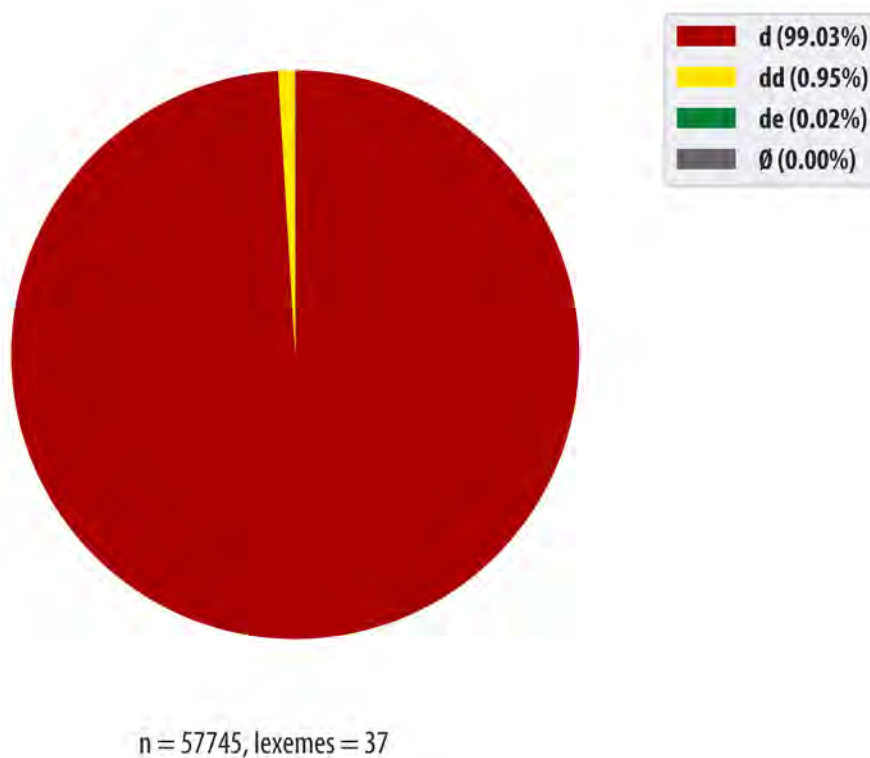


(b) Survey data

Figure C.39: Distribution of graphemes for phoneme /d/

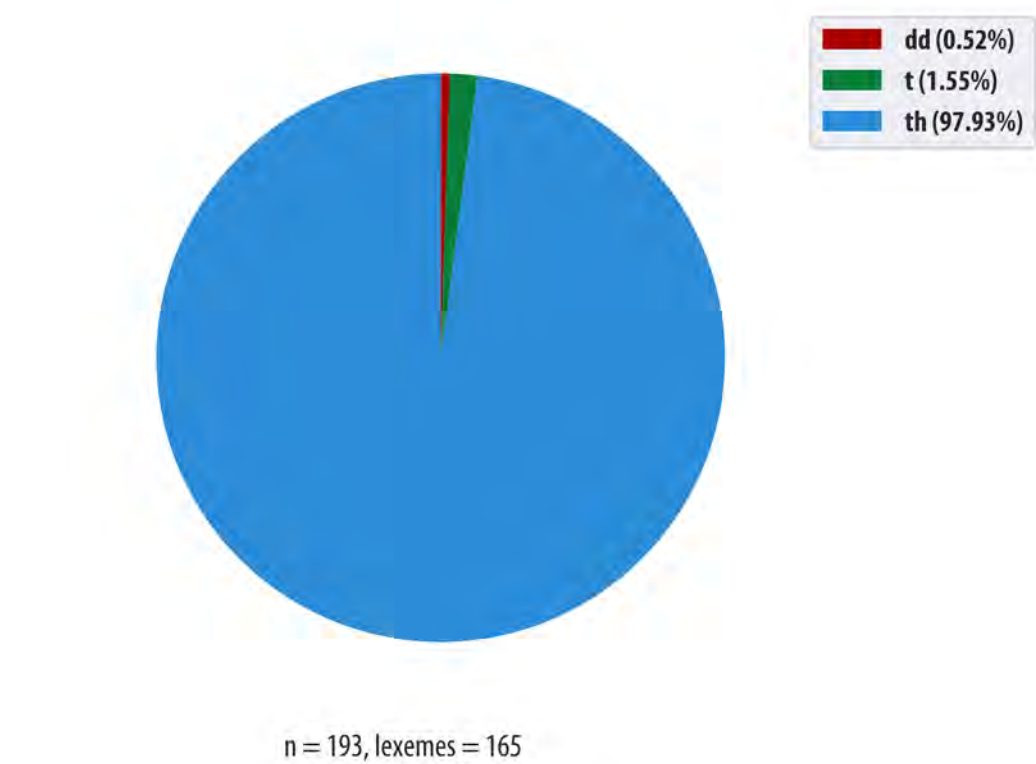


(c) Twitter (personal) data

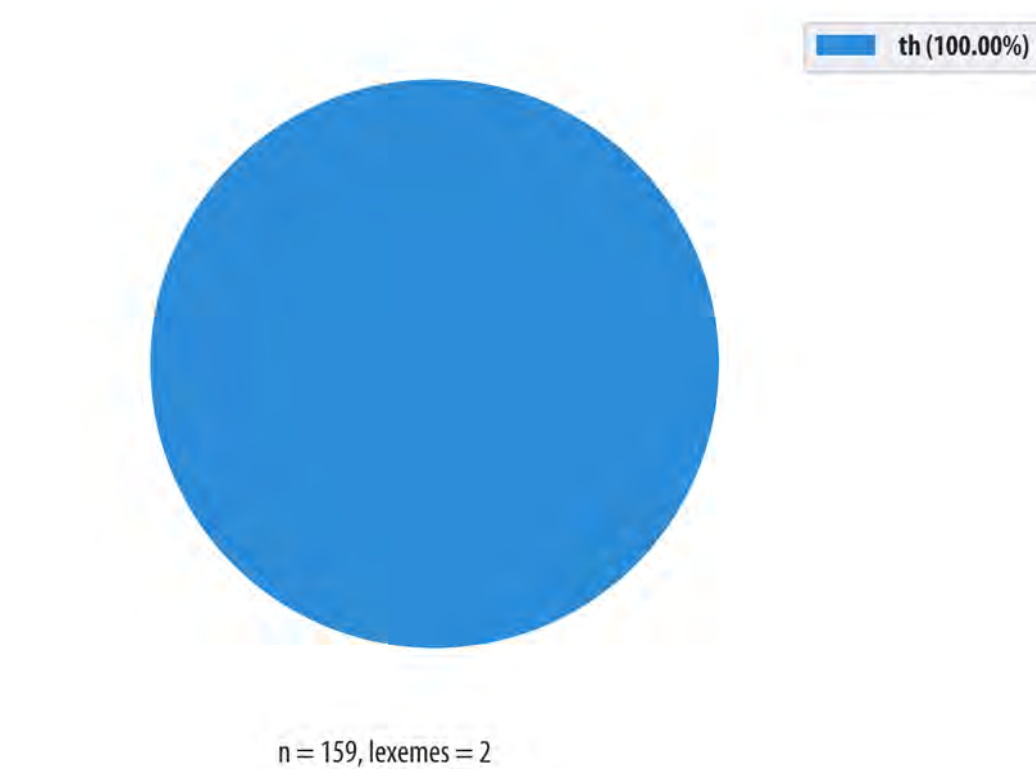


(d) Twitter (official) data

Figure C.39: Distribution of graphemes for phoneme /d/ (cont.)

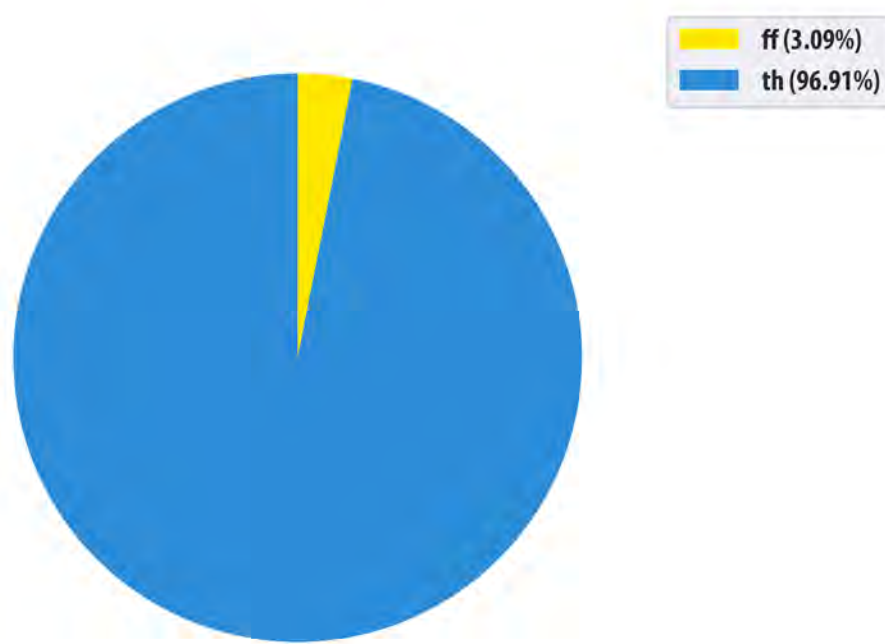


(a) GPC data



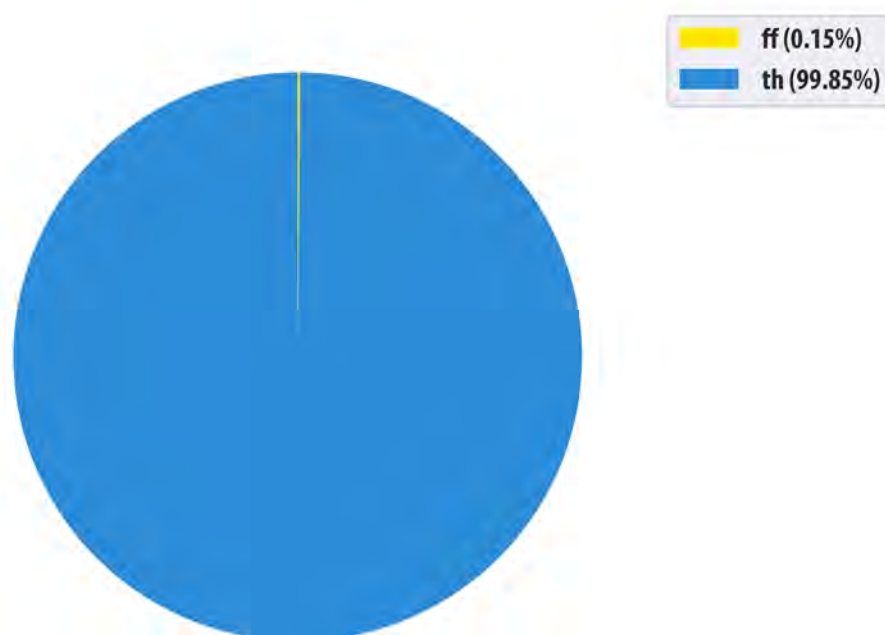
(b) Survey data

Figure C.40: Distribution of graphemes for phoneme /θ/



n = 1293, lexemes = 5

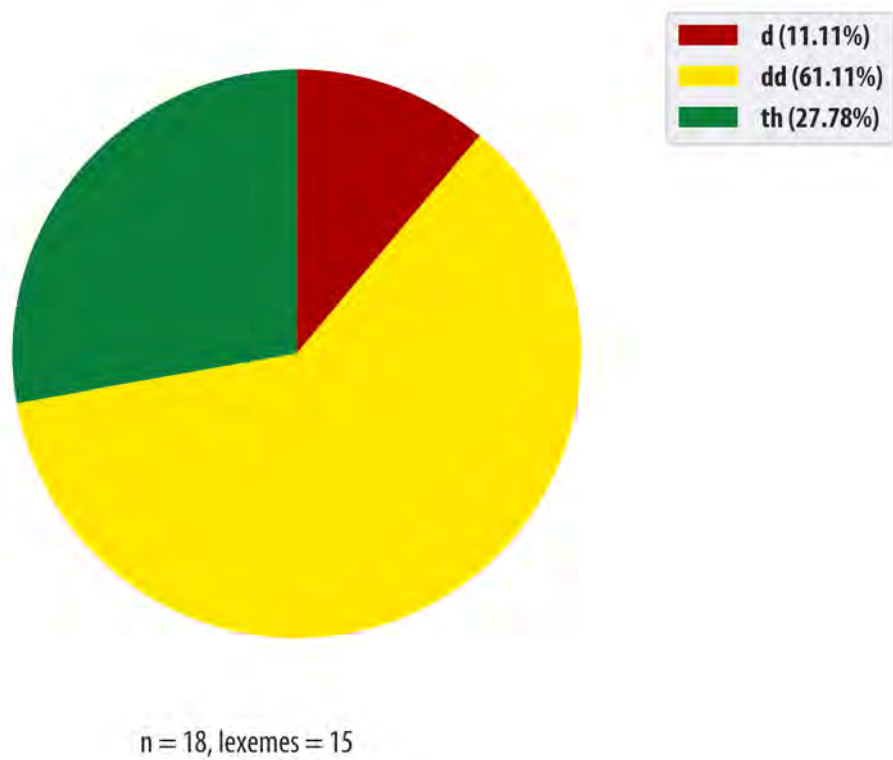
(c) Twitter (personal) data



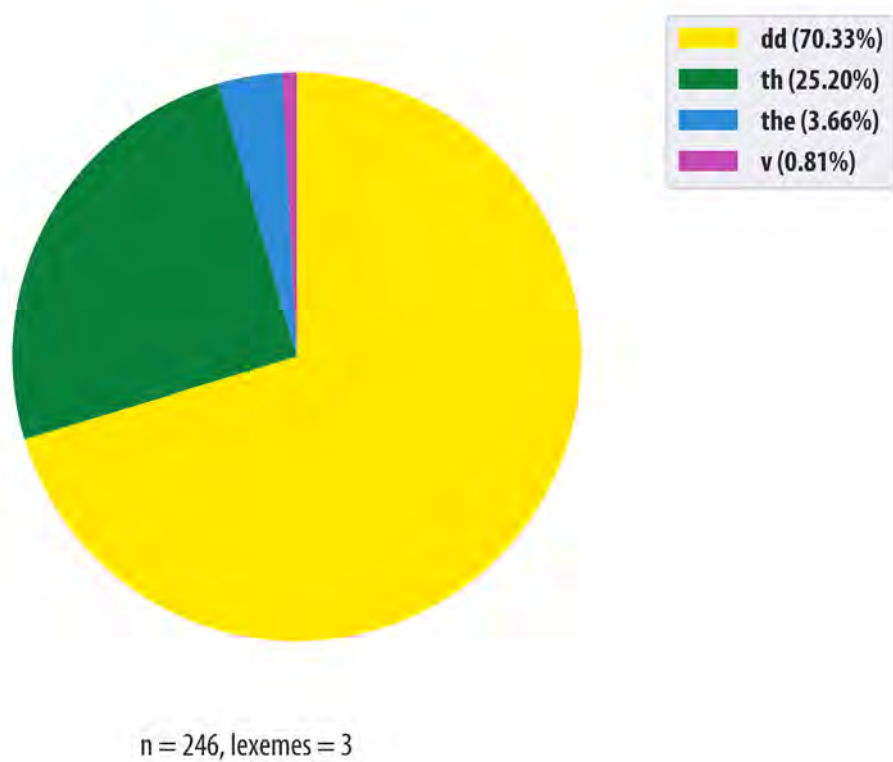
n = 1376, lexemes = 4

(d) Twitter (official) data

Figure C.40: Distribution of graphemes for phoneme /θ/ (cont.)

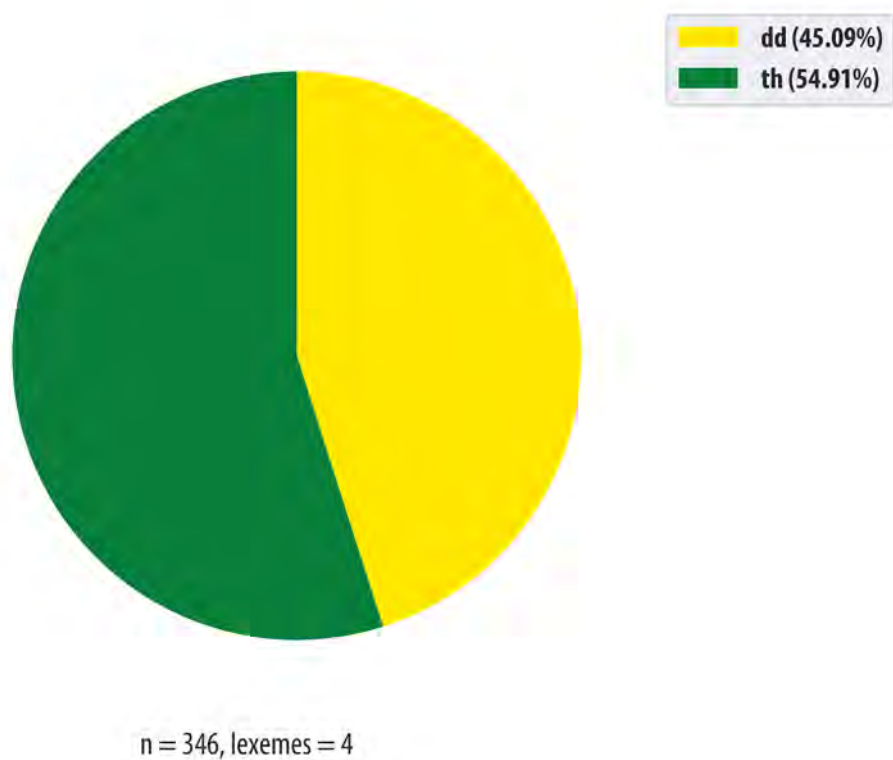


(a) GPC data

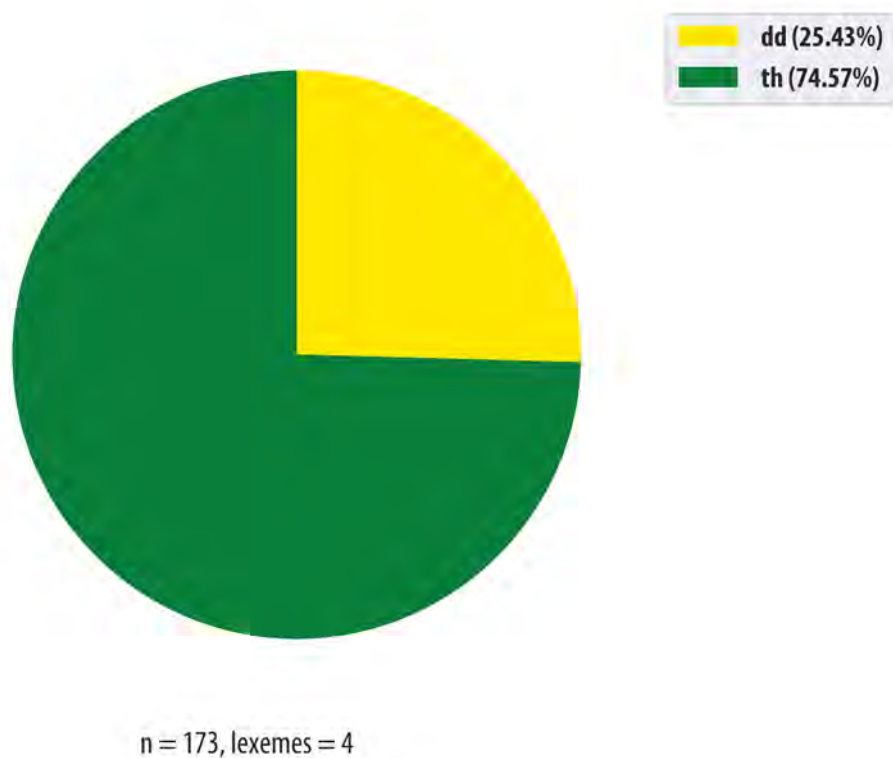


(b) Survey data

Figure C.41: Distribution of graphemes for phoneme /ð/

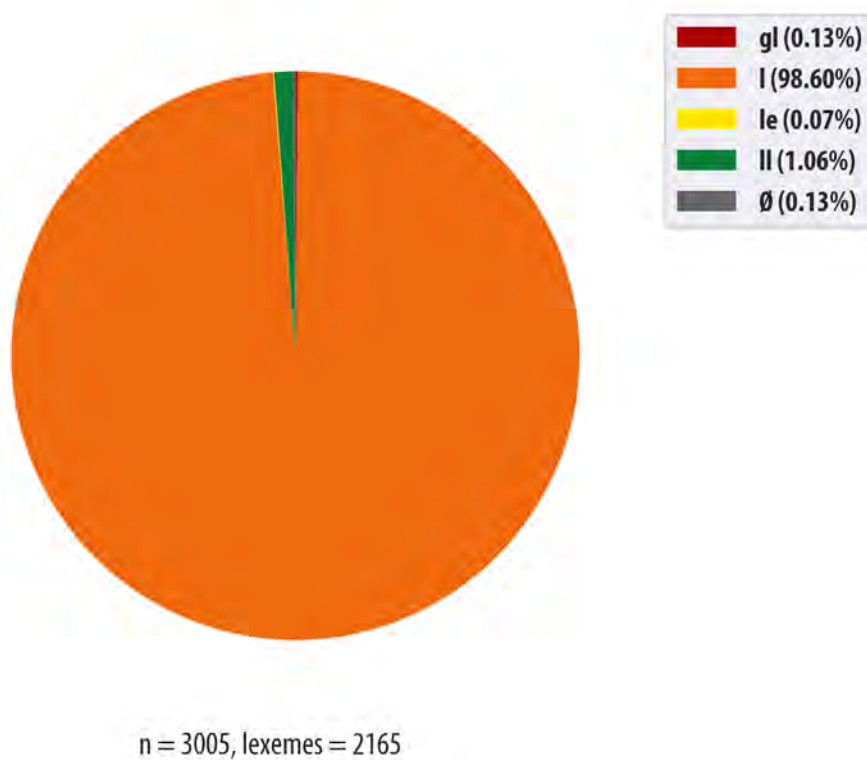


(c) Twitter (personal) data

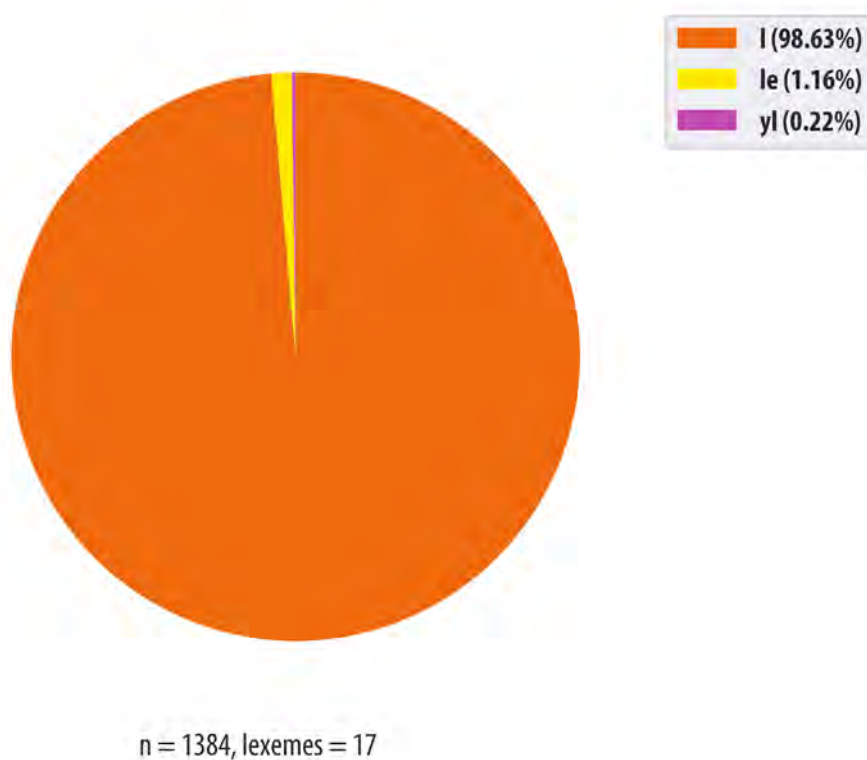


(d) Twitter (official) data

Figure C.41: Distribution of graphemes for phoneme /ð/ (cont.)

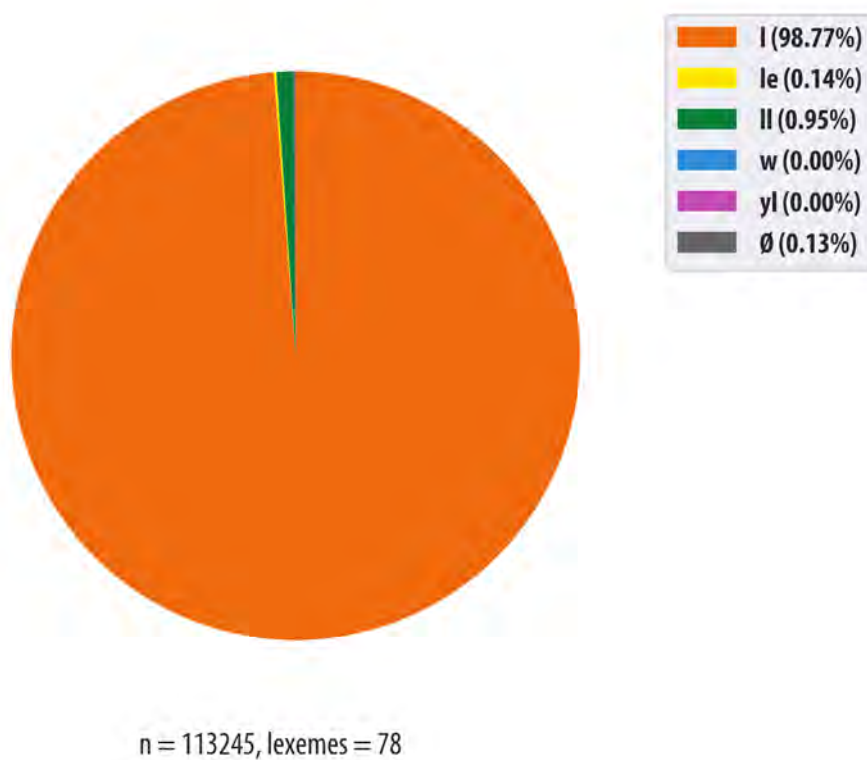


(a) GPC data

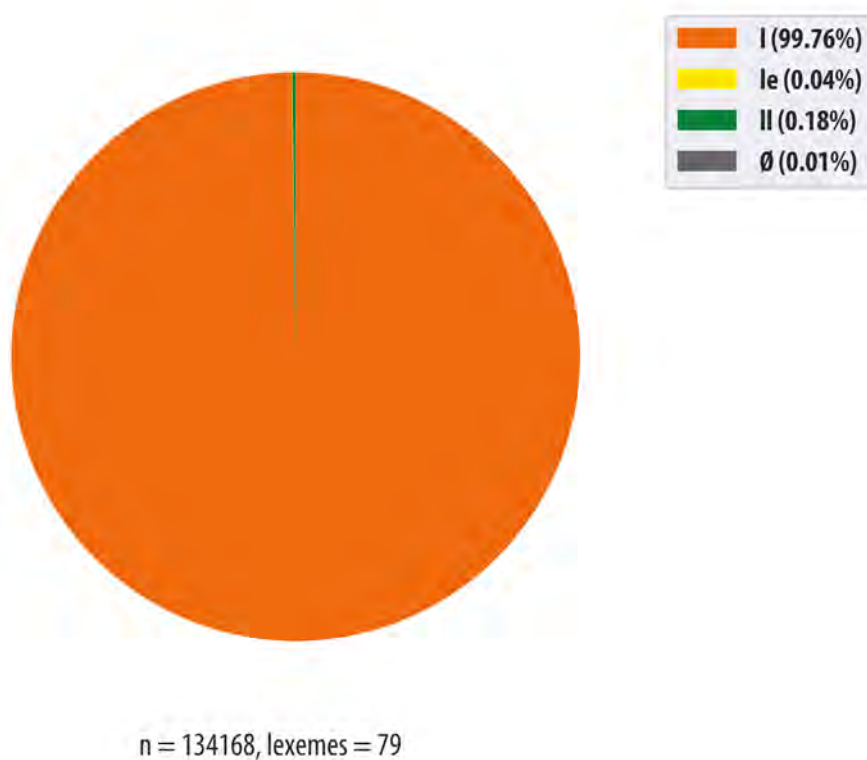


(b) Survey data

Figure C.42: Distribution of graphemes for phoneme /l/

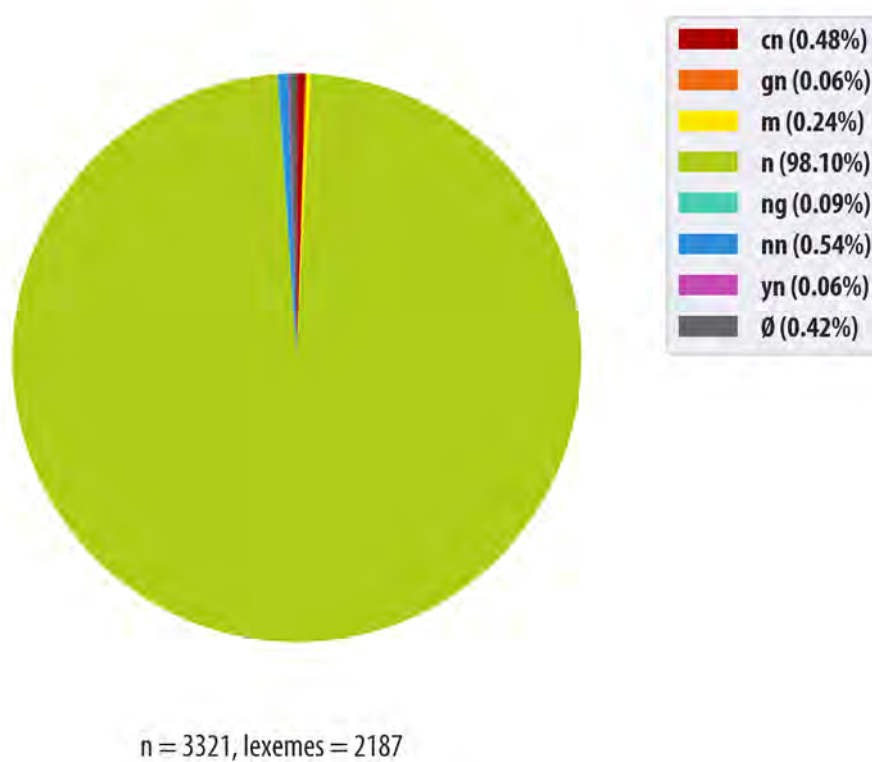


(c) Twitter (personal) data

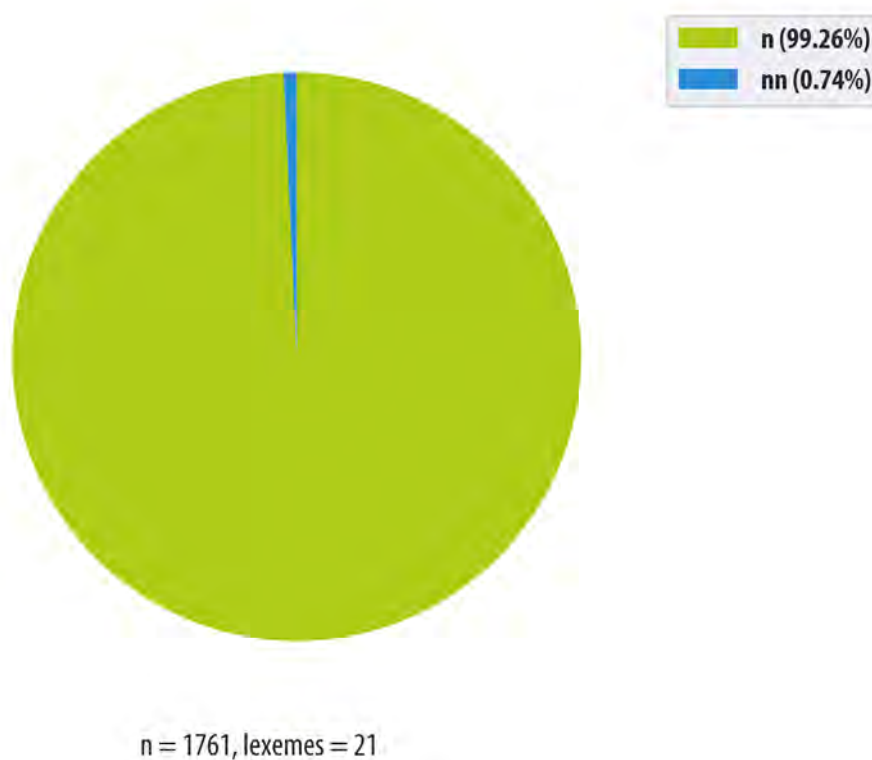


(d) Twitter (official) data

Figure C.42: Distribution of graphemes for phoneme /l/ (cont.)

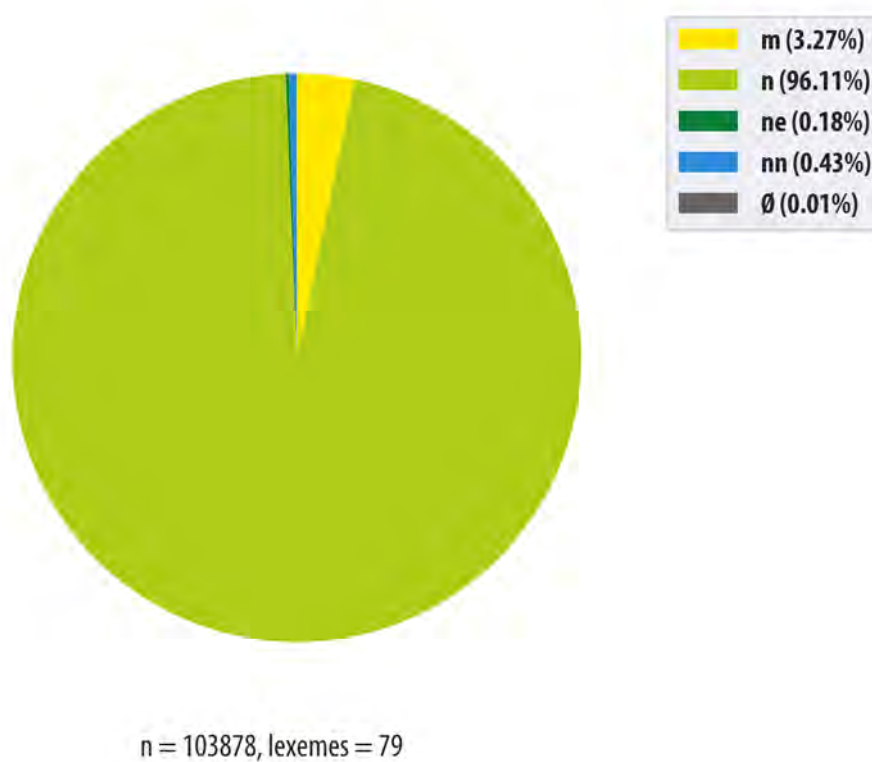


(a) GPC data

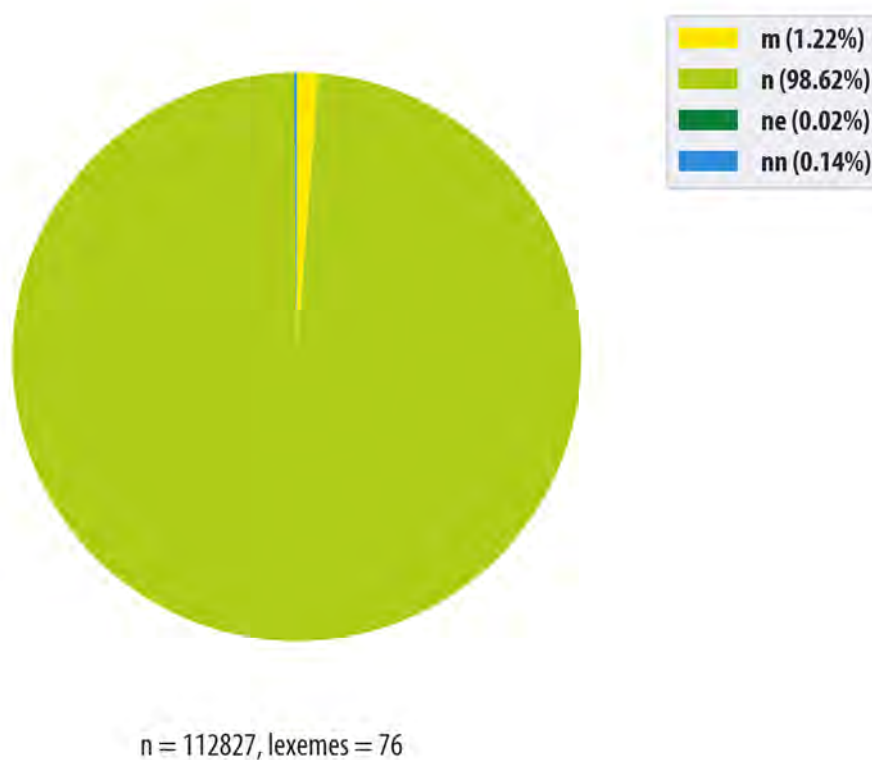


(b) Survey data

Figure C.43: Distribution of graphemes for phoneme /n/

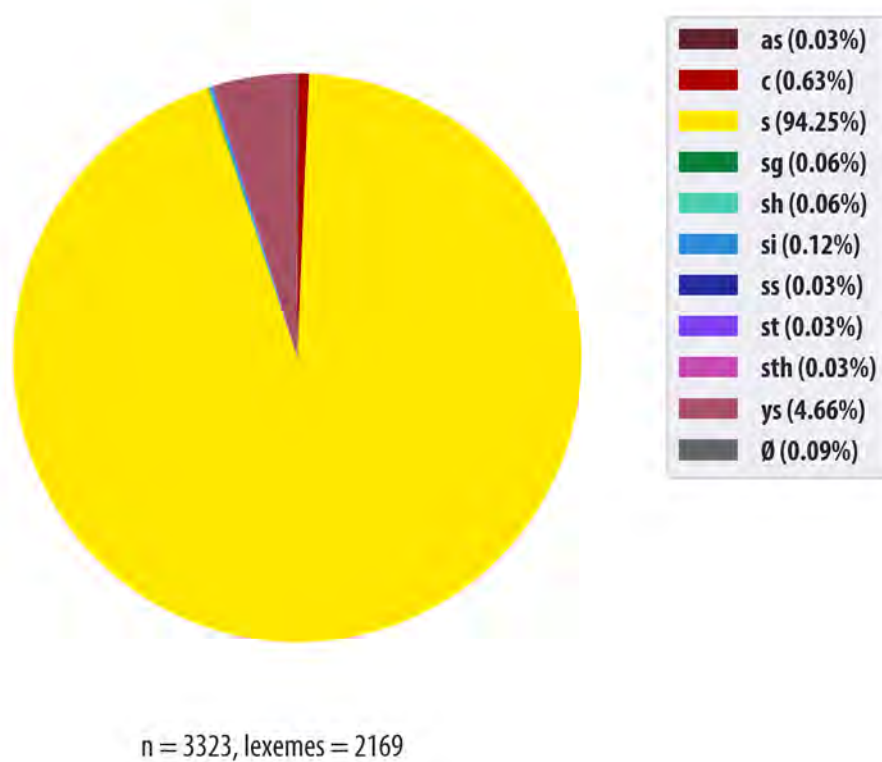


(c) Twitter (personal) data

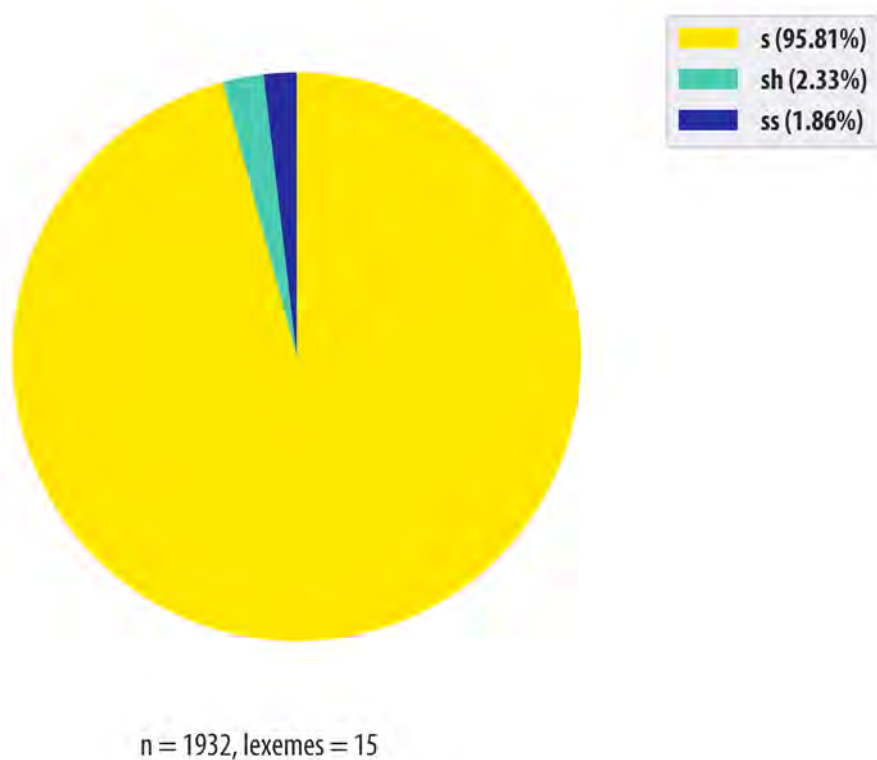


(d) Twitter (official) data

Figure C.43: Distribution of graphemes for phoneme /n/ (cont.)

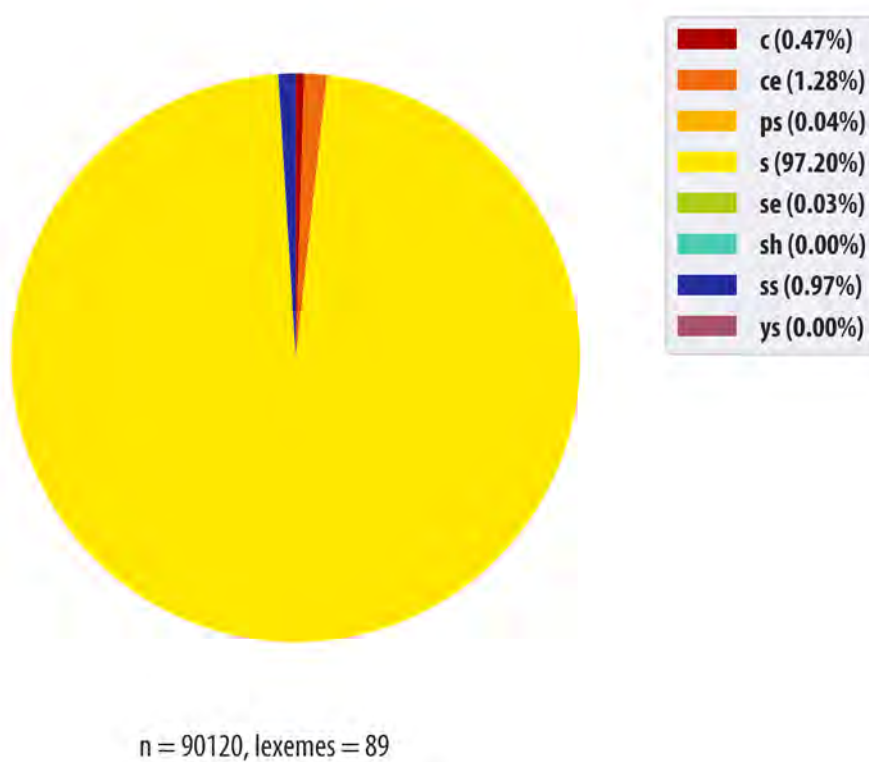


(a) GPC data

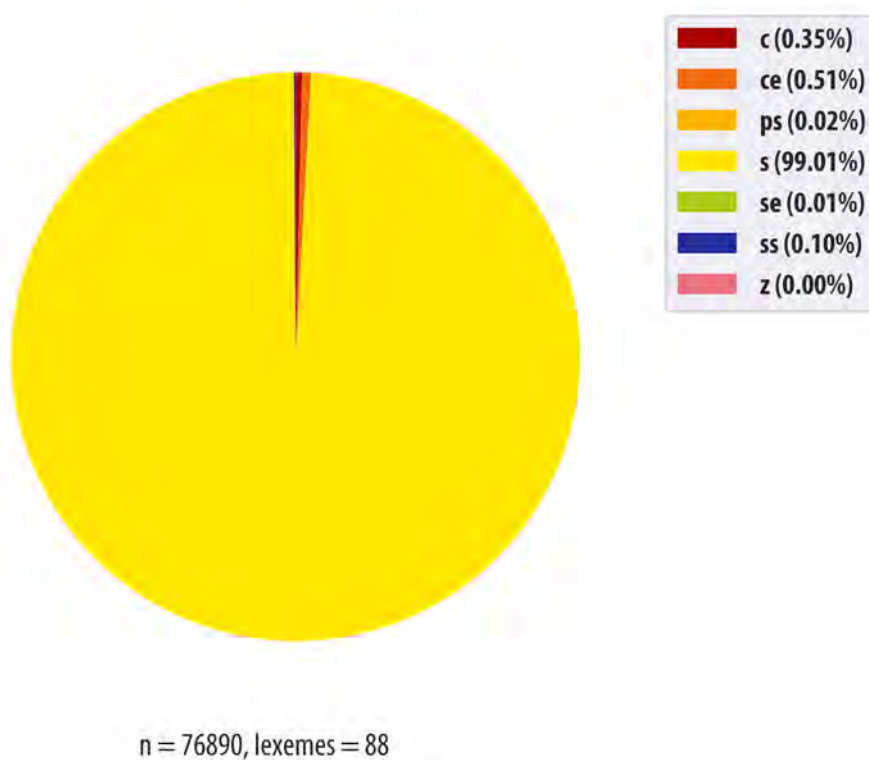


(b) Survey data

Figure C.44: Distribution of graphemes for phoneme /s/

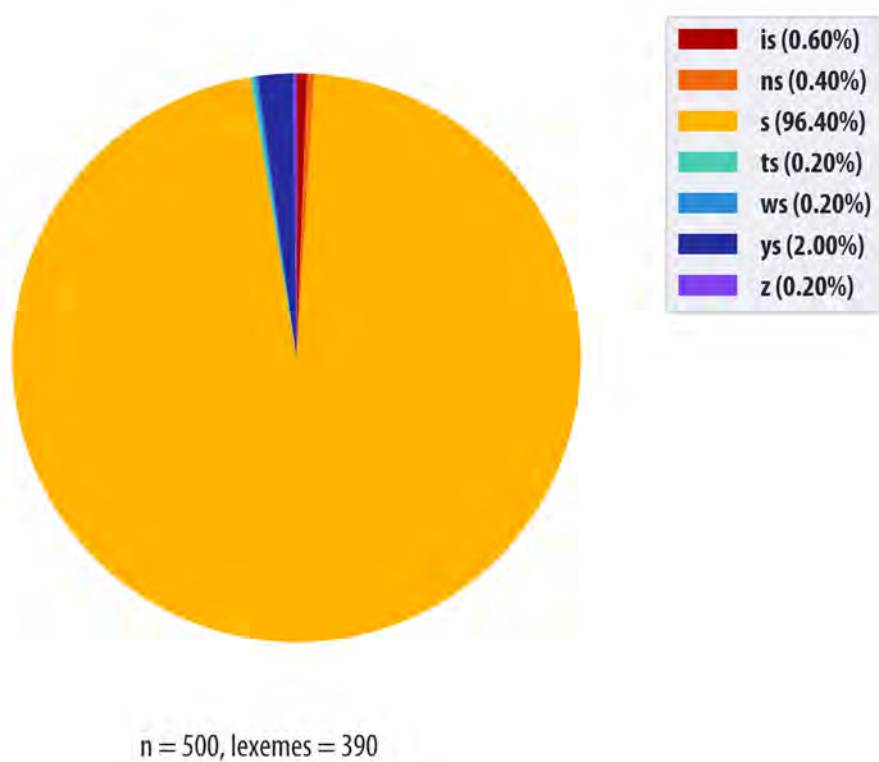


(c) Twitter (personal) data

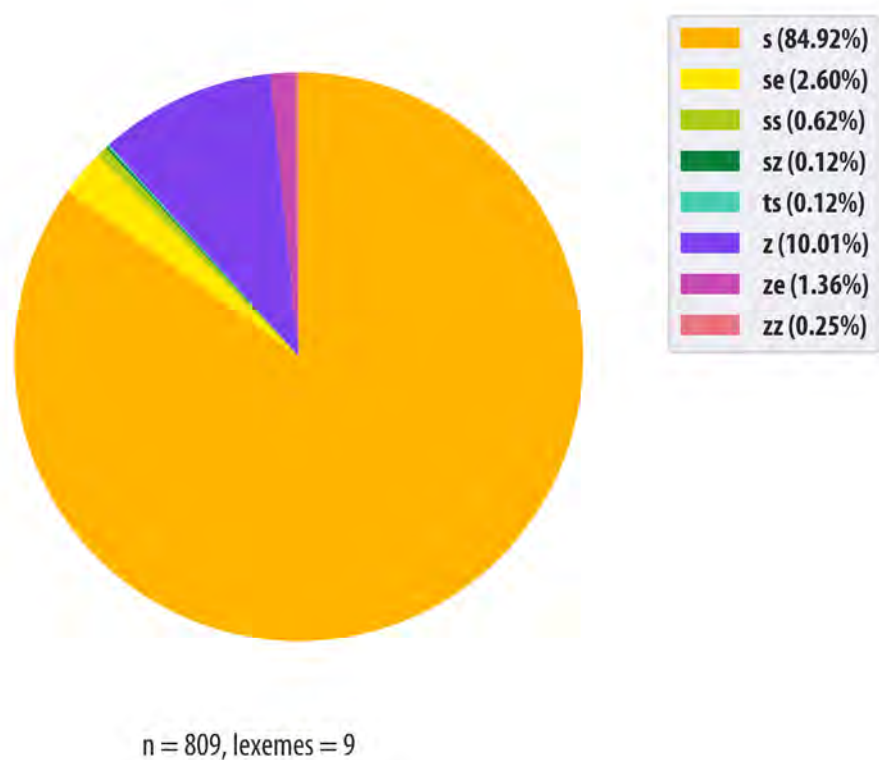


(d) Twitter (official) data

Figure C.44: Distribution of graphemes for phoneme /s/ (cont.)

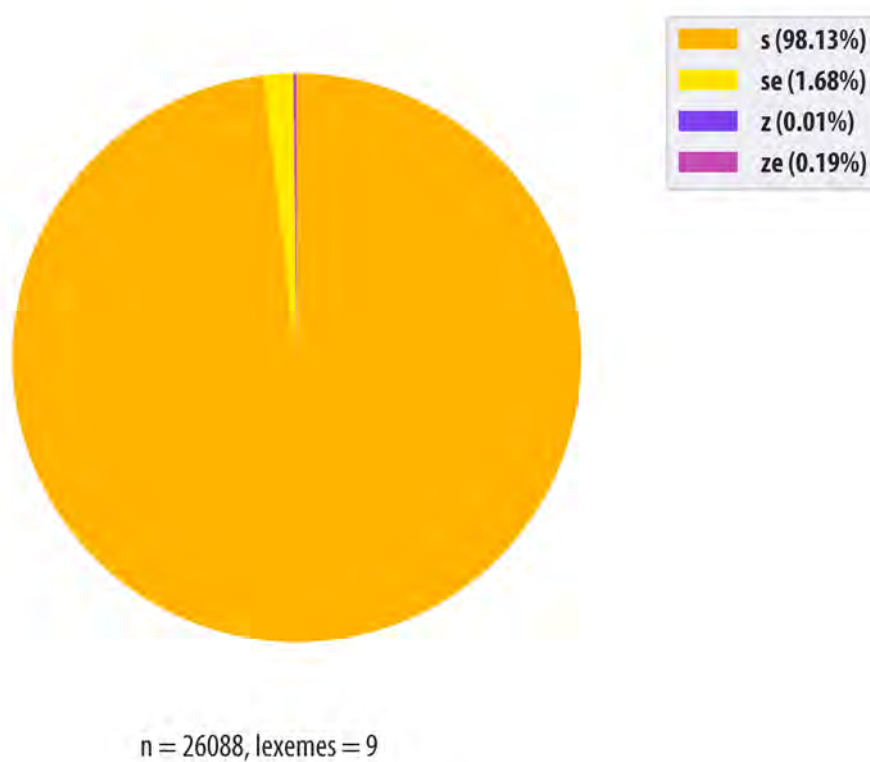


(a) GPC data

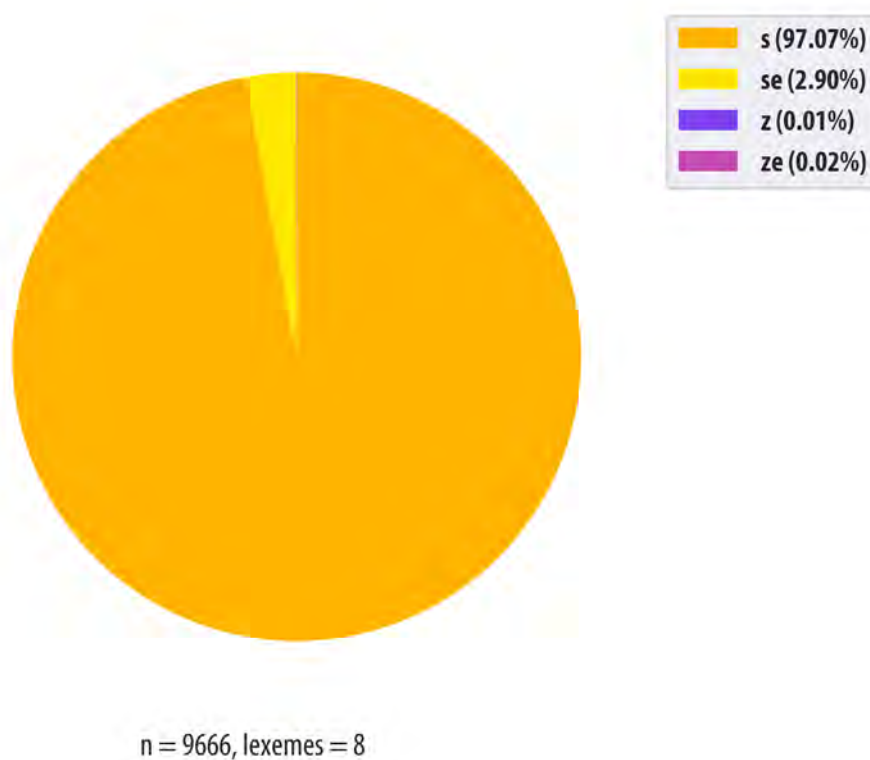


(b) Survey data

Figure C.45: Distribution of graphemes for phoneme /z/

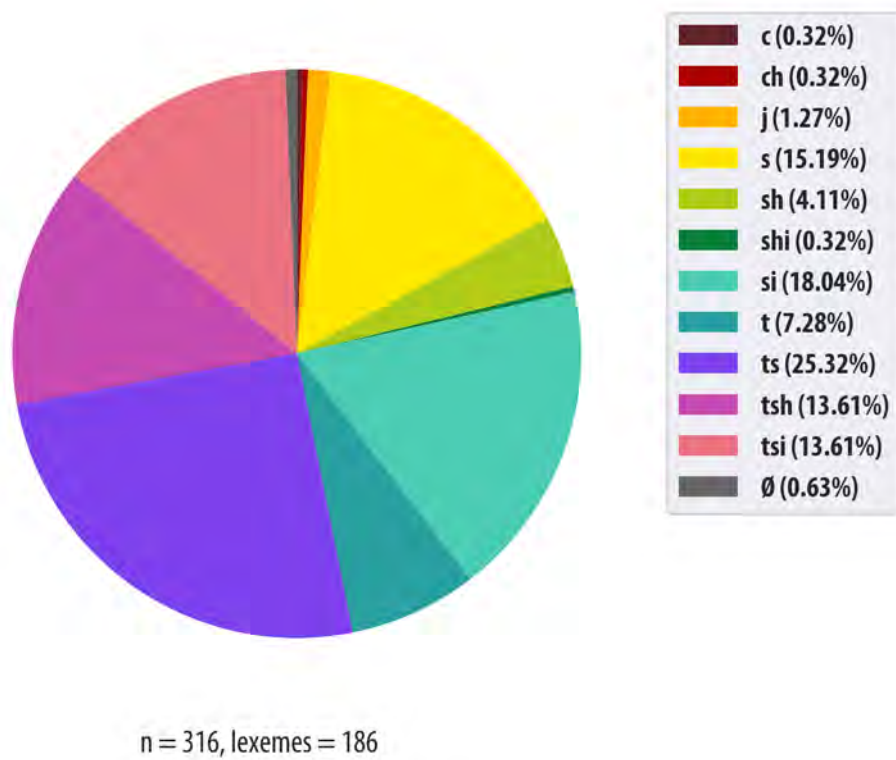


(c) Twitter (personal) data

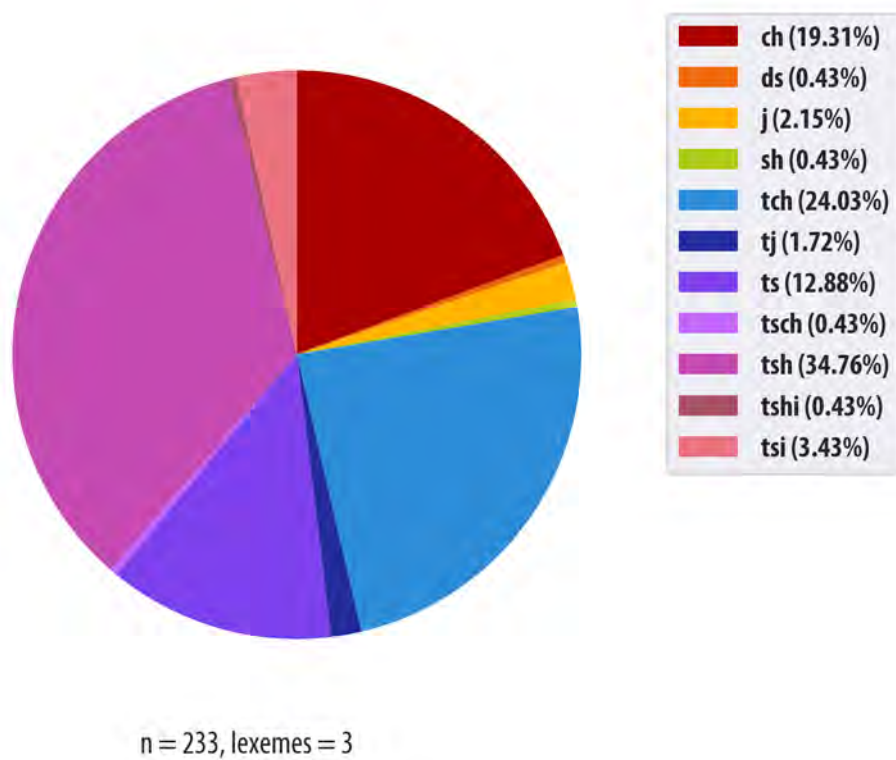


(d) Twitter (official) data

Figure C.45: Distribution of graphemes for phoneme /z/ (cont.)

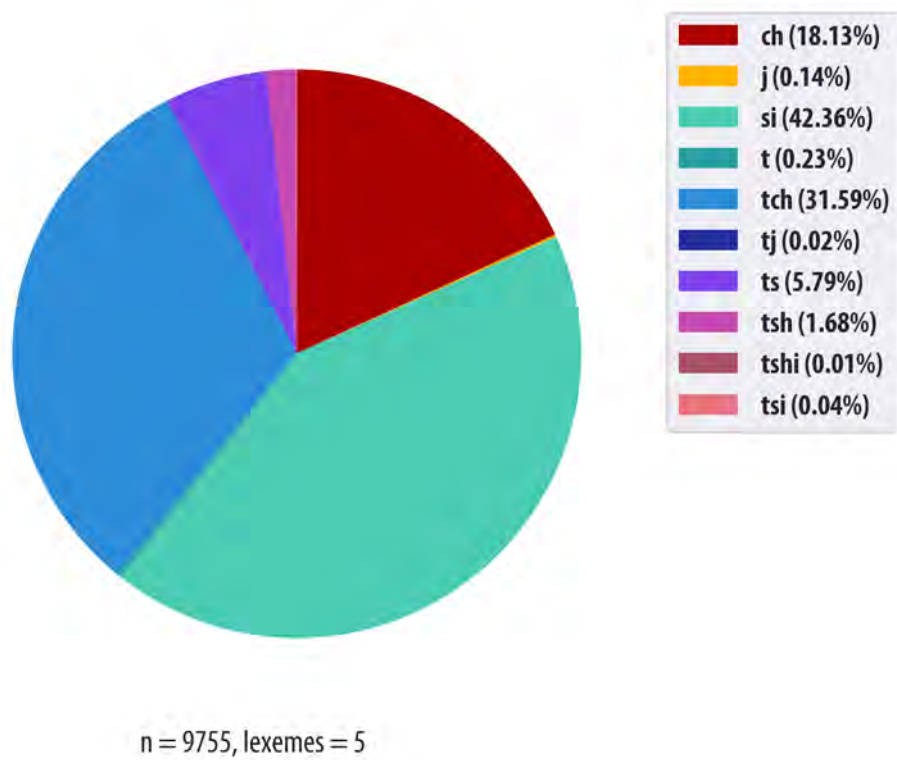


(a) GPC data

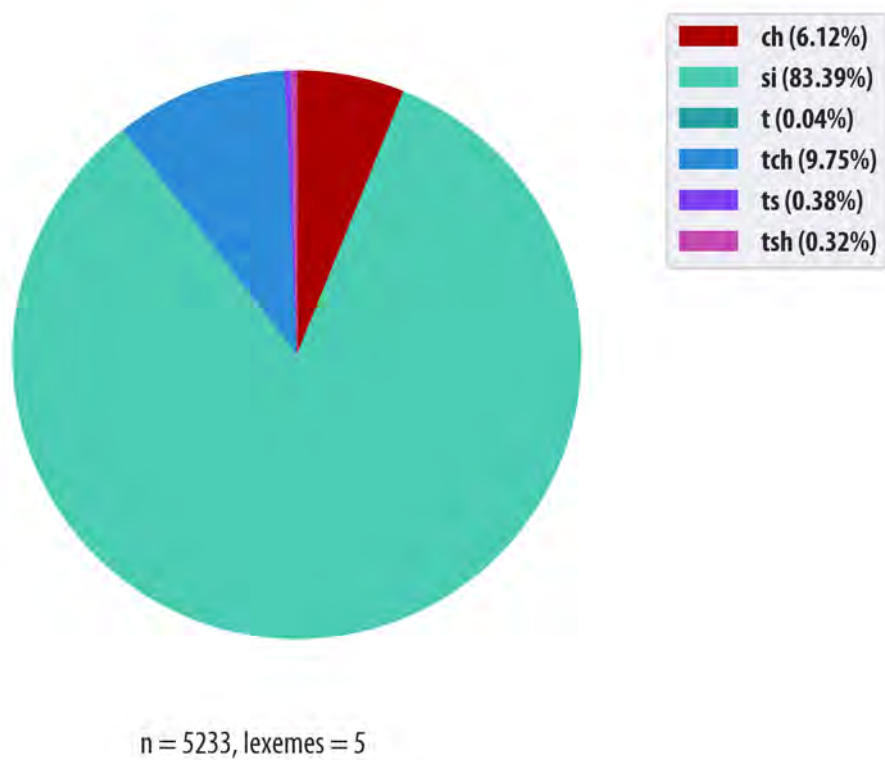


(b) Survey data

Figure C.46: Distribution of graphemes for phoneme /tʃ/

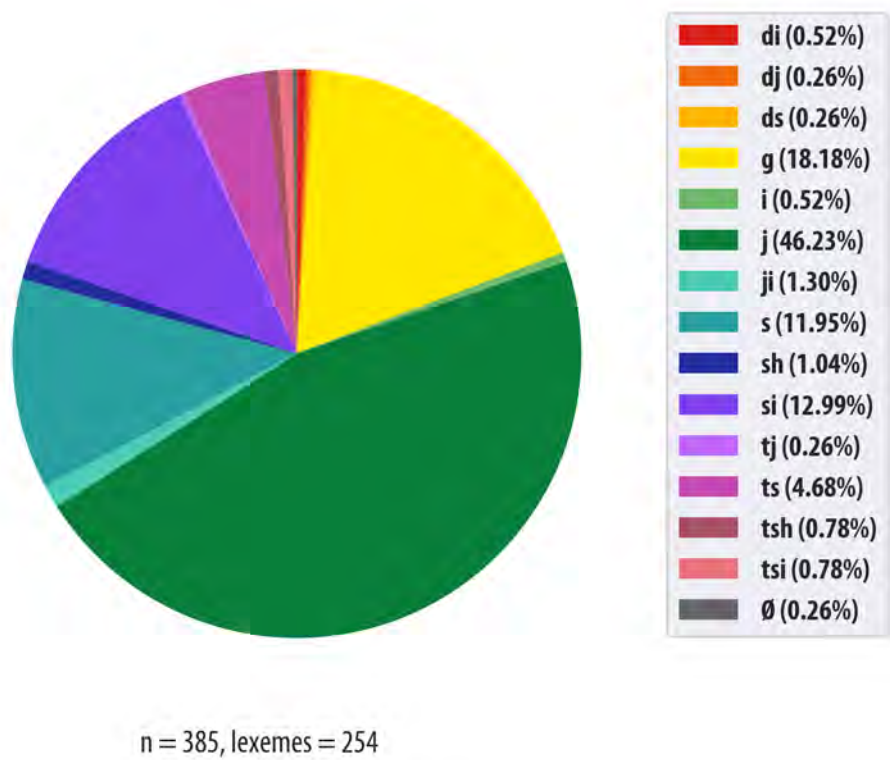


(c) Twitter (personal) data

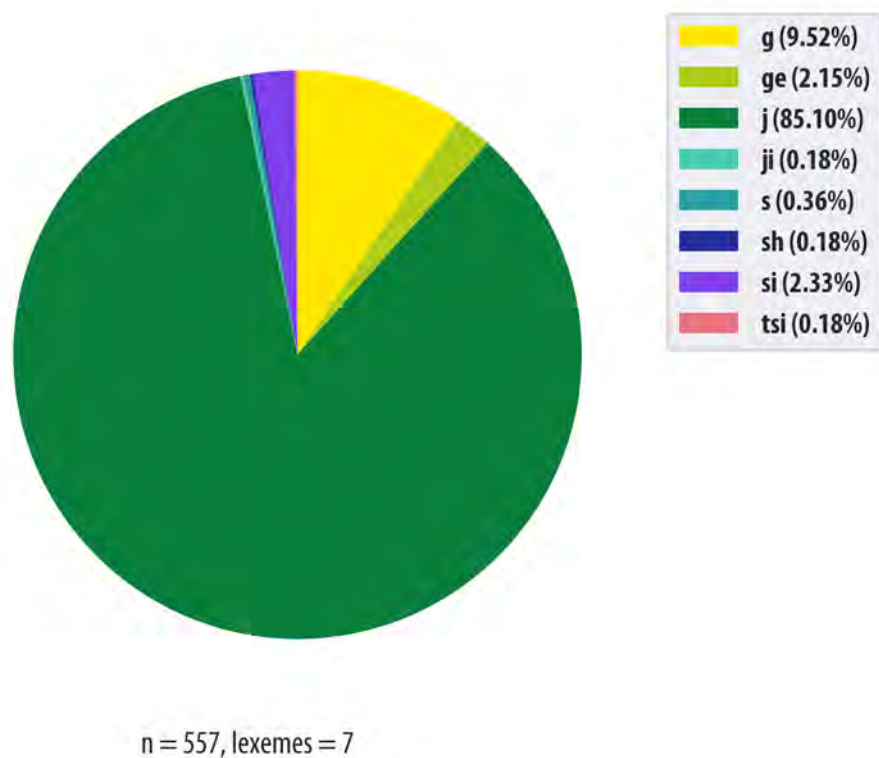


(d) Twitter (official) data

Figure C.46: Distribution of graphemes for phoneme /tʃ/ (cont.)

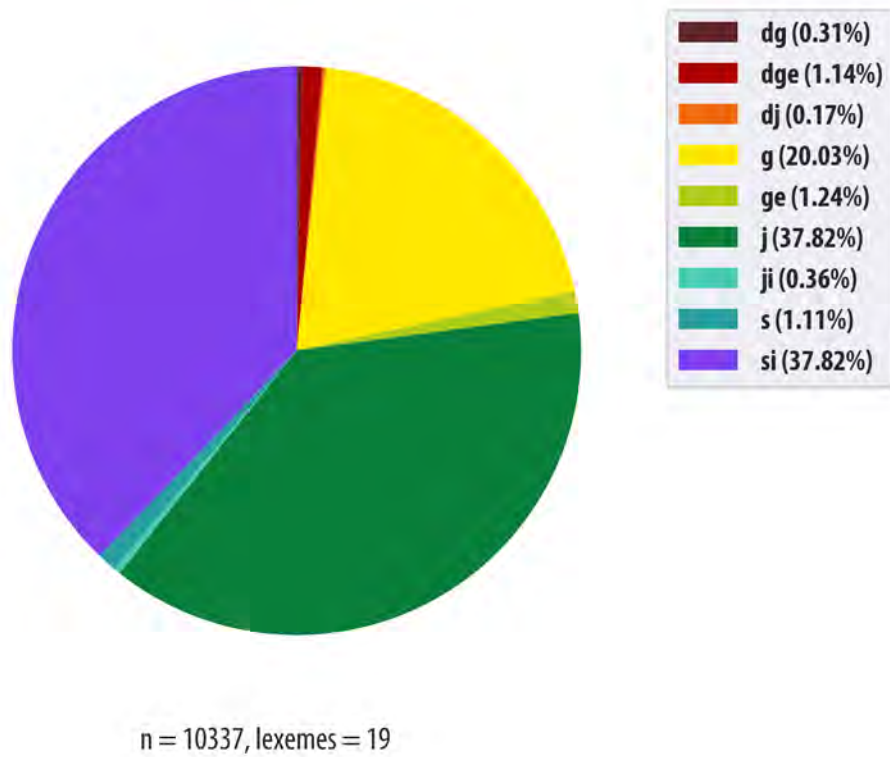


(a) GPC data

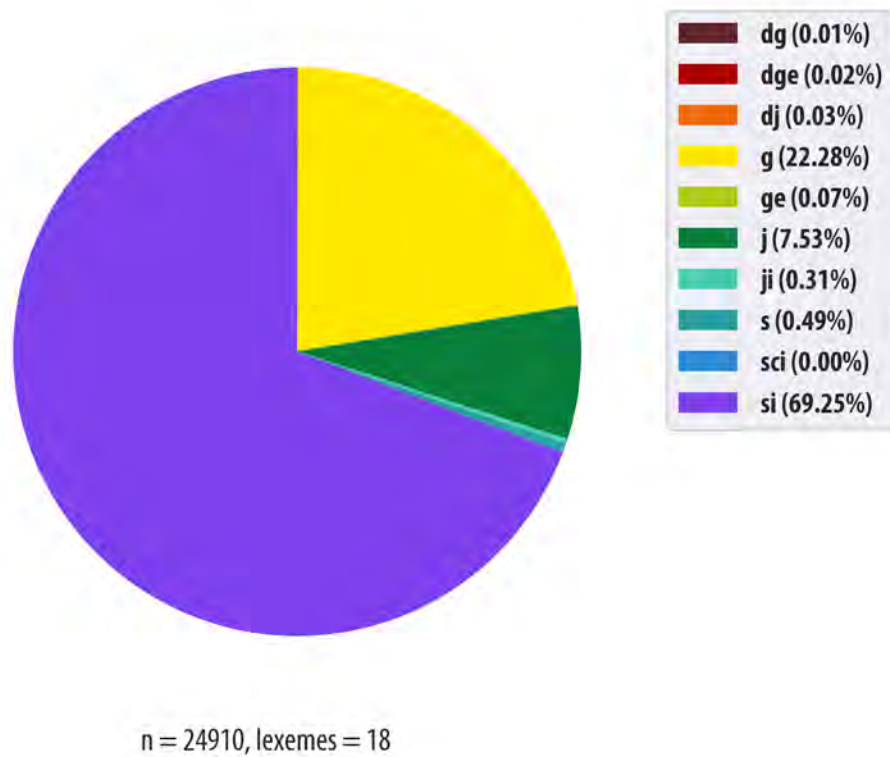


(b) Survey data

Figure C.47: Distribution of graphemes for phoneme /dʒ/

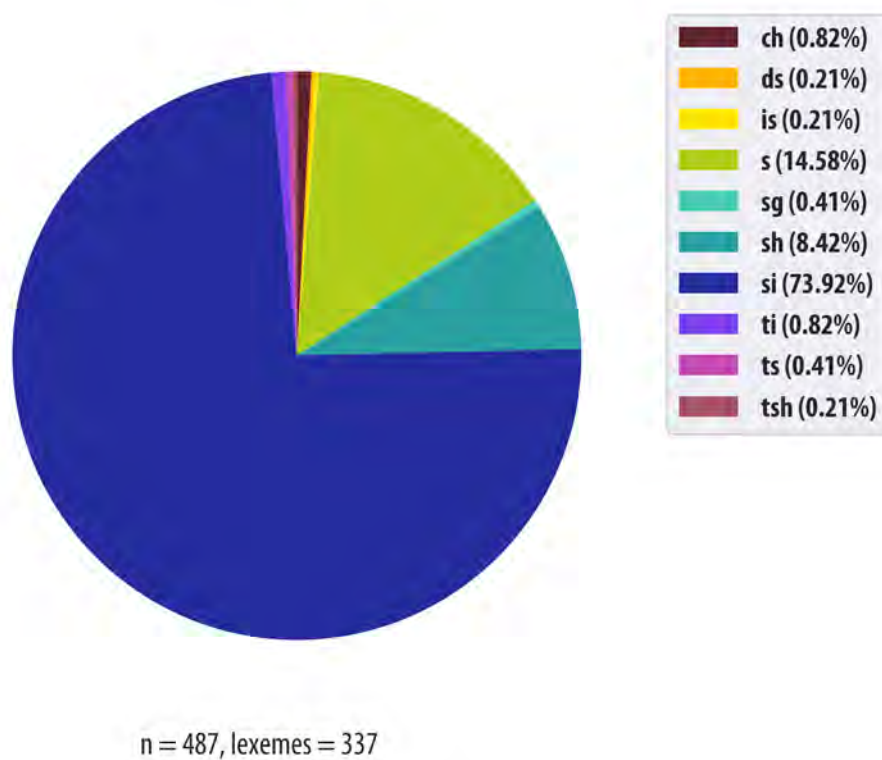


(c) Twitter (personal) data

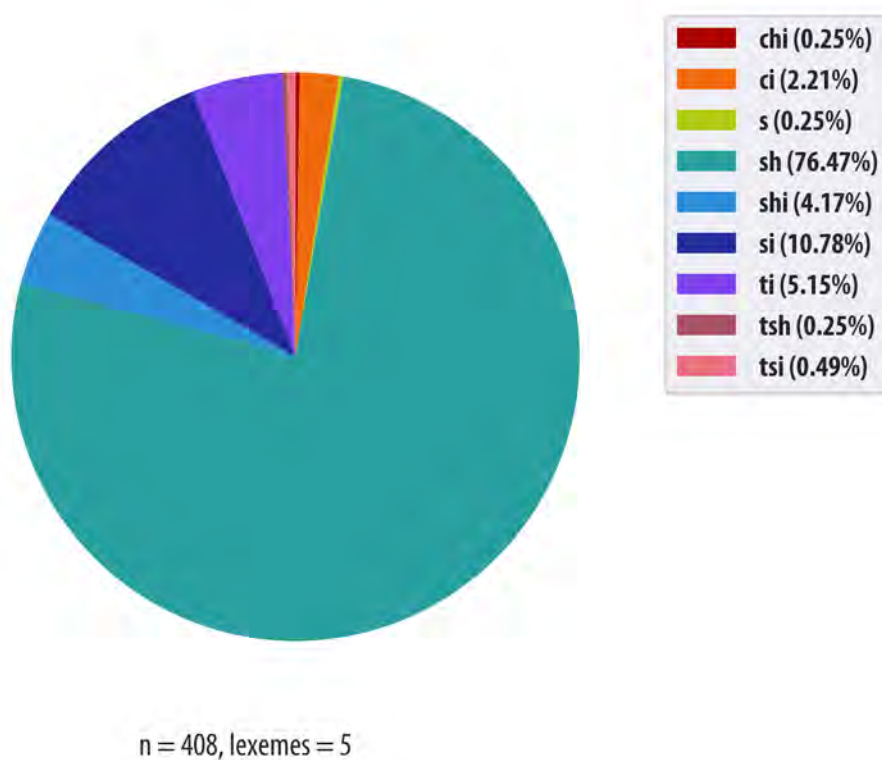


(d) Twitter (official) data

Figure C.47: Distribution of graphemes for phoneme /dʒ/ (cont.)

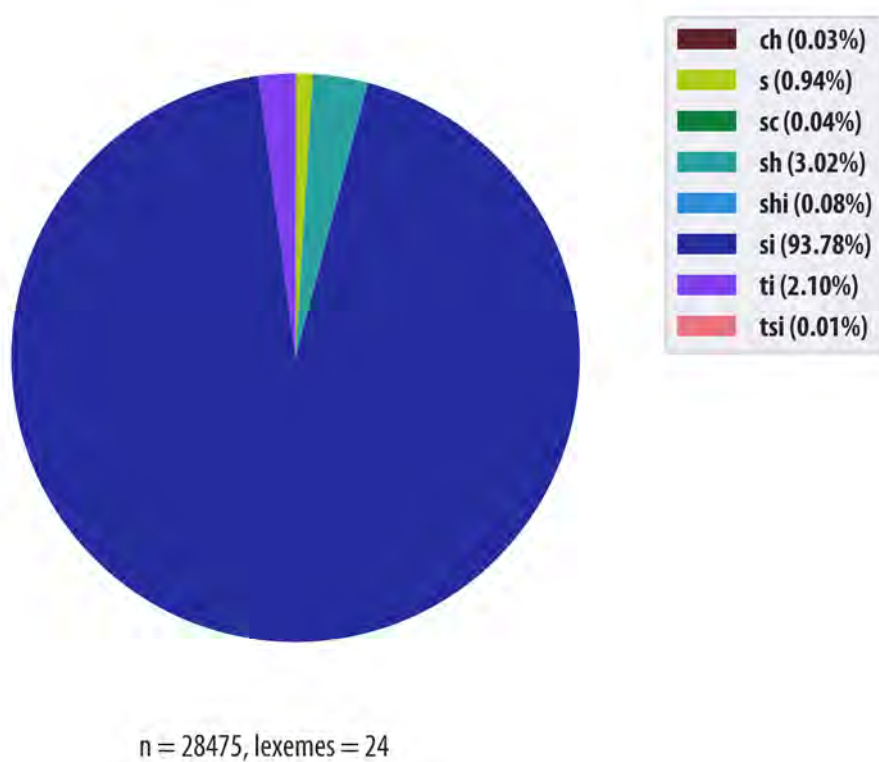


(a) GPC data

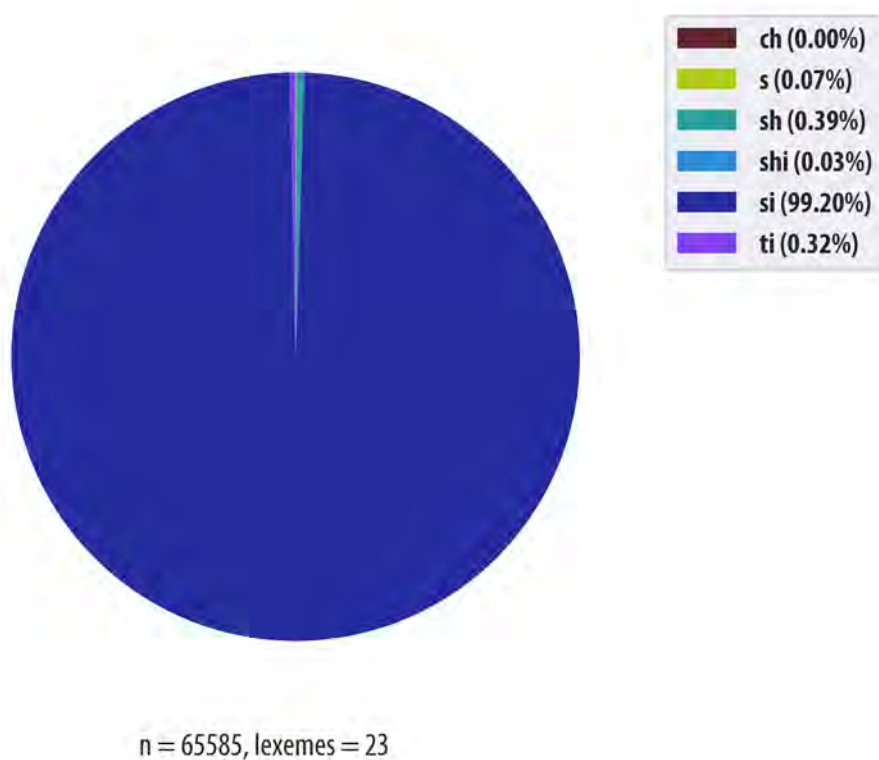


(b) Survey data

Figure C.48: Distribution of graphemes for phoneme /ʃ/

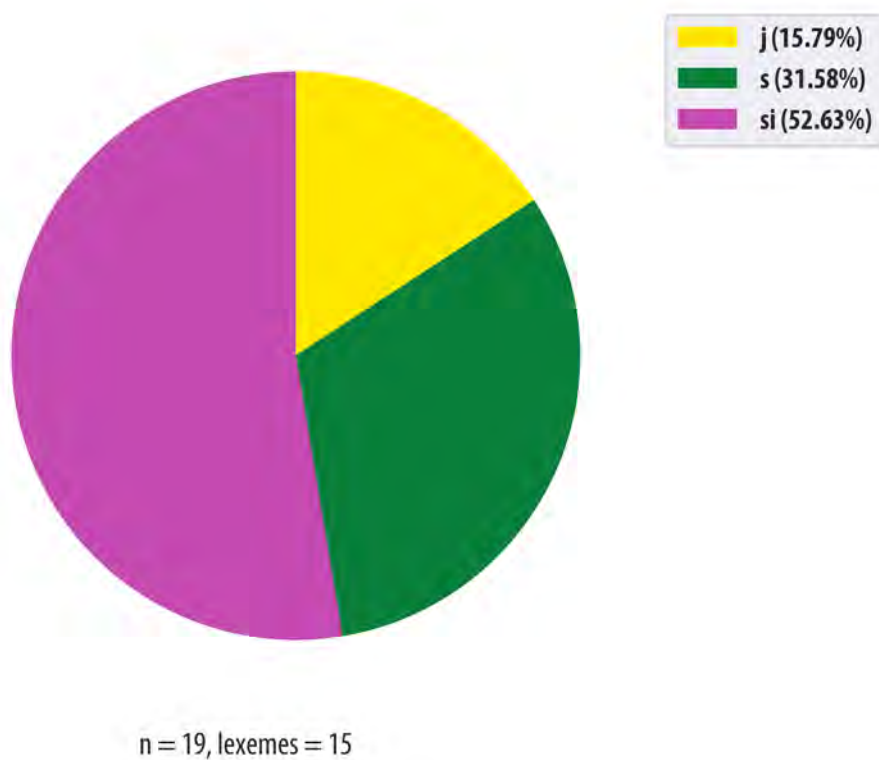


(c) Twitter (personal) data

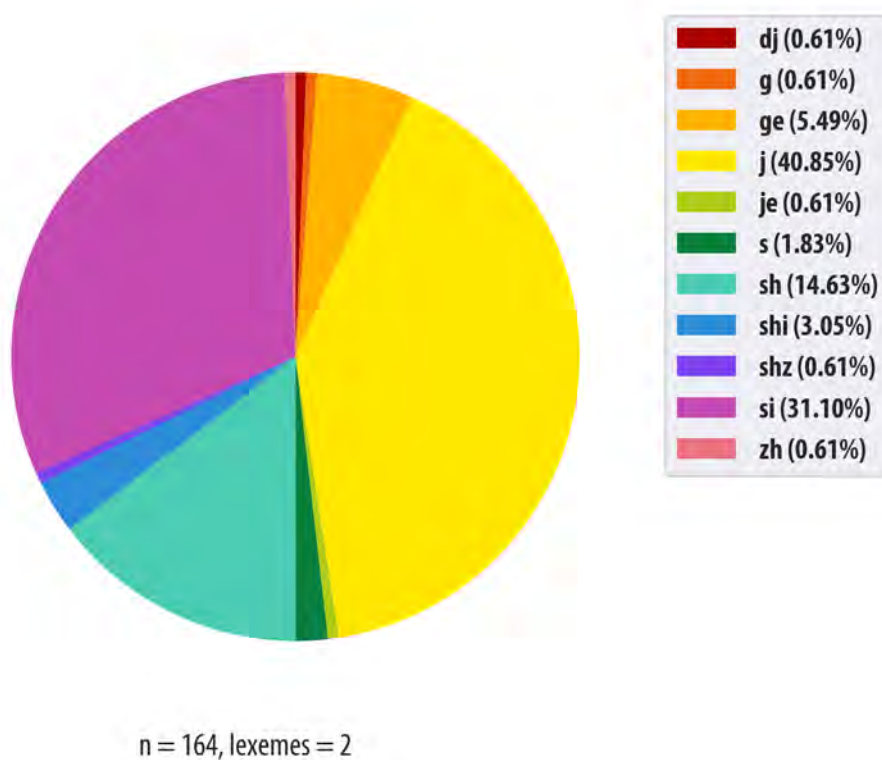


(d) Twitter (official) data

Figure C.48: Distribution of graphemes for phoneme /ʃ/ (cont.)

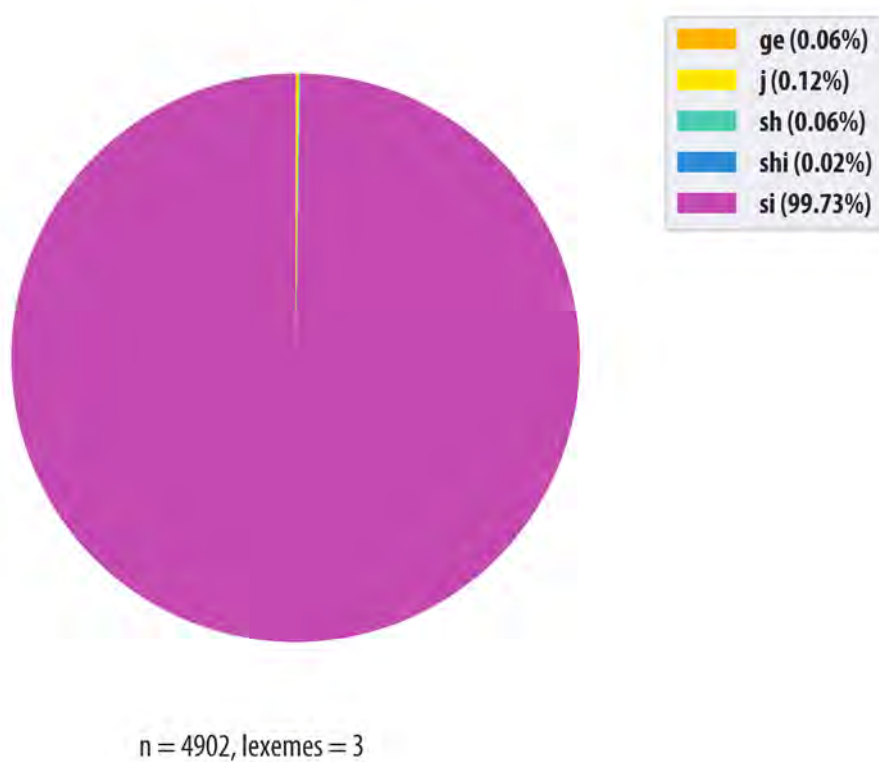


(a) GPC data

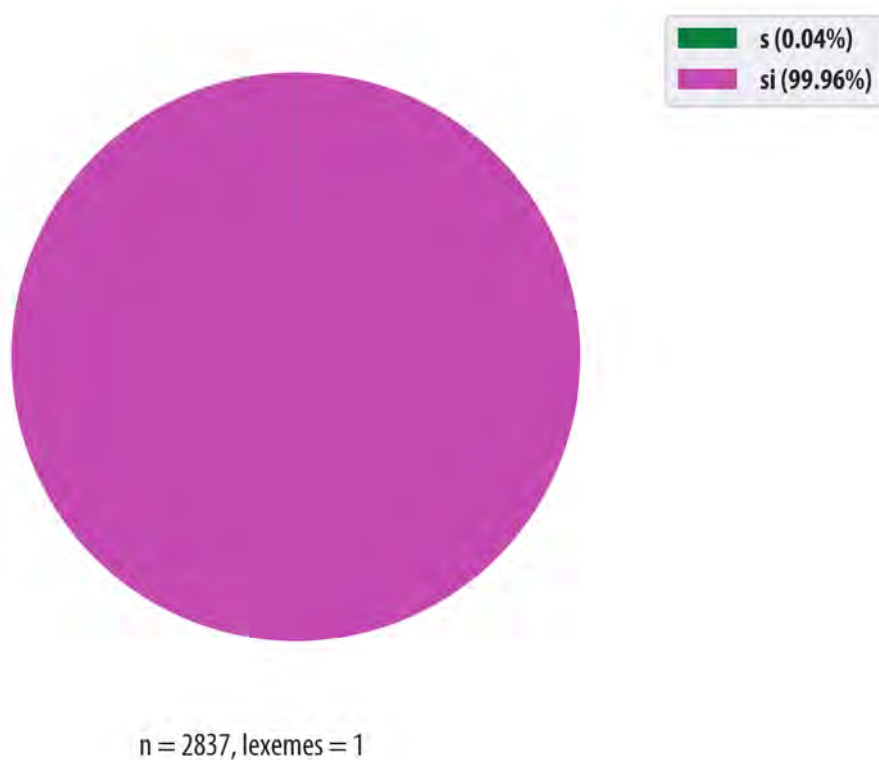


(b) Survey data

Figure C.49: Distribution of graphemes for phoneme /ʒ/

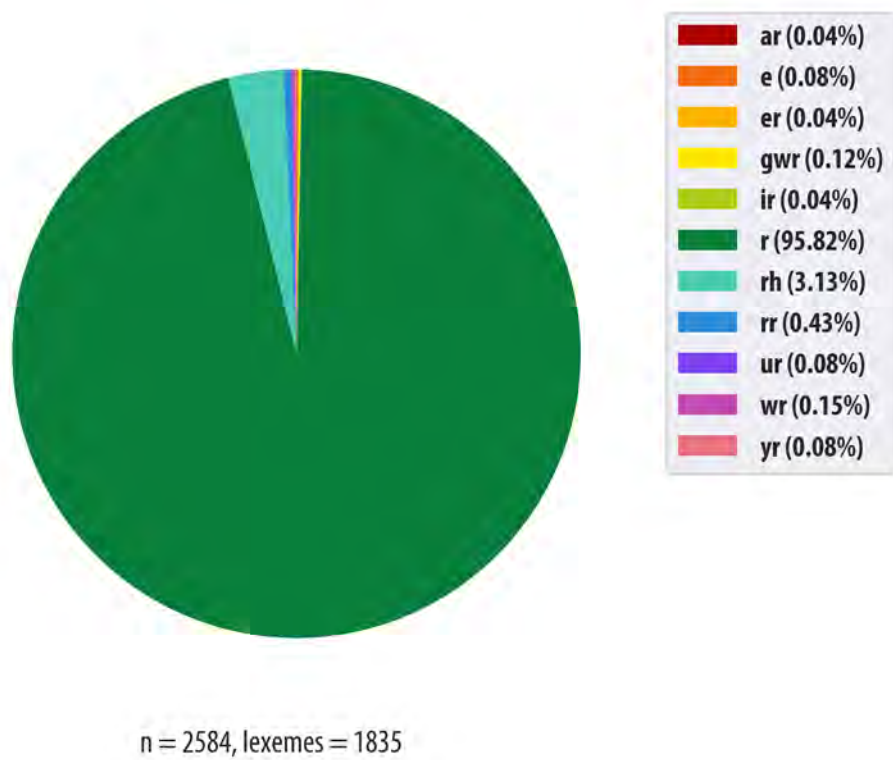


(c) Twitter (personal) data

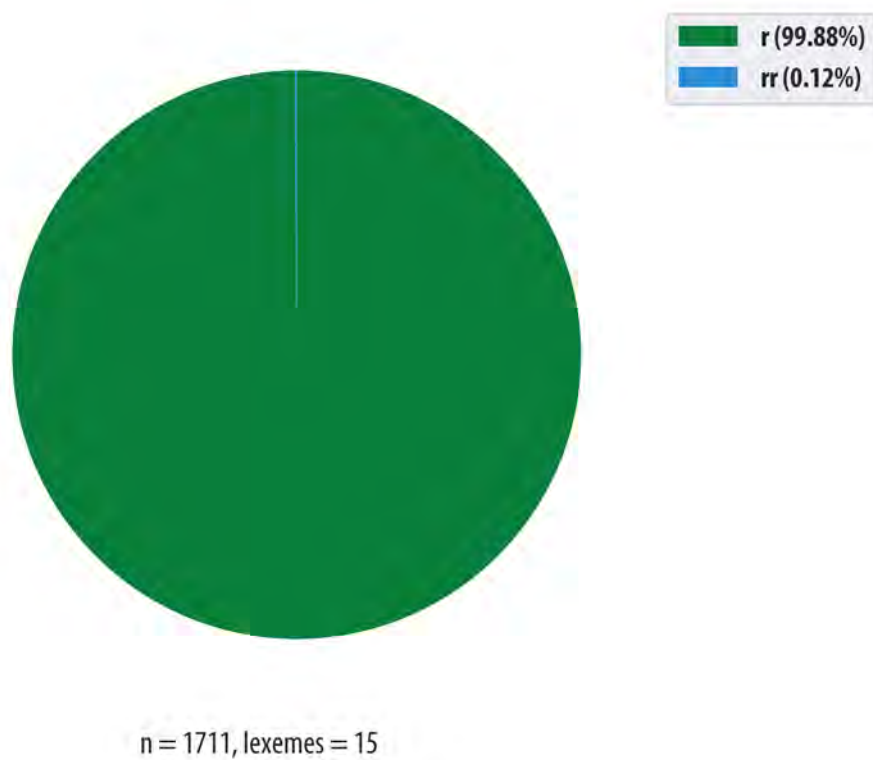


(d) Twitter (official) data

Figure C.49: Distribution of graphemes for phoneme /ʒ/ (cont.)

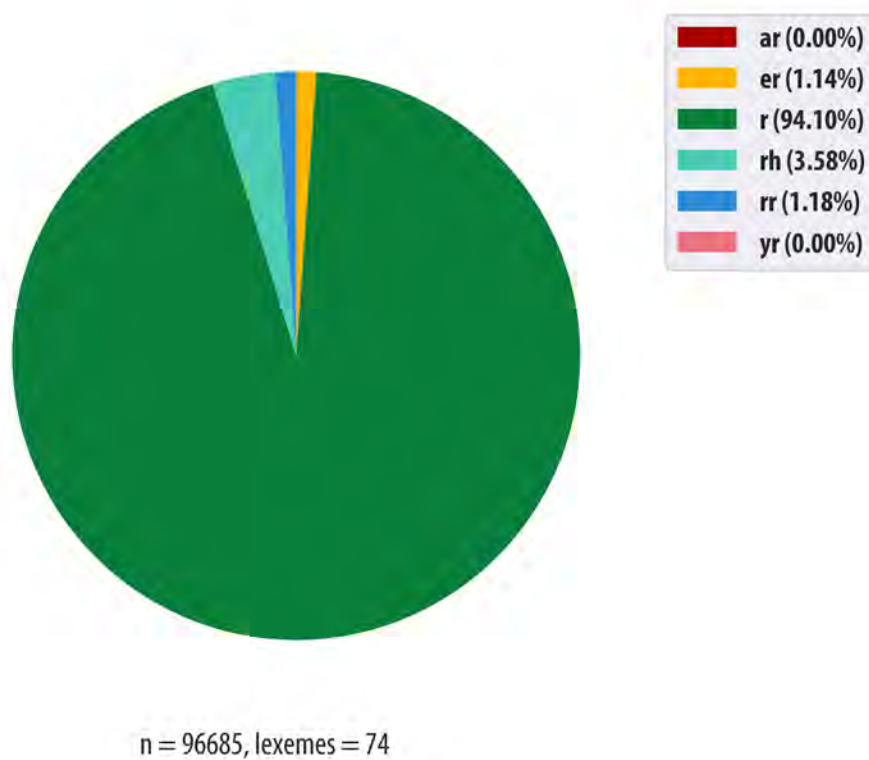


(a) GPC data

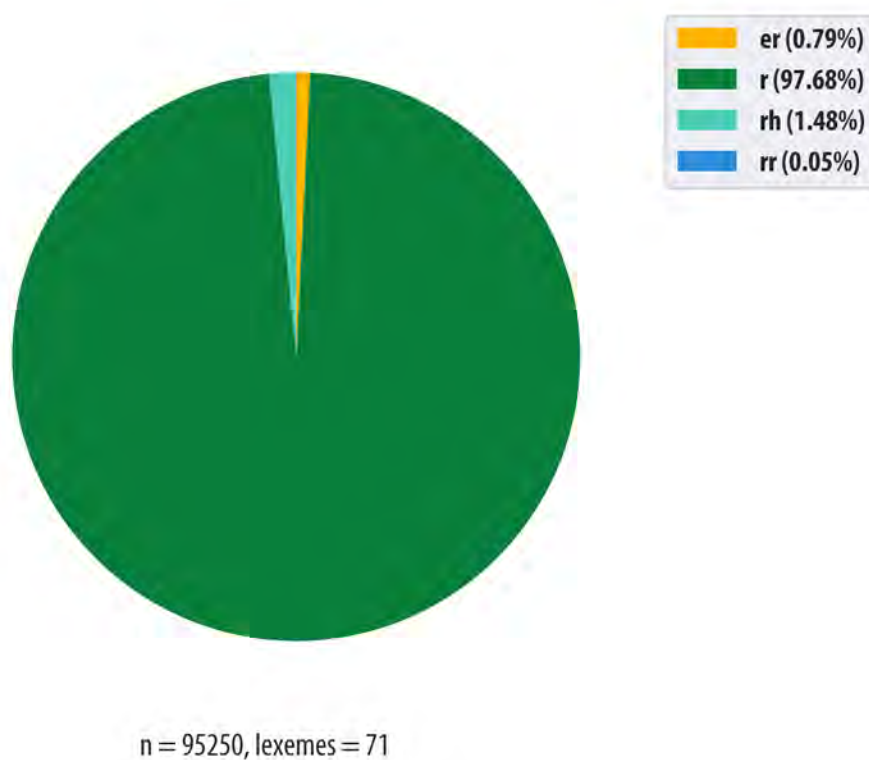


(b) Survey data

Figure C.50: Distribution of graphemes for phoneme /ɹ/

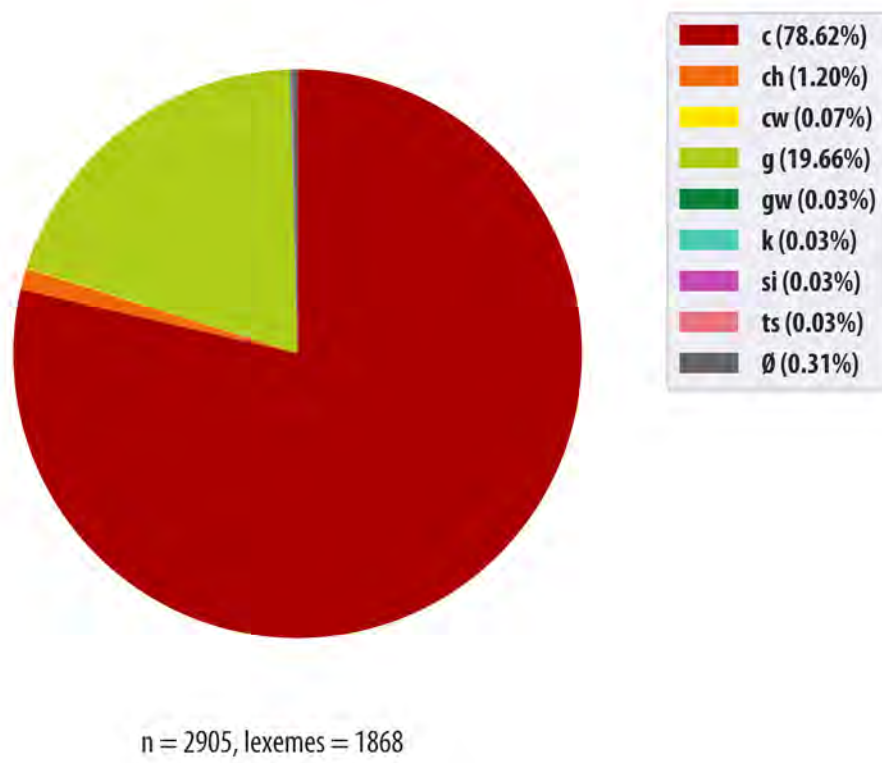


(c) Twitter (personal) data

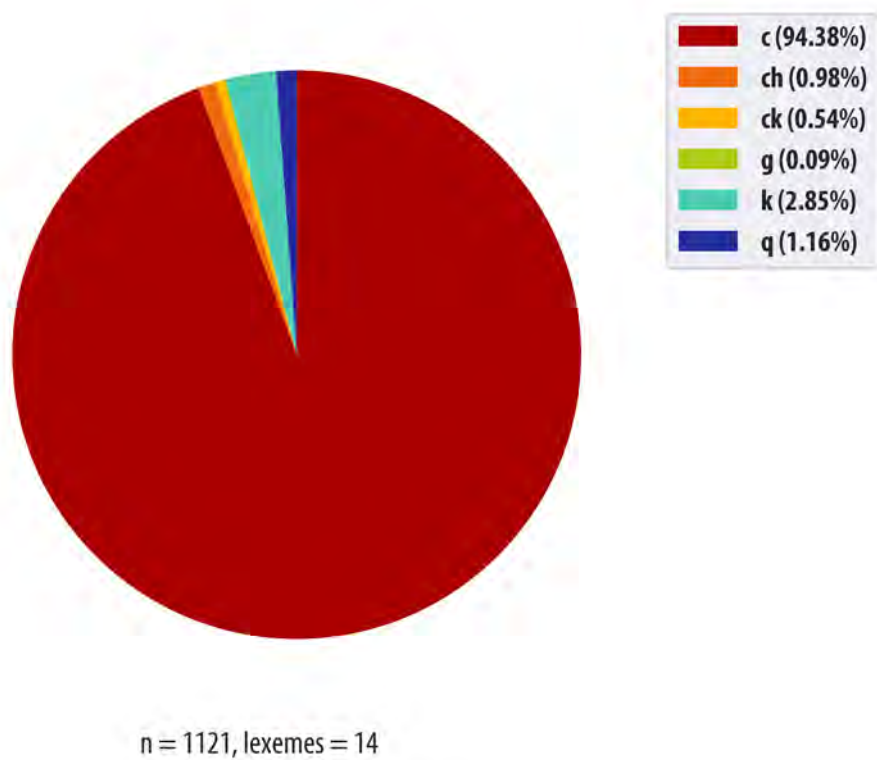


(d) Twitter (official) data

Figure C.50: Distribution of graphemes for phoneme /ɹ/ (cont.)

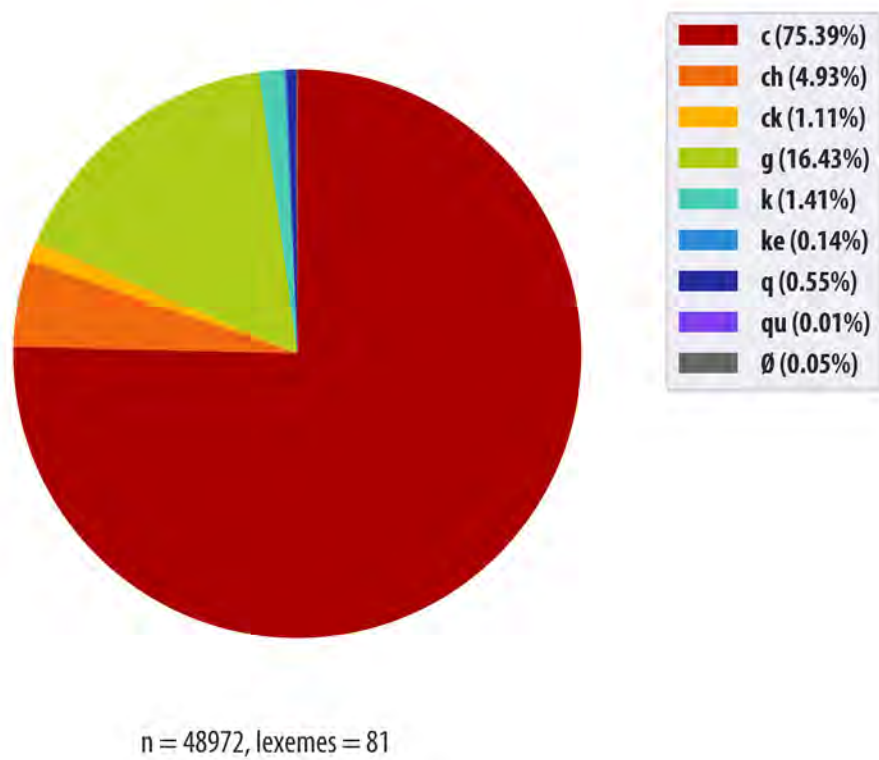


(a) GPC data

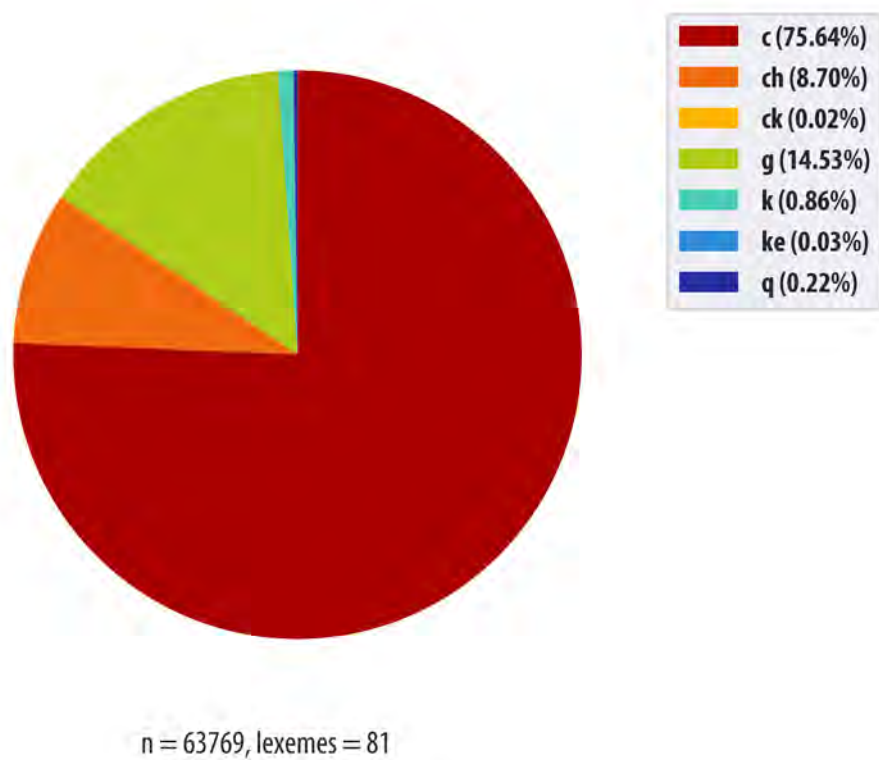


(b) Survey data

Figure C.51: Distribution of graphemes for phoneme /k/

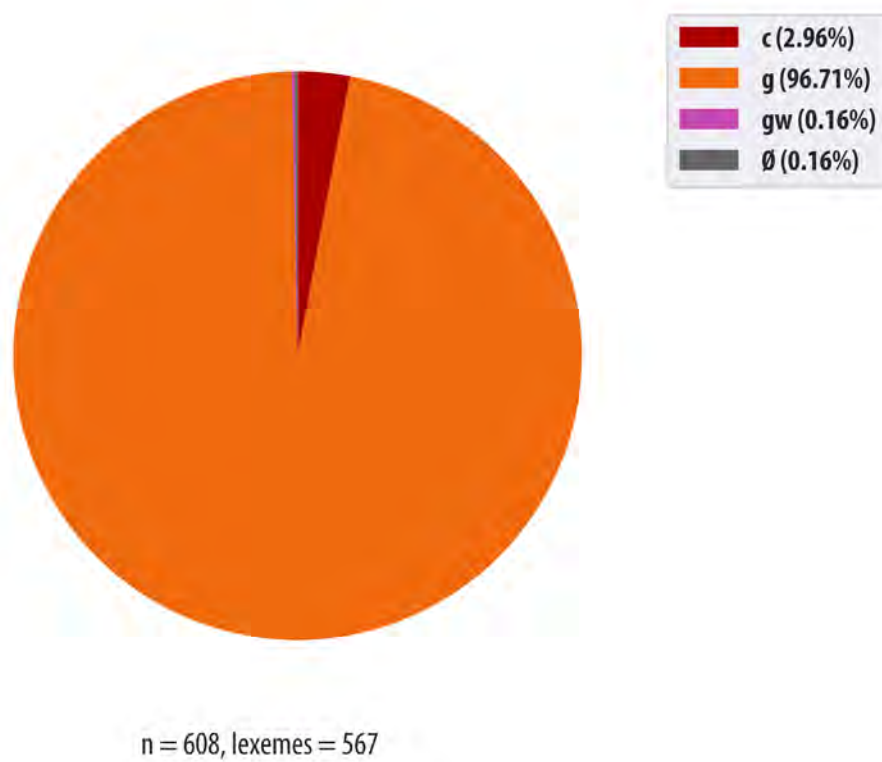


(c) Twitter (personal) data

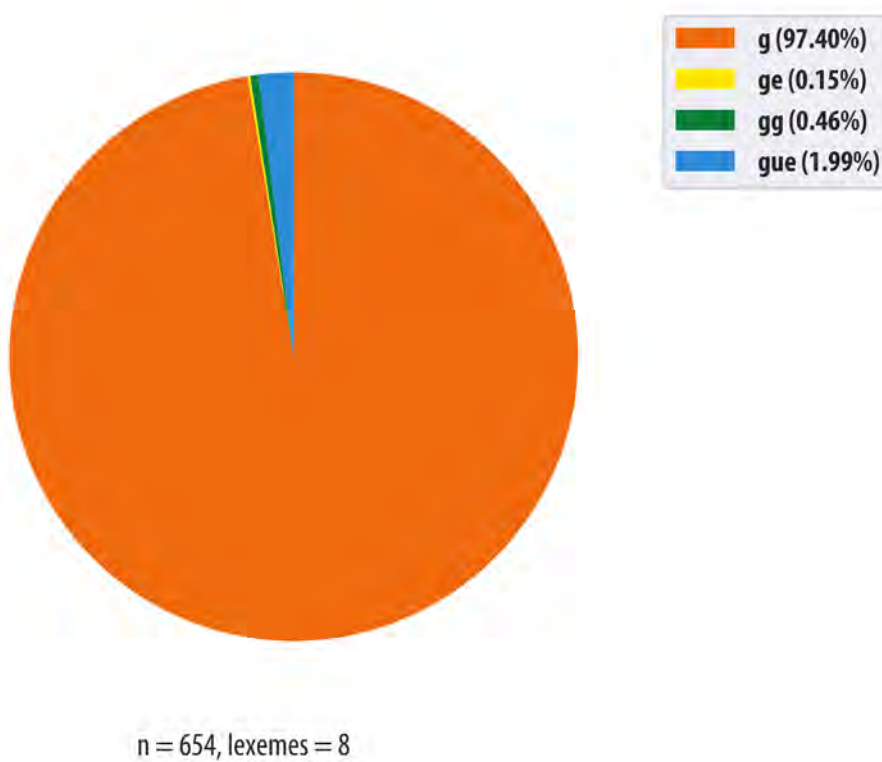


(d) Twitter (official) data

Figure C.51: Distribution of graphemes for phoneme /k/ (cont.)

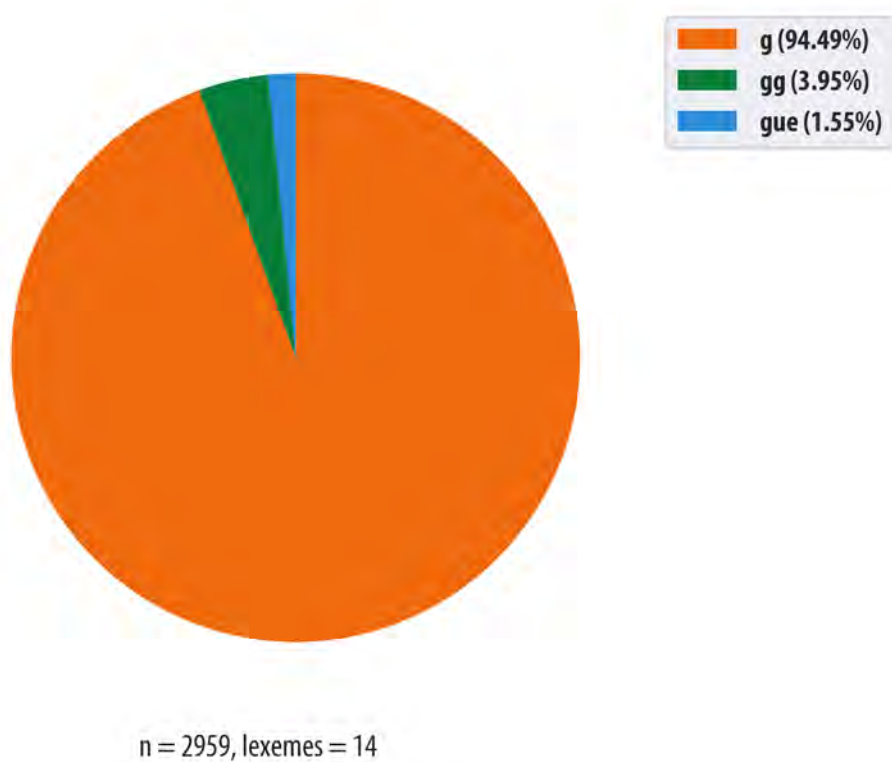


(a) GPC data

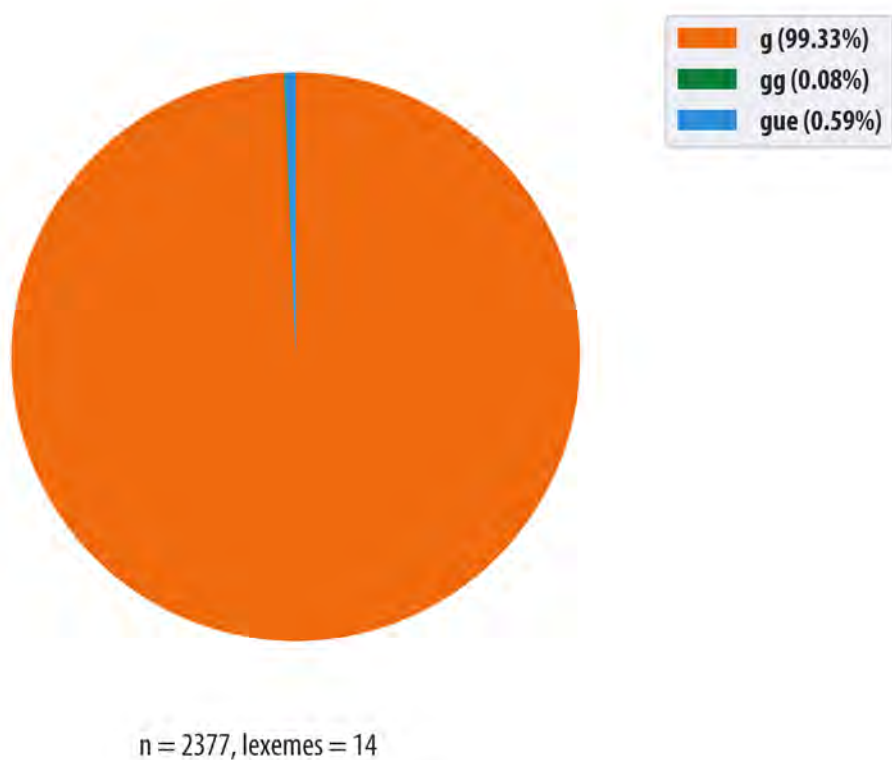


(b) Survey data

Figure C.52: Distribution of graphemes for phoneme /g/

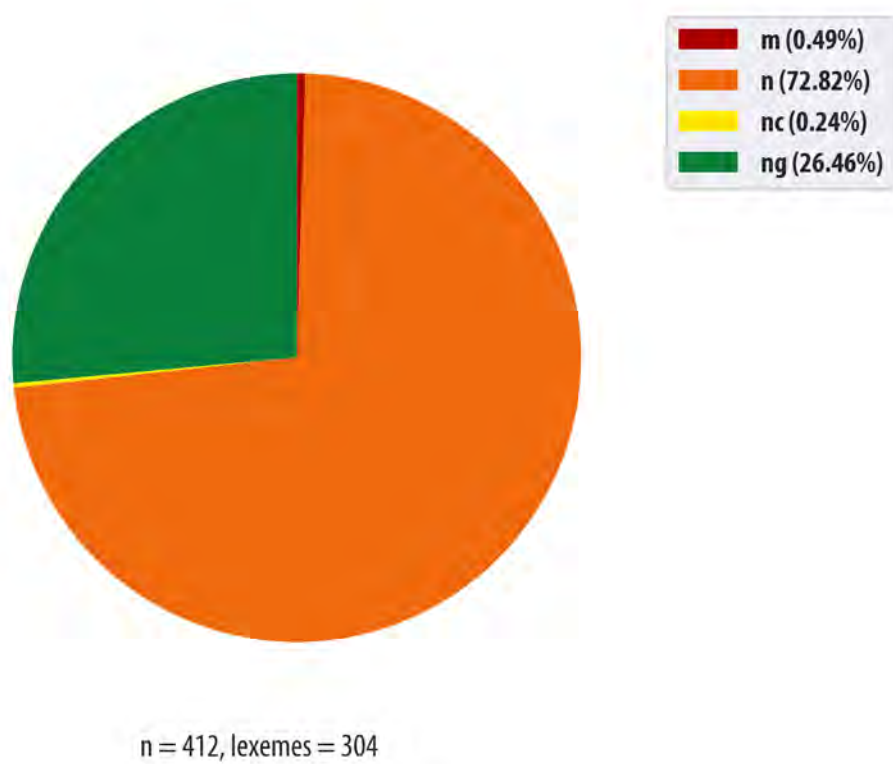


(c) Twitter (personal) data

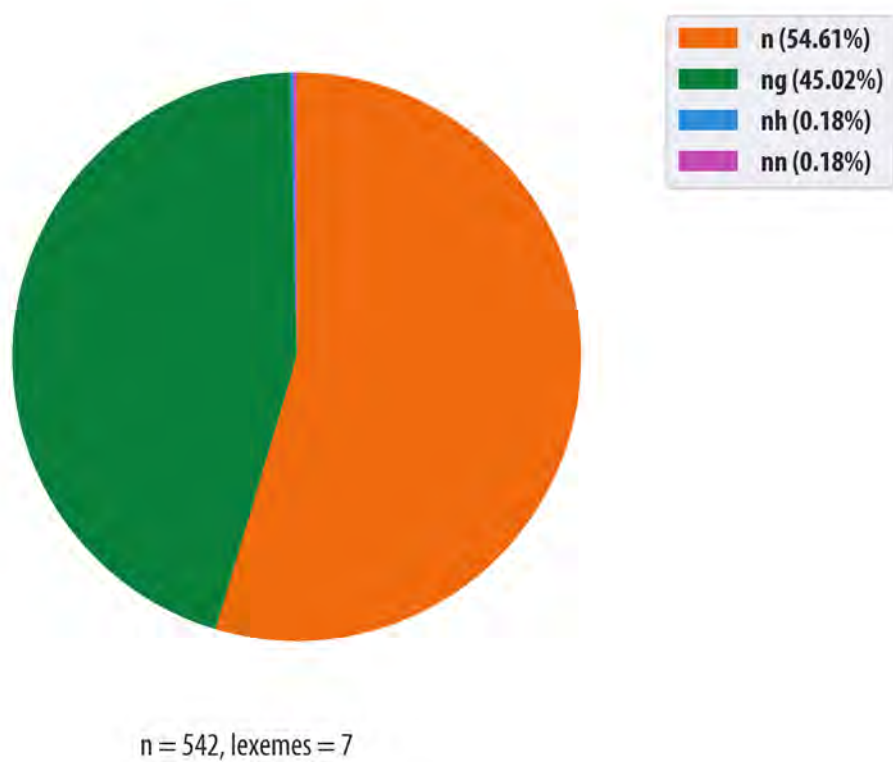


(d) Twitter (official) data

Figure C.52: Distribution of graphemes for phoneme /g/ (cont.)

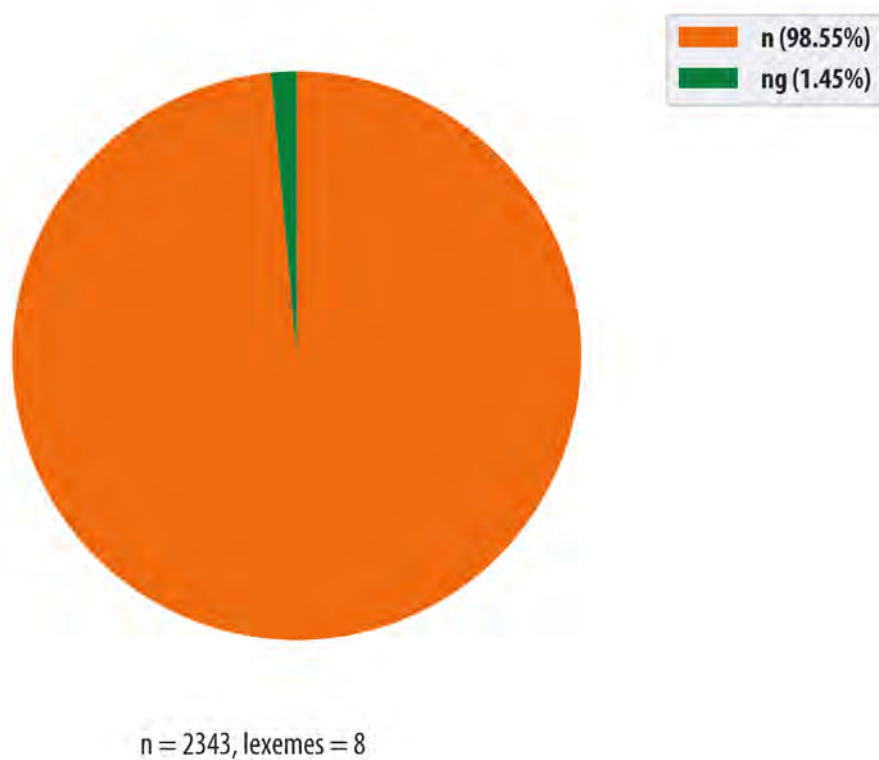


(a) GPC data

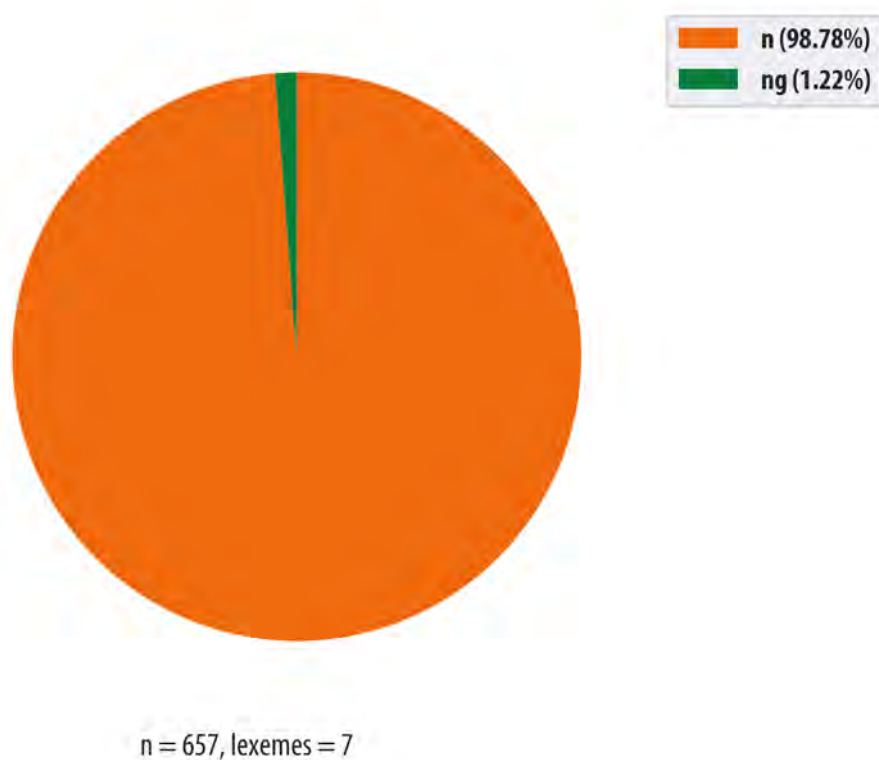


(b) Survey data

Figure C.53: Distribution of graphemes for phoneme /ŋ/

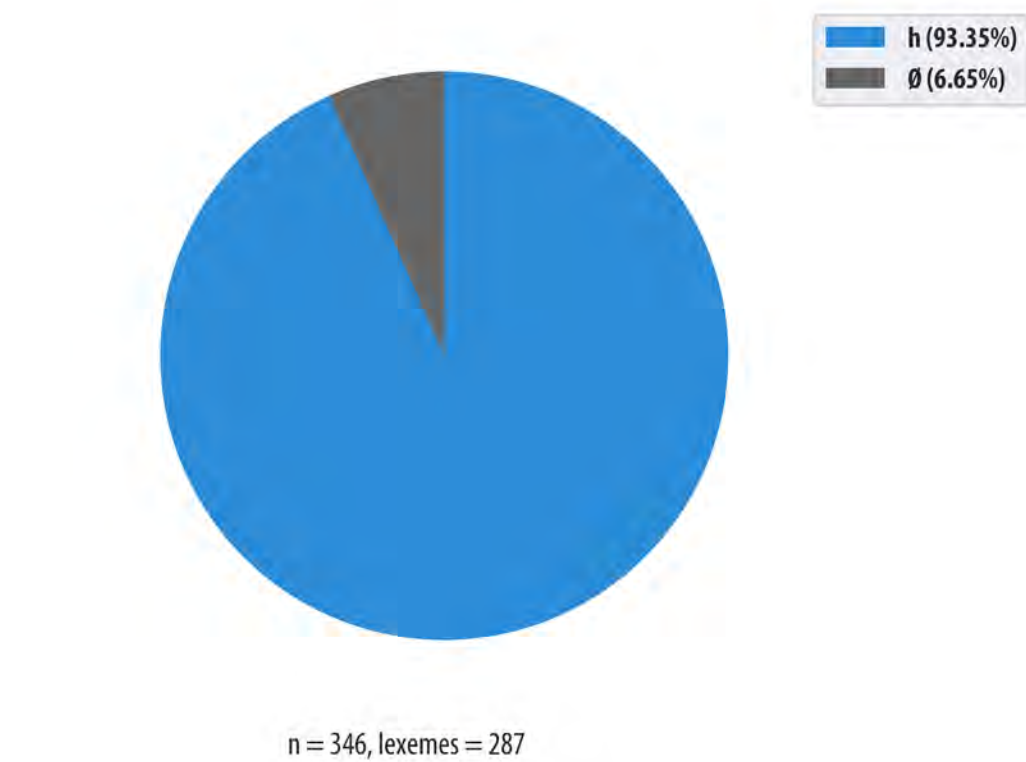


(c) Twitter (personal) data

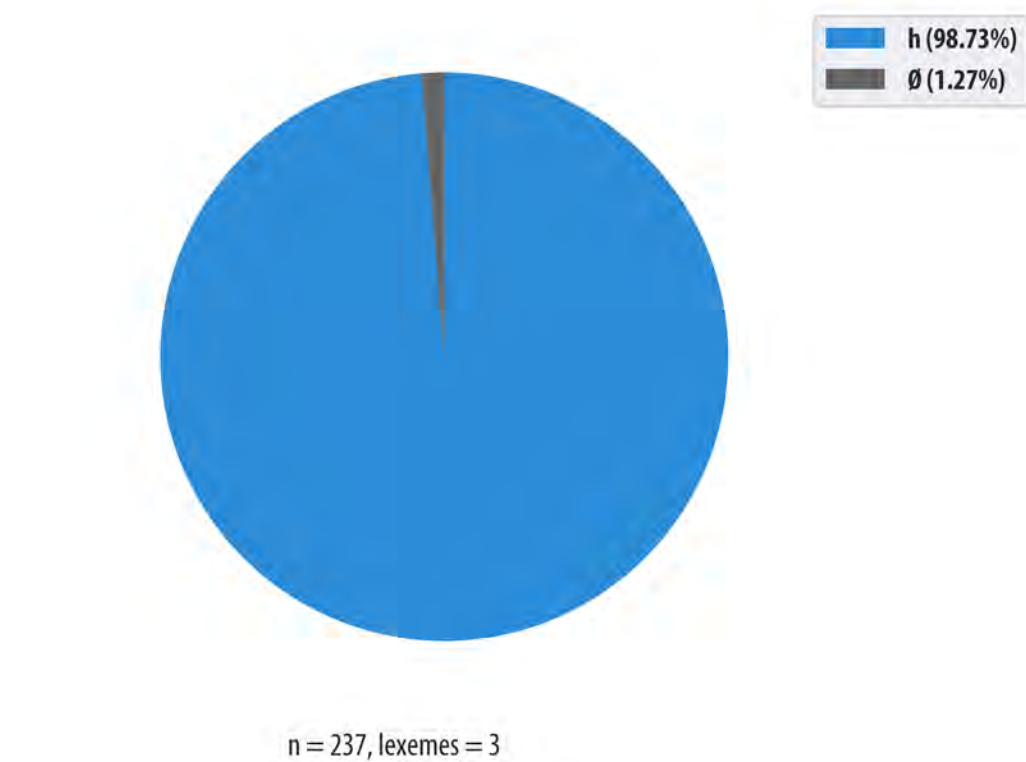


(d) Twitter (official) data

Figure C.53: Distribution of graphemes for phoneme /ŋ/ (cont.)

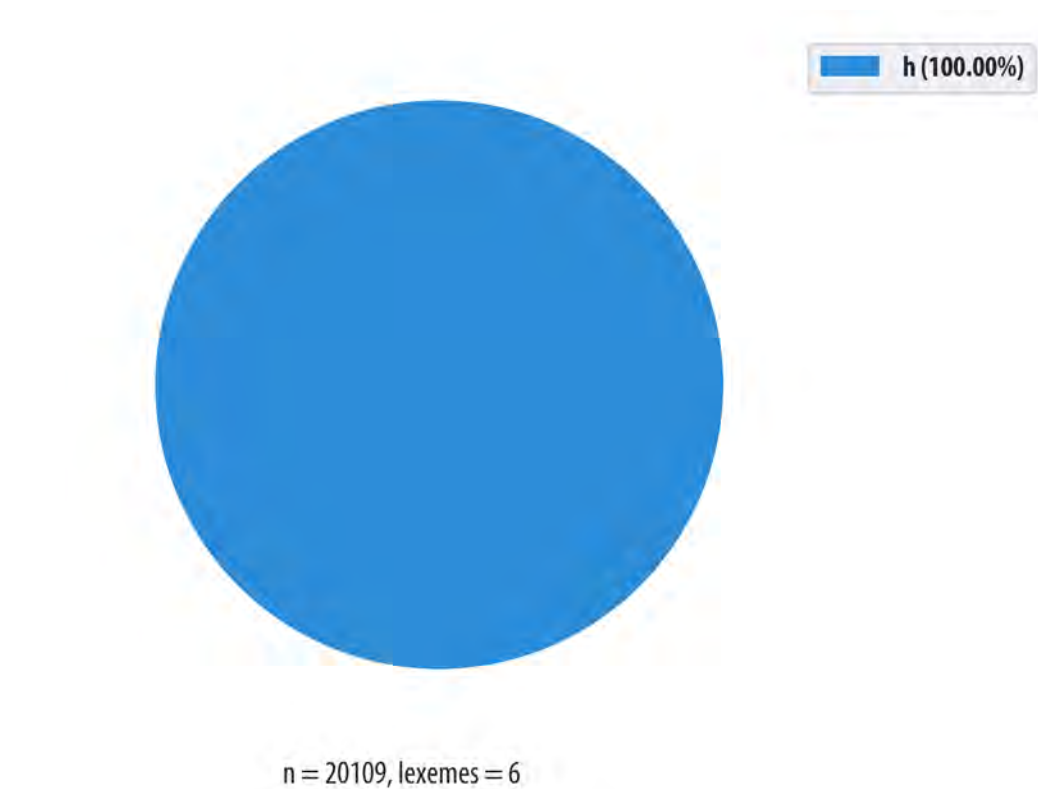


(a) GPC data

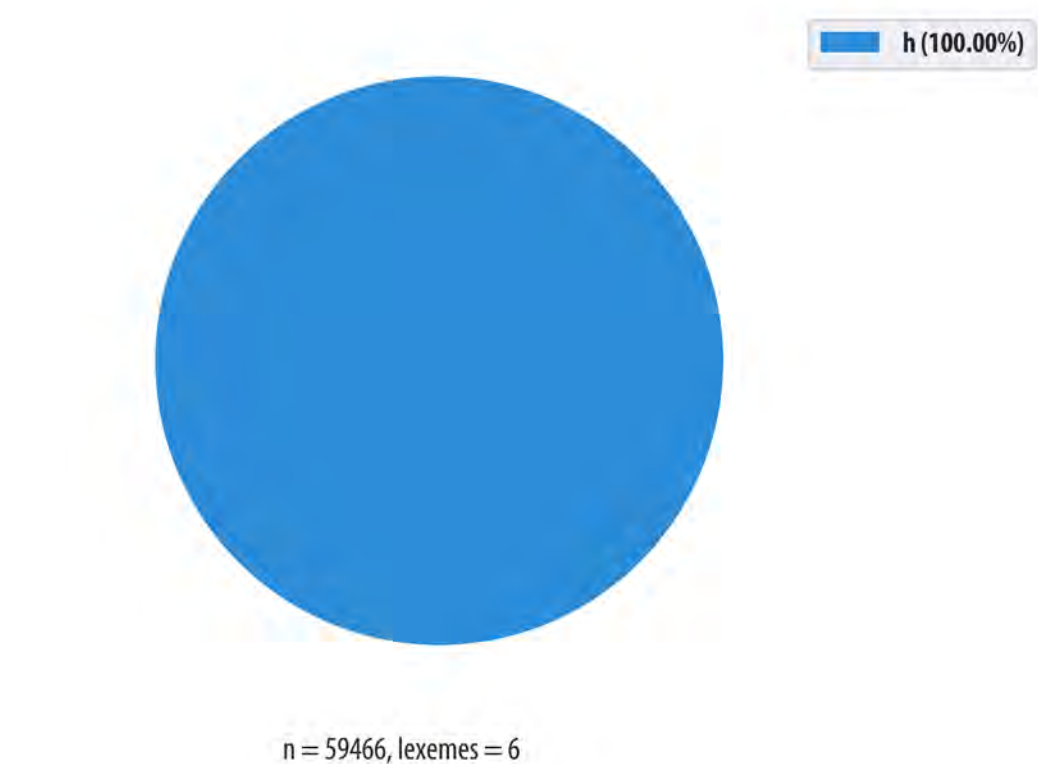


(b) Survey data

Figure C.54: Distribution of graphemes for phoneme /h/



(c) Twitter (personal) data



(d) Twitter (official) data

Figure C.54: Distribution of graphemes for phoneme /h/ (cont.)

Appendix D

Loanwords in Geiriadur Prifysgol Cymru

D.i GPC: summary statistics

Table D.1: GPC – summary statistics

Number of variants	Headword count
1	5201
2	1240
3	251
4	84
5	27
6	8

Table D.2: Phoneme totals and numbers of unique grapheme pairs in adapted forms and Standard English forms

Phoneme	Total	Unique Welsh graphemes	Unique English graphemes
j	130	4	5
ɪ	2557	14	7
ɛ	1470	14	8
æ	2000	16	5
ɔ	1361	16	9
ʌ	963	14	4
ʊ	96	11	6

Table D.2: Phoneme totals and numbers of unique grapheme pairs in adapted forms and Standard English forms

Phoneme	Total	Unique Welsh graphemes	Unique English graphemes
i	1567	22	17
e	841	18	12
a	752	16	9
o	98	8	3
au	170	12	4
əu	904	17	12
ɸu	267	14	12
w	427	6	5
jɸu	401	17	10
jʊə	34	7	3
ɪə	63	13	6
ɛə	81	14	13
aə	99	25	16
ɑə	17	10	5
ʊə	15	6	8
ɑː	528	16	11
oː	493	23	14
ɜː	303	18	10
ə	3407	18	21
ə	1073	22	18
p	2058	3	4
b	1247	5	5
m	2076	7	8
f	989	6	7
v	555	6	4
t	3701	6	13
d	1487	4	3
θ	193	3	1
ð	18	3	1
l	3005	5	9
n	3321	8	10

Table D.2: Phoneme totals and numbers of unique grapheme pairs in adapted forms and Standard English forms

Phoneme	Total	Unique Welsh graphemes	Unique English graphemes
s	3323	11	16
z	500	7	9
tʃ	316	12	7
dʒ	385	15	7
ʃ	487	10	14
ʒ	19	3	4
ɹ	2584	11	9
k	2905	9	15
g	608	4	7
ŋ	412	4	4
h	346	2	2

Appendix E

Survey of orthographic adaptation

E.i Survey: text



Ysgrifennu geiriau benthyg o'r Saesneg

Gwybodaeth am yr ymchwil a ffurflen gydsynio

Bethan Tovey-Walsh ydw i. Rwy'n fyfyrwraig PhD yng Ngholeg y Celfyddydau a'r Dyniaethau ym Mhrifysgol Abertawe. Mae'r holiadur hwn yn casglu data i gyfrannu at fy mhroject ymchwil PhD.

Os oes gen chi unrhyw gwestiynau, cysylltwch â fi drwy e-bost: 

Pwrpas yr ymchwil

Pwrpas fy ymchwil ydy deall sut mae siaradwyr Cymraeg yn defnyddio geiriau benthyg o'r Saesneg yn eu hiaith arferol. Rwy'n gobeithio cyfrannu at ein dealltwriaeth o ddatblygiad yr iaith fodern, a thaflu goleuni ar iaith anffurfiol yn ei holl amrywiaeth.

Eich cyfraniad chi

Os dewiswch chi gwblhau'r holiadur, fe fydd gofyn i chi ail-ysgrifennu geiriau Saesneg fel petaent yn eiriau Cymraeg (fe fydd mwy o gyfarwyddiadau yn yr holiadur ei hun). Dylai'r holiadur gymryd tua hanner awr i chi ei gwblhau. Gofynnaf hefyd am wybodaeth am eich cefndir fel siaradwr Cymraeg, er bod rho'r wybodaeth yma'n opsiynol.

Preifatrwydd

Byddaf yn defnyddio canlyniadau'r holiadur fel un ffynhonnell o wybodaeth ymysg eraill yn fy nhraethawd PhD. Mae'n bosib y byddaf hefyd yn trafod rhai atebion penodol yno, neu mewn cyhoeddiadau eraill. Serch hynny, ni fydd hi'n bosib i unrhyw un eich gwahaniaethu chi'n bersonol fel awdur eich atebion, gan nad yw'r holiadur yn gofyn am eich enw, eich cyfeiriad e-bost, neu unrhyw ddata eraill y gall eich adnabod chi fel unigolyn.

Goruchwyliaeth yr ymchwil

Fy ngoruchwylyfyr ydy Mr. Steven Morris ([REDACTED]) a'r Athro Tess Fitzpatrick ([REDACTED]). Croeso i chi gysylltu â nhw os oes gennych gwestiynau neu bryderon yn ymwneud â'ch cyfraniad i'r project ymchwil hwn.

Os ydych chi'n hapus i barhau i'r holiadur, darllenwch y datganiad isod, a chliciwch y blwch i ddangos eich bod yn cytuno.

- * 1. Rydw i'n 16 oed neu'n hŷn. Rydw i'n hapus i barhau gyda'r holiadur ymchwil hwn, ar sail y gwybodaeth uchod. Deallaf fod gennyf yr hawl i dynnu'n ôl o'r ymchwil ar unrhyw adeg cyn gwblhau'r holiadur. Deallaf hefyd y caiff fy hawl i fod yn ddienw a'm hawl i gyfrinachedd eu parchu.

- ☐ Cytuno
- ☐ Dim yn cytuno



Ysgrifennu geiriau benthyg o'r Saesneg

Diolch o galon i chi am ddewis cymryd yr amser i lenwi'r holiadur hwn. Fe fydd eich ymatebion yn cyfrannu'n sylweddol at fy ngwaith ymchwil, sy'n archwilio defnydd anffurfiol o'r iaith Gymraeg.

Gan mai iaith anffurfiol yw targed yr holiadur, peidiwch â phoeni pa mor "cywir" yw'ch ymatebion. **Eich iaith bob-dydd, naturiol sydd o diddordeb i fi.**

Cyn dechrau, gofynnir i chi ateb ychydig o gwestiynau yn ymwneud â chi'ch hunan. **Does dim rhaid i chi ateb unrhyw gwestiwn os nad ydych yn gyfforddus i wneud hynny.** Serch hynny, cofiwch fo'ch ymatebion yn gwbl gyfrinachol, ac yn ddefnyddiol iawn i fi fel ymchwilydd. Does dim modd i unrhyw un gysylltu gwybodaeth amdanoch chi (megis eich enw neu'ch lleoliad) gyda'ch ymatebion.

* 1. Beth yw'ch oedran?



* 3. Edrychwch ar y mapiau uchod. Mae'r map ar y chwith yn dangos enwi trefi er mwyn eich helpu i leoli'r ffiniau (y llinellau du). Mae'r map ar y dde yn dangos ardaloedd bras, gyda rhif yn gysylltiedig â phob un.

Gyda pha ardal fydddech chi'n cysylltu'ch Cymraeg llafar, ar y cyfan? (Ystyriwch eich acen, eich tafodiaith, a.y.y.b.)

☐ Wellgen i beidio ateb

☐ Ardal 1

☐ Ardal 2

☐ Ardal 3

☐ Ardal 4

☐ Ardal 5

☐ Ardal 6

☐ Ardal 7

☐ Ardal 8

☐ Ardal 9

* 4. Faint o'ch addysg cyn 18 oed ddaeth drwy gyfrwng y Gymraeg?

☐ Wellgen i beidio ateb

☐ Dim addysg ffurfiol drwy gyfrwng y Gymraeg

☐ Peth addysg gynradd yn unig drwy gyfrwng y Gymraeg

☐ Peth addysg uwchradd yn unig drwy gyfrwng y Gymraeg

☐ Peth addysg gynradd ac uwchradd drwy gyfrwng y Gymraeg

☐ Addysg gynradd i gyd drwy gyfrwng y Gymraeg

☐ Addysg uwchradd i gyd drwy gyfrwng y Gymraeg

☐ Addysg gynradd a pheth addysg uwchradd drwy gyfrwng y Gymraeg

☐ Addysg gynradd ac addysg uwchradd i gyd drwy gyfrwng y Gymraeg

* 5. Beth oedd brif iaith/prif ieithoedd eich cartref yn ystod eich plentyndod?

- ☐ Wellgen i beidio a le b
- ☐ Cymraeg
- ☐ Saesneg
- ☐ Cymraeg a Saesneg
- ☐ Cymraeg + arall/eraill
- ☐ Saesneg + arall/eraill
- ☐ Cymraeg, Saesneg + arall/eraill
- ☐ Iaith/ieithoedd heb gynnwys Cymraeg na Saesneg

* 6. Beth yw brif iaith/prif ieithoedd eich cartref nawr?

- ☐ Wellgen i beidio a le b
- ☐ Cymraeg
- ☐ Saesneg
- ☐ Cymraeg a Saesneg
- ☐ Cymraeg + arall/eraill
- ☐ Saesneg + arall/eraill
- ☐ Cymraeg, Saesneg + arall/eraill
- ☐ Iaith/ieithoedd heb gynnwys Cymraeg na Saesneg

* 7. Ydych chi'n defnyddio'r Gymraeg yn eich bywyd proffesiynol neu mewn addysg - e.e. yn y gwaith, yr ysgol, neu'r brifysgol?

- ☐ Wellgen i beidio a le b
- ☐ Ydw, yn gyson
- ☐ Ydw, ambeill waith
- ☐ Nac ydw

* 8. Ydych chi'n defnyddio'r Gymraeg yn eich bywyd cymdeithasol - e.e. gyda ffrindiau, yn y capel, neu o fewn grwpiau cymdeithasol?

- ☐ Well gen i beidio a le b
- ☐ Ydw, yn gyson
- ☐ Ydw, ambe ll waith
- ☐ Nac ydw

* 9. Ydych chi'n defnyddio'r Gymraeg i ysgrifennu dogfennau ffurfiol - e.e. traethodau, adroddiadau, papurau ymchwil, datganiadau i'r wasg, a.y.y.b.?

- ☐ Well gen i beidio a le b
- ☐ Ydw, yn gyson
- ☐ Ydw, ambe ll waith
- ☐ Nac ydw

* 10. Ydych chi'n defnyddio'r Gymraeg i ysgrifennu'n anffurfiol - e.e. e-byst, negeseuon testun, lythyrau at ffrindiau, a.y.y.b.?

- ☐ Well gen i beidio a le b
- ☐ Ydw, yn gyson
- ☐ Ydw, ambe ll waith
- ☐ Nac ydw

* 11. Ydych chi'n defnyddio'r Gymraeg ar y cyfryngau cymdeithasol, e.e. Twitter, Facebook, Reddit?

- ☐ Well gen i beidio a le b
- ☐ Ydw, yn gyson
- ☐ Ydw, ambe ll waith
- ☐ Nac ydw, ond mi ydw i'n defnyddio Saesneg a/neu iaith arall
- ☐ Nac ydw; does gen i ddim cyfrifon ar y cyfryngau cymdeithasol



Ysgrifennu geiriau benthyg o'r Saesneg

Cyflwyniad

Ar adegau, fe welwn air Saesneg o fewn brawddeg Cymraeg, a'r gair wedi ei addasu o ran sillafu. Wrth gwrs, mae yna nifer o eiriau benthyg sefydledig sydd wedi'u addasu yn y fath modd, e.e. "cwestiwn", "siwgr", a "gêm". Mae newid sillafu geiriau Saesneg ar hap yn arfer gyffredin iawn mewn iaith anffurfiol, a fe'i gwelir mewn cyd-destunau ffurfiol, hefyd - er enghraifft, pan ddefnyddir term technegol sydd heb gyfieithiad Cymraeg. Dyma enghreifftiau o'r hyn a olygir:

"Eniwê, fydda' i ddim yn y parti ar y wfcend."
(Anyway, fydda' i ddim yn y parti ar y weekend.)

"Nifer bach o bobl sy'n dioddef o affantasia, yn ôl ymchwilyr."
(...dioddef o a phantasia...)

Cyfarwyddiadau

Ar y tudalen nesaf, fe welwch gyfres o eiriau Saesneg gyda brawddegau Cymraeg yn eu dilyn. Mae yna fwch wag ymhob brawddeg. Gofynnir i chi ail-ysgrifennu'r gair Saesneg fel petai'n air Cymraeg, gan ddefnyddio rheolau sillafu Cymraeg, er mwyn ei ffitio yn y bwch yn y frawddeg. Ystyriwch pa mor ddealladwy fydd y gair i siaradwyr Cymraeg eraill fyddai'n ei darllen.

Dyma enghraifft o'r hyn a olygir:

peacock
Mae e'n gwisgo fel ____

Ymatebion posib: "pîcoc", "picoc", a.y.y.b.

Os ydych yn teimlo fod angen treigladau ar y gair yn y cyd-destun, croeso i chi dreiglo'ch ymateb. Gofynnir i chi **beidio** â chywiro'ch treigladau drwy gyfeirio at unrhyw lyfr/wefan, fodd hynny. Eich teimlad naturiol chi ynglŷn â chywirdeb treiglo'r gair sydd o ddiddordeb. Cofiwch fo'ch ymatebion yn hollol ddienw; fydd neb yn barnu cywirdeb eich iaith!

Dyma esiampl arall, mewn cyd-destun lle mae treiglo'n bosib:

television

Ydy dy ____ di'n gweithio?

Ymatebion posib: "delifisiyn"; "television", "delevisiyn" a.y.y.b.

Mae'n bosib na fyddwch yn teimlo fod angen newidiadau ar y gair er mwyn iddo gydymffurfio â rheolau sillafu Cymraeg. Ar y llaw arall, efallai na fyddwch yn teimlo fod modd i newid y sillafu heb wneud y gair yn annealladwy i'ch darllenwyr. Os felly, ymatebwch drwy deipio'r gair yn union fel y mae hi yn ymddangos yn y Saesneg. Dyma enghraifft:

dogs

Pwy biau'r ____ 'na?

Ymateb posib: "dogs"

Cofiwch, nid oes un ateb "cywir" i unrhyw gwestiwn. Eich dewis chi ynglŷn â sut i newid y geiriau sydd yn bwysig.



Ysgrifennu geiriau benthyg o'r Saesneg

Isod, fe welwch 75 gair Saesneg, a brawddeg Cymraeg yn dilyn pob un.

Ysgrifennwch eich fersiwn chi o'r gair Saesneg yn y blwch sy'n dilyn y frawddeg gyda'r cyd-destun Cymraeg. Dylai'ch ymateb cynnwys eich fersiwn chi o'r gair yn unig - **does dim eisiau ail-ysgrifennu'r frawddeg iawn.**

Cofiwch, nid cyfieithiad i air Cymraeg sydd angen, ond addasiad o'r gair i system sillafu'r Gymraeg. Dwi'n deall fod yna eiriau na fydddech chi'n dewis eu defnyddio yn eich Cymraeg, ond ceisiwch eich gorau i'w haddasu nhw fodd bynnag. Cofiwch fod hi'n iawn i beidio addasu gair os fyddai'r canlyniad, yn eich barn chi, yn annealladwy. Os felly, ymatebwch gyda'r gair Saesneg gwreiddiol.

Os ydych chi eisiau defnyddio llythyren gydag acen grom (â, ê, î, ô, ŷ, w), ac yn cael trafferth ei teipio ar eich cyfrifiadur, teipiwch ^ ar ôl y llythyren (e.e. 'yngly^n a^' yn lle 'ynglŷn â').

1. painting

Wyt ti wedi gweld ei ____ newydd hi?

2. boiler

Ydy dy ____ di wedi torri?

3. breadsticks

Ti'n hoffi'r ____ 'na ti'n cael yn y bwyty Eidaleg?

4. technology

Mae ____ yn hynod bwysig i'r byd heddi.

5. tractor

Mae ei ____ newydd hi'n cyrraedd fory.

6. table

Mae ei ____ e'n llawn papurau.

7. dozens

Ces i ____ ohonyn nhw.

8. doubt

Ei ____ mwyaf e yw gallu pasio'r arholiad.

9. cable

Dwi 'di colli fy ____ i chargo'r iPhone.

10. cardigan

Mae'r ____ ma'n gynnes.

11. goodness

Wel, ____ - dyna oedd sioc!

12. goal

Ei ____ e enillodd y gêm.

13. cheer

Rhowch ____ mawr i'r enillydd!

14. journey

Siwd oedd dy ____ di i'r Alban?

15. themes

Wy isie llyfrau sy'n cynnwys ____ addas at blant.

16. themselves

'Mond ____ sydd ar fai.

17. sawdust

Ma' ngheg i'n teimlo fel ____!

18. zero

Ma' bateri'r ffôn 'ma bron lawr at ____.

19. shoot

Ma'r pobl teledu'n defnyddio'r tŷ mawr am ____.

20. hatching

Defnyddiwch techneg ____ yn eich darlun.

21. major

Y broblem mwya' ____ heddiw yw newid hinsawdd.

22. nasty

Ma' fe mor ____ at bobl eraill.

23. yesterday

Peidiwch poeni am ____!

24. forward

Ma' fe'n blentyn ____ iawn.

25. violent

Troseddau ____ sy'n ofni pobl.

26. warfare

Ma' 'na ryw ____ yn mynd ymlaen yn y cyfarfod!

27. rough

'Sdim eisie chwarae'n ____.

28. lounge

Eistedd yn y ____ ma' nhw.

29. angler

Ar ôl ymddeol, ____ fydda' i!

30. ah-ha

____ -wela' i'r broblem nawr!

31. opera

Noson yn yr ____ oedd fy anrheg pen-blwydd.

32. ought to

"Ydy e'n gwneud ymarfer corff?"

"____, ond pwy a wyr?"

33. issue

Cartrefi fforddadwy yw'r ____ mawr o flaen y Cyngor.

34. envelope

Oes gen ti ____ mawr?

35. onion

Mae angen ____ arna' i er mwyn 'neud pasta heno.

36. untied

Ma' dy lasys di'n ____ i gyd!

37. oops

____ - ddaru fi anghofio f'allweddii!

38. Asia

I ____ 'dwi eisie teithio nesa'.

39. overtaking

Ma' ____ fan 'na'n beryglus.

40. equal

Does dim yn ____ i gwrw o Wlad Belg.

41. ooze

Ych! Ma' 'na ryw ____ yn dod o waelod y sinc!

42. Ivy League

Ai prifysgolion yr ____ ____ yw'r orau yn yr UDA?

43. oyster

Bwytes i un ____ gwael, a fues i'n sâl am wythnos!

44. outfit

Hoffi dy ____ dil

45. earth

Lliwiau ____ sy' hoff gen i.

46. armour

Roedd y marchogion yn gwisgo ____ trwm.

47. orange

Blas ____ sydd ar y siocled.

48. drink

Ble dodes i fy ____ i, te?

49. orphan

Ma' hi'n ____ nawr, druan.

50. earwig

Ych! Ma' 'na ____ yn y sinc!

51. aircraft

Faint o ____ fydd yn hedfan yn y sioe?

52. tourist

Dwi ddim yn hoffi bod yn ____.

53. avoidant

Ma' ymddygiad y ci yn ____ iawn.

54. original

Ai'r ____ sy' gyda ti?

55. otherwise

Paid! ____ gei di stŵr gan mam!

56. breathe

Os ti'n cael pwl o banic, jyst ____!

57. survive

____ tan fory - dyna i gyd sydd angen!

58. wardrobe

Ydy fy nghrys i yn y ____?

59. fetching

Dyna sgert ____ ti'n gwisgo!

60. vicious

Mae'r ci 'na'n un ____.

61. goldfish

Mae fy ____ wedi marw.

62. camouflage

Mae angen ____ ar filwyr.

63. vacuum

Dydy swm ddim yn teithio mewn ____.

64. English

Dim ond ____ mae nhw'n siarad yn y siop.

65. bizarre

Wel, am ddydd ____!

66. enjoy

Dyma dy fwyd - ____!

67. anyhow

____, mae'r amser wedi pasio nawr.

68. connoisseur

Ma' fe'n ____ o winoedd Ffrainc.

69. reservoir

Dwi'n byw'n agos i'r ____.

70. karma

Fydd ____ yn dal lan gyda hi rhyw ddydd!

71. hemisphere

Dau ____ sydd gan ein Daear.

72. semaphore

Defnyddion nhw ____ i gyfathrebu.

73. variation

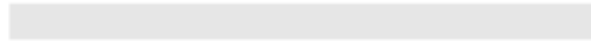
'Sdim ____ mawr ymysg yr aelodau.

74. materials

Pa ____ defnyddiwyd i greu'r car?

75. carnivore

_____ oedd y T-Rex.



Pan gliciwch y botwm "Cyflwyno fy atebion" isod, caiff eich atebion eu cadw a'u hanfon i'r ymchwilydd, p'un ai a fyddwch chi wedi cwblhau pob cwestiwn ai beidio.

Ni fydd hi'n bosib i chi newid eich atebion na'u tynnu 'nôl ar ôl clicio'r botwm.

E.ii Survey: text (English translation)

Writing English-origin loanwords

I'm Bethan Tovey-Walsh. I'm a PhD student in the College of Arts and Humanities in Swansea University. This questionnaire collects data to contribute to my PhD research project.

If you have any questions, contact me through e-mail: [REDACTED].

The purpose of the research:

The purpose of my research is to understand how Welsh speakers use loanwords from English in their everyday language. I hope to contribute to our understanding of the development of the contemporary [Welsh] language, and to illuminate colloquial language in all its variety.

Your contribution

If you decided to complete the questionnaire, you'll be asked to re-write English words as though they were Welsh words (there will be further directions within the questionnaire itself). The questionnaire should take around half an hour to complete. I will also ask for information about your background as a Welsh speaker, although giving this information is optional.

Privacy

I will use the results of this questionnaire as one amongst many sources of information in my PhD thesis. I may also discuss some specific responses there, or in other publications. Nonetheless, it will not be possible for anyone to identify you personally as the author of your responses, since the questionnaire does not ask for your name, your email address, or any other data that could identify you as an individual.

Research supervision

My supervisors are Mr. Steven Morris ([REDACTED]) and Professor Tess Fitzpatrick ([REDACTED]). You are welcome to contact them if you have any questions or concerns relating to your contribution to this research project.

If you are happy to continue to the questionnaire, read the statement below, and click the box to show that you agree.

1. I am age 16 or older. I am happy to continue with this research questionnaire, on the basis of the information above. I understand that I have the right to withdraw from participation at any time before completing the questionnaire. I also understand that my right to be anonymous and my right to privacy will be respected.

Agree

Do not agree

Writing English-origin loanwords

Heartfelt thanks to you for deciding to take the time to fill out this questionnaire. Your responses will make a significant contribution to my research, which investigates informal use of the Welsh language.

Since the target of the questionnaire is informal language, don't worry about the "correctness" of your responses. It's your everyday, natural language that interests me.

Before starting, you'll be asked to answer a few questions about yourself. You are not obliged to answer any question if you do not feel comfortable doing so. But do remember that your answers are entirely anonymous, and will be very useful to me as a researcher. Nobody can connect information about you, such as your name or location, with your answers.

1. What is your age?

2. Since when have you spoken Welsh?

Prefer not to answer

Since I learned to speak

Since young childhood (8 years old or younger)

Since older childhood (between 9 and 12 years old)

Since my teens (between 12 and 19 years old)

Since 20 years old or above

3. Look at the maps above. The left-hand map shows the names of towns in order to help you locate the borders (the black lines). The right-hand map shows general areas, with a number assigned to each one.

With which area do you associate your spoken Welsh, on the whole? (Consider your accent, dialect, and do on).

Prefer not to answer

ETC

4. How much of your education before 18 years of age was Welsh-medium?

Prefer not to answer

No formal Welsh-medium education

Only some primary education through the medium of Welsh

Only some secondary education through the medium of Welsh

Some primary and secondary education through the medium of Welsh

All primary education through the medium of Welsh

All secondary education through the medium of Welsh

All primary education and some secondary education through the medium of Welsh

All primary and secondary education through the medium of Welsh

5. What was/were the main language(s) of your home during childhood?

Prefer not to answer

Welsh

English

Welsh and English

Welsh and other(s)

English and other(s)

Welsh, English, and other(s)

Language(s) not including Welsh or English

6. What is/are the main language(s) of your home now?

Prefer not to answer

Welsh

English

Welsh and English

Welsh and other(s)

English and other(s)

Welsh, English, and other(s)

Language(s) not including Welsh or English

7. Do you use Welsh in your professional life or in education – e.g. at work, school, or university?

Prefer not to answer

Yes, regularly

Yes, sometimes

No

8. Do you use Welsh socially – e.g. with friends, at church, or in social groups?

Prefer not to answer

Yes, regularly

Yes, sometimes

No

9. Do you use Welsh to write formal documents – e.g. essays, reports, research papers, press releases etc.?

Prefer not to answer

Yes, regularly

Yes, sometimes

No

10. Do you use Welsh to write informally – e.g. e-mails, text messages, letters to friends, etc.?

Prefer not to answer

Yes, regularly

Yes, sometimes

No

11. Do you use Welsh on social media – e.g. Twitter, Facebook, Reddit?

Yes, regularly

Yes, sometimes

No, but I do use English and/or (an) other language(s)

No; I have no social media accounts

Introduction

Sometimes, we'll see an English word in a Welsh sentence, and the word's spelling

has been adapted. Of course, there are many settled borrowings which have been adapted in this way, e.g. “cwestiwn” [*question*], “siwgr” [*sugar*], and “gêm” [*game*]. Changing the English spelling as a one-off is a very common practice in informal language, and we also see it in formal contexts – for example when a technical term is used which has no Welsh translation. Here are examples of the kind of thing I mean:

”Eniwê, fydda’ i ddim yn y parti ar y wîcend.”

Anyway, I won’t be at the party on the weekend.

(Anyway, fydda’ i ddim yn y parti ar y weekend.)

”Nifer bach o bobl sy’n dioddef o affantasia, yn ôl ymchwilyr.”

Only a small number of people suffer from aphantasia, according to researchers.

(...dioddef o aphantasia...)

Instructions

On the next page, you’ll see a series of English words, with Welsh sentences following them. There is a blank space in every sentence. Please re-write the English word as though it were a Welsh word, using Welsh spelling rules, in order to fit it into the gap in the sentence. Consider how comprehensible the word will be to other Welsh speakers who might read it.

Here is an example of what I mean:

peacock

”Mae e’n gwisgo fel _____”

He dresses like a _____

Possible responses: “pîcoc”; “picoc”, etc.

If you feel that the word in context requires treiglâd, you’re welcome to apply one to your response. You are asked not to correct your treigladau with reference to any book or website, however. My interest is in your natural sense about the correctness of applying treiglâd to the word. Remember that your response is totally anonymous nobody will be judging the correctness of your language!

Here’s another example, in a context where treiglâd is possible:

television

Ydy dy ____ di’n gweithio?

Is your _____ working? – N.B. the context triggers soft treigladd, making the first grapheme <d> instead of <t>.

Possible responses: “delifisiyn”; “telefision”, “delefisiyn”, etc.

It’s possible that you will not feel that the word needs any changes in order to conform to Welsh spelling rules. On the other hand, perhaps you will not feel there is any way to change the spelling without making the word incomprehensible to your readers. If so, respond by typing the word exactly as it appears in English. Here’s an example:

dogs

Pwy biau’r _____ ’na?

Who owns those _____?

Possible response: “dogs”.

Remember there is no one “correct” answer to any question. Your own decision about how to adapt the words is what’s important.

Below, you will see 75 English words, with a Welsh sentence following each one.

Write your version of the English word in the box that follows the sentence giving the Welsh context. Your response should include only your version of the word – you needn’t rewrite the entire sentence.

Remember, what’s needed is not a Welsh translation, but an adaptation to the Welsh spelling system. I understand that there are words you would not choose to use in your Welsh, but try your best to adapt them anyway. Remember that it’s fine not to adapt a word if the result, in your estimation, would be incomprehensible. If so, respond with the original English word.

If you wish to use a letter with a circumflex (â, ê, î, ô, û, ð), and have difficulty typing it on your computer, type â after the letter (e.g. “ynglyñ â” instead of “yngllyn â”).

75 English words are listed, with a Welsh sentence containing a blank to provide context for each word.

When you click the “Upload my answers” button below, your answers will be saved and sent to the researcher, whether or not you have completed every question.

It will not be possible for you to change or withdraw your answers after clicking the button.

E.iii Survey: prompt list

- aircraft
- angler
- anyhow
- armour
- asia
- avoidant
- bizarre
- boiler
- breadsticks
- breathe
- cable
- camouflage
- cardigan
- carnivore
- cheer
- connoisseur
- doubt
- dozens
- drink
- earth
- earwig
- english
- enjoy
- envelope
- equal
- fetching
- forward
- goal
- goldfish
- goodness
- hatching
- hemisphere
- issue
- ivy league
- journey
- karma
- lounge
- major
- materials
- nasty
- onion
- oops
- ooze
- opera
- orange
- original
- orphan
- otherwise
- ought to
- outfit
- overtaking
- oyster
- painting
- reservoir
- rough
- sawdust
- semaphore
- shoot
- survive
- table
- technology
- themes
- themselves
- tourist
- tractor
- untied

• vacuum	• violent	• yesterday
• variation	• wardrobe	
• vicious	• warfare	• zero

E.iv Survey: variant forms

• aircraft	• enjoy	• orange
• angler	• envelope	• original
• anyhow	• equal	• orphan
• armour	• fetching	• otherwise
• asia	• forward	• ought to
• avoidant	• goal	• outfit
• bizarre	• goldfish	• overtaking
• boiler	• goodness	• oyster
• breadsticks	• hatching	• painting
• breathe	• hemisphere	• reservoir
• cable	• issue	• rough
• camouflage	• ivy league	• sawdust
• cardigan	• journey	• semaphore
• carnivore	• karma	• shoot
• cheer	• lounge	• survive
• connoisseur	• major	• table
• doubt	• materials	• technology
• dozens	• nasty	• themes
• drink	• onion	• themselves
• earth	• oops	• tourist
• earwig	• ooze	• tractor
• english	• opera	• untied

• vacuum	• violent	• yesterday
• variation	• wardrobe	
• vicious	• warfare	• zero

E.v Survey: demographic and Welsh usage graphs

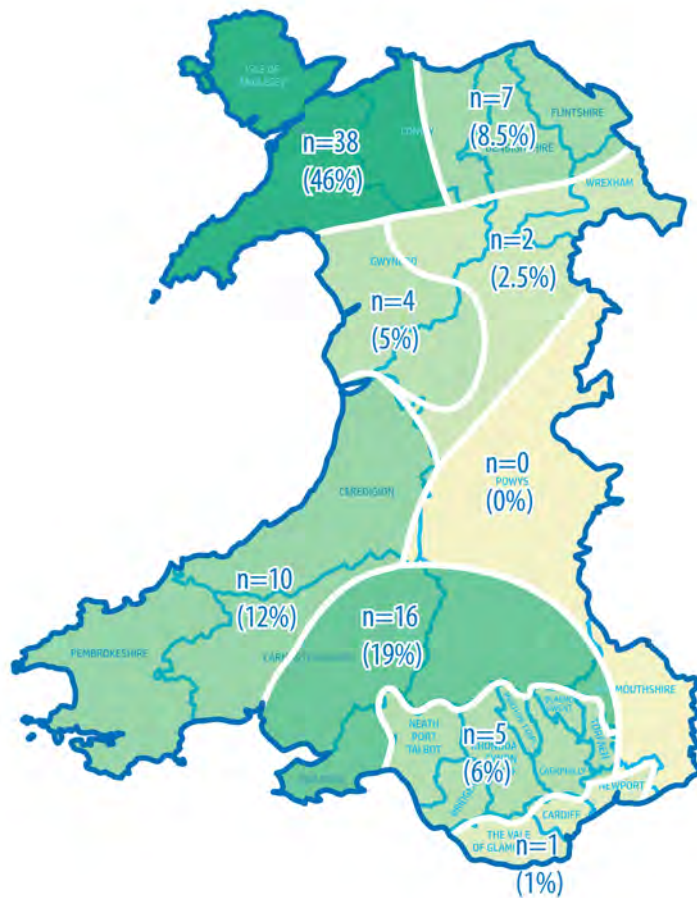


Figure E.1: Geographic distribution of survey respondents.

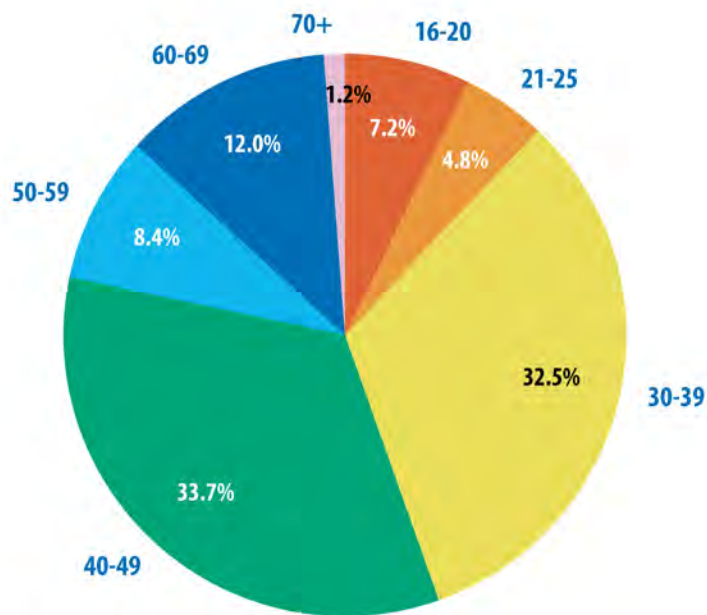


Figure E.2: Survey respondents' age at time of survey.

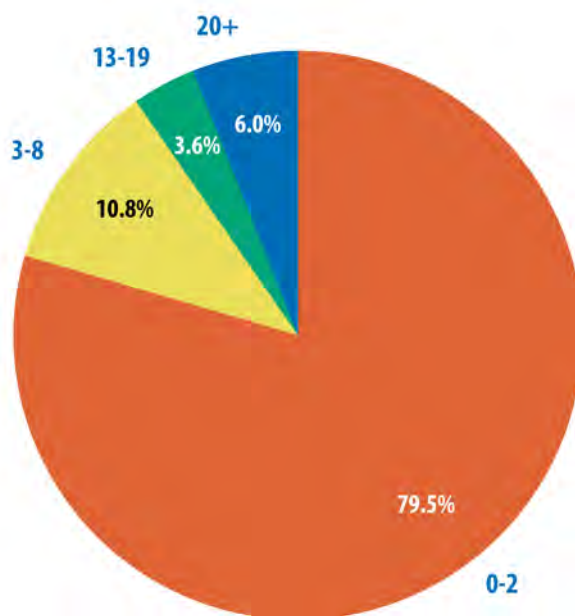


Figure E.3: Survey respondents' age of acquisition of Welsh.

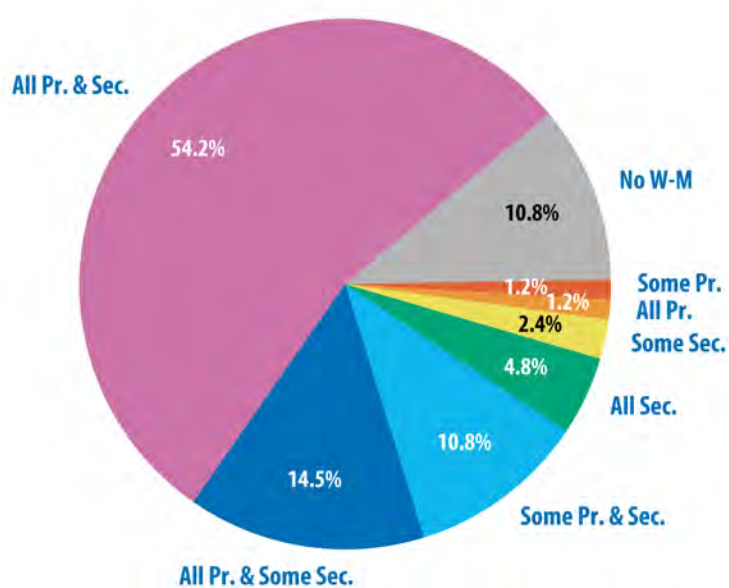


Figure E.4: Survey respondents' Welsh-medium primary and secondary education.

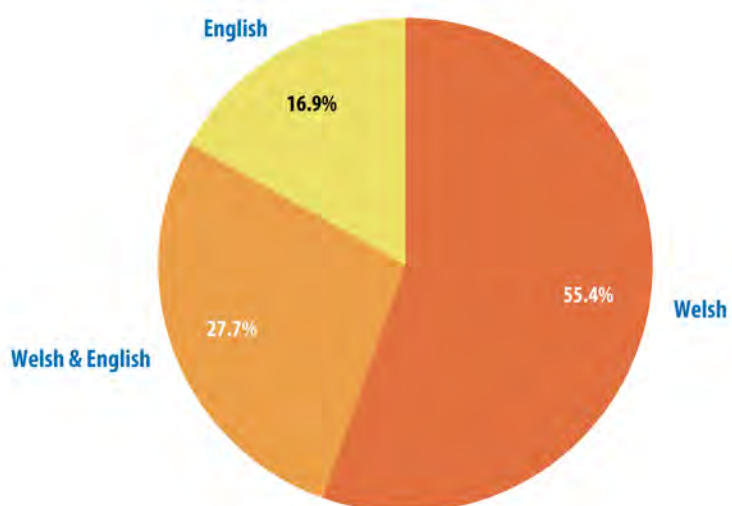


Figure E.5: Survey respondents' reported home language(s) in childhood.

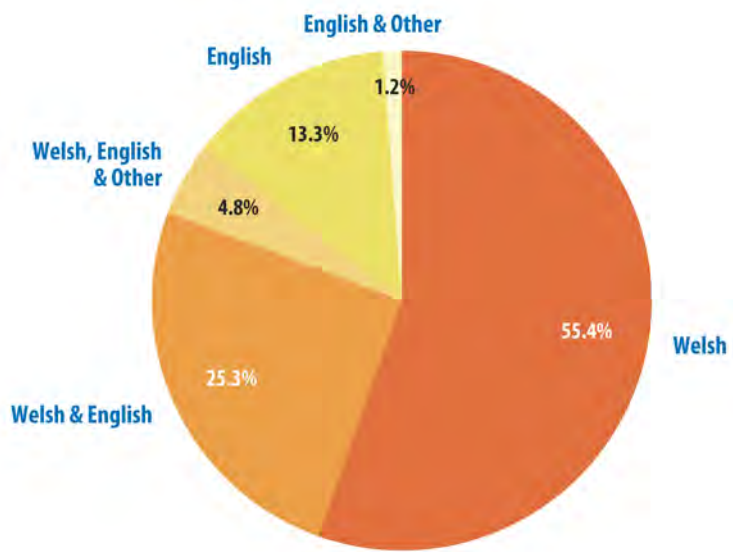


Figure E.6: Survey respondents' reported current home language(s).

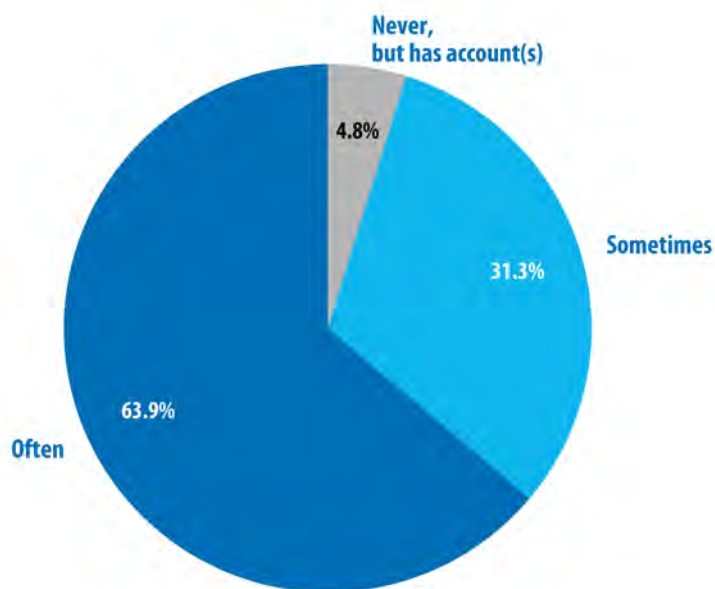


Figure E.7: Survey respondents' reported use of Welsh on social media.

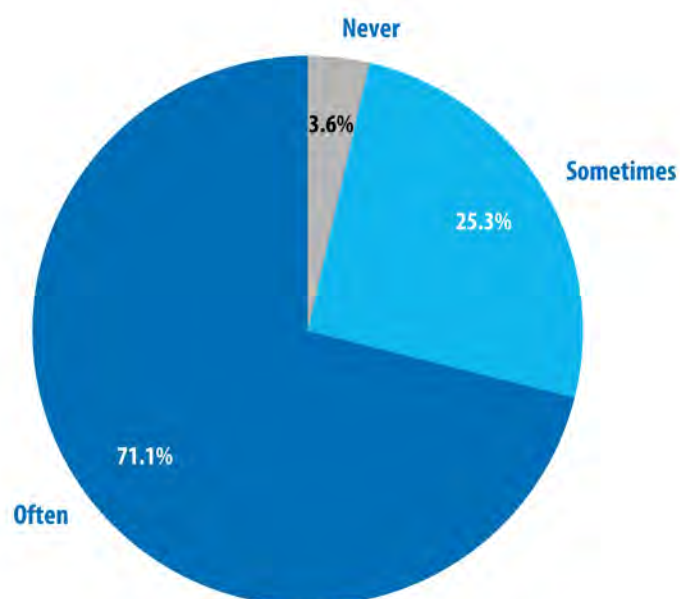


Figure E.8: Survey respondents' reported social use of spoken Welsh.

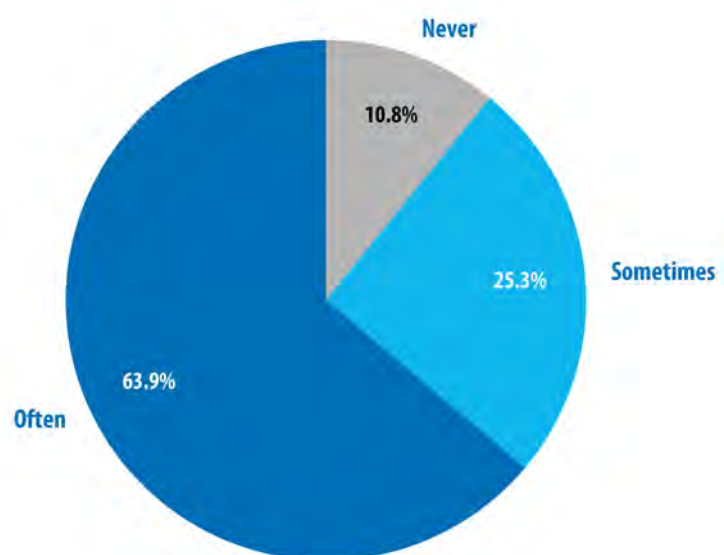


Figure E.9: Survey respondents' reported professional use of spoken Welsh.

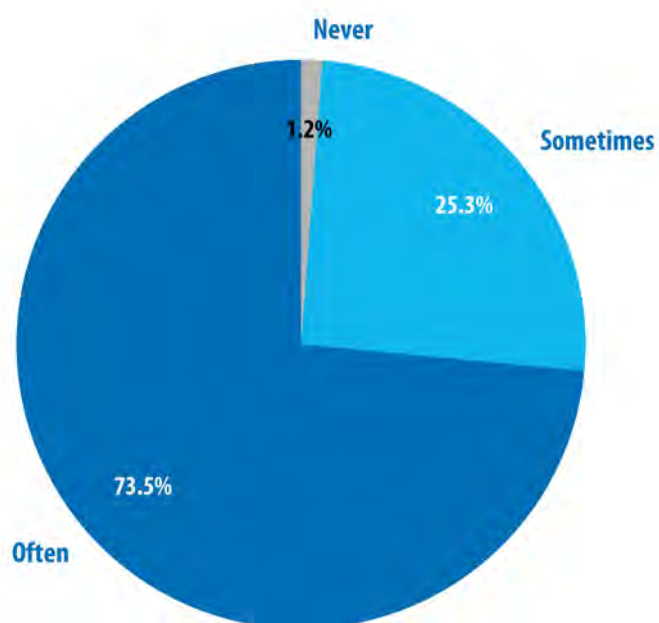


Figure E.10: Survey respondents' reported informal use of written Welsh.

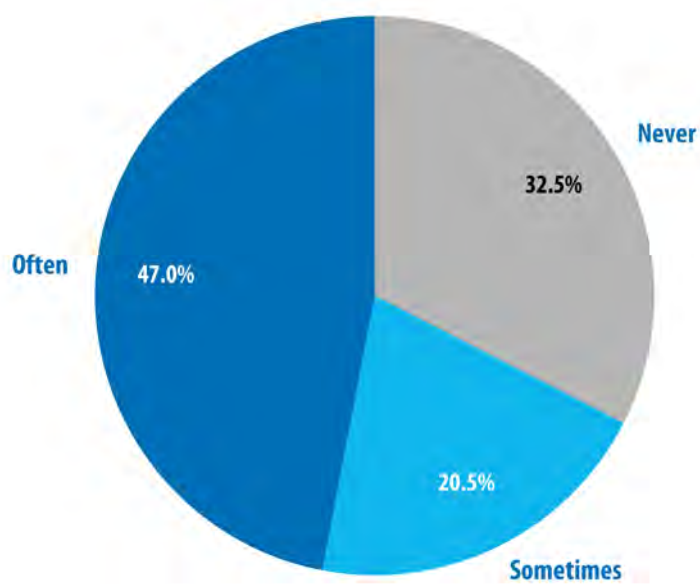


Figure E.11: Survey respondents' reported formal use of written Welsh.

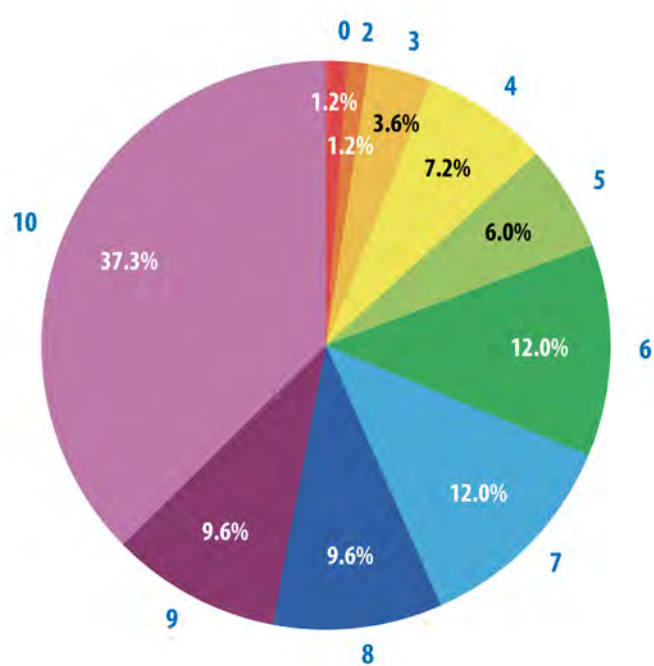


Figure E.12: Survey respondents' reported total use of Welsh.

E.vi Survey: summary statistics

Table E.1: Numbers of variant forms and total number of lexemes with that number of variants.

Number of variants	Lexemes count
4	1
5	2
7	3
8	6
9	5
10	2
11	7
12	7
13	5
14	5
15	4
16	2
17	7
18	3
20	4
21	1
22	1
25	2
26	1
28	1
32	1
33	1
34	1
35	1
38	1

Table E.2: Phoneme totals and numbers of unique grapheme pairs in adapted forms and Standard English forms

Phoneme	Total	Unique Welsh graphemes	Unique English graphemes
j	161	4	2
ɪ	1444	7	1
ɛ	968	10	3
æ	565	6	1
ɔ	480	6	1
ʌ	551	11	3
ʊ	88	5	3
ɪi	1141	15	5
ei	634	12	3
ai	301	12	2
oi	320	8	2
au	378	12	2
əu	532	7	2
ɸu	407	11	2
w	564	3	2
jɸu	165	12	2
jʊə	0	0	0
ɪə	246	26	3
ɛə	247	12	3
aə	83	10	1
ɑə	0	0	0
ʊə	118	5	1
ɑː	659	9	3
oː	723	14	5
ɜː	240	20	3
ə	1955	23	8
ə	956	23	6
p	408	2	2
b	640	4	2
m	814	3	2
f	948	4	3

Table E.2: Phoneme totals and numbers of unique grapheme pairs in adapted forms and Standard English forms

Phoneme	Total	Unique Welsh graphemes	Unique English graphemes
v	1021	5	2
t	2023	4	3
d	1113	3	1
θ	159	1	1
ð	246	4	2
l	1384	3	2
n	1761	2	2
s	1494	3	2
z	809	8	5
tʃ	233	11	2
dʒ	557	8	3
ʃ	408	9	3
ʒ	164	11	2
ɹ	1335	2	1
k	1121	6	5
g	654	4	2
ŋ	542	4	2
h	237	2	1

E.vii Survey: variant forms

{aircraft}

<êrcrafft>	33
<ercrafft>	12
<aircrafft>	11
<aircraft>	10
<êcrafft>	3
<eircrafft>	2
<êcraffd>	2
<êrcraffd>	2
<aercrafft>	1
<aercraft>	1
<eircrefft>	1
<ercraft>	1
<êcraft>	1
<êcrâfft>	1

{angler}

<anglyr>	35
<angler>	27
<anglar>	7
<angly>	3
<angeler>	2
<anglêr>	2
<anggler>	1
<anggly>	1
<anglr>	1
<anglur>	1
<anglwr>	1
<anglâr>	1
<annglyr>	1

{anyhow}

<enihow>	35
<enihaw>	25
<anyhow>	9
<enihow>	3
<enîhow>	2
<eniaw>	1
<enihoe>	1
<eniháw>	1
<enihôw>	1
<ennihow>	1

	<enîhaw>	1
	<énihaw>	1
	<ênihow>	1
	<ênïow>	1
{ armour }		
	<armyr>	34
	<armor>	15
	<armwr>	12
	<armour>	8
	<armôr>	3
	<âmy>	3
	<ârmyr>	2
	<amyr>	1
	<armer>	1
	<armw̃r>	1
	<armÿr>	1
	<armýr>	1
	<âmyr>	1
{ asia }		
	<asia>	40
	<asha>	11
	<eisha>	8
	<esia>	5
	<ashia>	4
	<esia>	3
	<êisa>	2
	<êsha>	2
	<êsia>	2
	<eisa>	1
	<eishza>	1
	<eisiâ>	1
	<esha>	1
	<eshia>	1
{ avoidant }		
	<afoidant>	42
	<afoidynt>	14
	<avoidant>	13
	<yfoidynt>	4
	<afoident>	3
	<yfoidant>	3
	<afoudant>	1
	<afwydant>	1
	<avoident>	1

{bizarre}

<bisâr>	44
<bisar>	9
<bizâr>	6
<bizarre>	5
<bisâr>	3
<bizar>	3
<bissâr>	2
<bisaar>	1
<bisarê>	1
<busâr>	1
<buzâ>	1
<bysâr>	1
<byzar>	1
<byzzar>	1
<bîsar>	1
<bîsâr>	1
<bîzar>	1

{boiler}

<boilyr>	26
<boiler>	25
<boilar>	20
<boily>	3
<bolier>	2
<boaelr>	1
<boilwr>	1
<boliyr>	1
<boylar>	1
<boîlêr>	1
<bwyler>	1

{breadsticks}

<bredstics>	56
<bredsics>	7
<breadstics>	6
<breadsticks>	3
<brêdstics>	3
<bredsticks>	2
<bredstycs>	2
<bredstucs>	1
<brêdsticks>	1

{breathe}

<brîdd>	56
<breathe>	9
<bridd>	7
<brîth>	3
<bredd>	1
<breedd>	1
<briidd>	1
<brudd>	1
<brêdd>	1
<brîv>	1
<brÿdd>	1

{ cable }

<cêbl>	30
<cebl>	16
<cable>	7
<ceibl>	7
<ceibyl>	7
<cabl>	3
<cebyl>	3
<cêbyl>	3
<cabyll>	1
<caibl>	1
<ceibwl>	1
<ceybl>	1
<ceybyl>	1
<câbl>	1
<câbêl>	1

{ camouflage }

<camofflaj>	27
<camofflâj>	12
<camoflaj>	5
<camouflage>	5
<camyfflâj>	5
<camyfflaj>	4
<camoflage>	2
<camoflâj>	2
<camwfflaj>	2
<camwfflâj>	2
<camafflaj>	1
<camaflage>	1
<camiffflaj>	1
<camoffladj>	1
<camofflag>	1
<camofflaje>	1

<camofflâsh>	1
<camowfflaj>	1
<camowflaj>	1
<camwflaj>	1
<camwflâsh>	1
<camyflaj>	1
<camôflaj>	1
<camwfflaj>	1
<camwflage>	1
<cámoofflâzh>	1

{cardigan}

<cardigan>	70
<cardigyn>	6
<cardigen>	2
<cadigyn>	1
<cardigân>	1
<cardîgan>	1
<cardîgen>	1
<câdigyn>	1

{carnivore}

<carnifôr>	39
<carnifor>	29
<carnivore>	4
<carnîfôr>	2
<carnifore>	1
<carnifwr>	1
<carnifõr>	1
<carnivor>	1
<càrnifor>	1
<cânifo>	1
<cânifor>	1
<cânifô>	1

{cheer}

<cheer>	27
<tshîyr>	6
<chîr>	4
<tsîr>	4
<tshîr>	3
<chiyr>	2
<chîyr>	2
<tshier>	2
<tshir>	2

<tshiy>	2
<tsier>	2
<tsiyr>	2
<chir>	1
<chêr>	1
<chîr>	1
<dsir>	1
<jîr>	1
<jîyr>	1
<shiy>	1
<tcheâr>	1
<tchir>	1
<tchîr>	1
<tjîy>	1
<tschîr>	1
<tshiar>	1
<tshiâr>	1
<tshêr>	1
<tshîar>	1
<tshîy>	1
<tshîar>	1
<tsîiyr>	1
<tsîÿr>	1
<tsîÿr>	1
<tsîÿr>	1
<tsêr>	1
<tsîÿr>	1

{connoisseur}

<conosÿr>	12
<conisyr>	8
<conisÿr>	8
<connoisseur>	7
<conosyr>	7
<conysÿr>	5
<conosiÿr>	2
<conosÿr>	2
<conosÿr>	2
<conosÿr>	2
<conessour>	1
<coniser>	1
<conisiwr>	1
<conisor>	1
<conisÿr>	1
<conniosseur>	1
<conniswr>	1
<connisyr>	1
<connosyr>	1
<conosiwr>	1
<conosiyr>	1
<conosy>	1
<conosÿ>	1
<conosÿr>	1

<conwasŷr>	1
<conwswr>	1
<conyser>	1
<conysiyŷr>	1
<conyswr>	1
<conysyr>	1
<conysyyŷr>	1
<conysŷ>	1
<conísŷr>	1

{doubt}

<dowt>	50
<dawt>	21
<doubt>	7
<dywt>	2
<dout>	1
<doŵt>	1
<dôut>	1

{dozens}

<dysyns>	34
<dozens>	9
<dysens>	9
<dosens>	4
<dysins>	4
<dyzyns>	4
<dosyns>	2
<dwsins>	2
<dossens>	1
<dosêns>	1
<dwsens>	1
<dysns>	1
<dysons>	1
<dysêns>	1
<dyzins>	1
<dyznz>	1
<dyzynz>	1
<dôsens>	1

{drink}

<drinc>	76
<drink>	4
<drunc>	2
<drync>	1

{earth}

<yrrth>	38
<earth>	23
<ÿrrth>	6
<ÿth>	3
<erth>	2
<urth>	2
<yth>	2
<êrth>	2
<yrrth>	1
<yyth>	1
<ÿrrth>	1

{earwig}

<iyrrwig>	18
<earwig>	15
<îrrwig>	10
<ÿyrrwig>	7
<iarwig>	6
<irwig>	5
<iywig>	4
<ierwig>	2
<êrrwig>	2
<îrrwyg>	2
<earwig>	1
<eiyrwig>	1
<erwig>	1
<iawwig>	1
<irwyg>	1
<irwig>	1
<iyrrwig>	1
<iêrrwig>	1
<iÿrrwig>	1
<yerrwig>	1
<îarwig>	1
<iywwug>	1

{english}

<inglish>	63
<english>	10
<inglysh>	2
<inglish>	2
<ingglish>	1
<inglisi>	1
<ingylish>	1
<ungglush>	1
<înglish>	1

{enjoy}

<enjoi>	49
<injoy>	15
<enjoy>	3
<enjou>	2
<enjói>	1
<ensioe>	1
<injoy>	1
<ynjoi>	1

{envelope}

<enfelôp>	22
<enfilop>	14
<enfelop>	11
<enfilôp>	11
<enfylop>	5
<envelope>	4
<enfelowp>	2
<envelop>	2
<enfelope>	1
<enfilowp>	1
<enfylywp>	1
<enfylôp>	1
<enfilôp>	1
<envelôp>	1
<onfelop>	1
<onfelôp>	1
<wnfolop>	1
<énfylowp>	1

{equal}

<icwal>	25
<îcwal>	18
<equal>	12
<ecwal>	8
<icwyl>	8
<îcwyl>	5
<icwâl>	2
<eecwal>	1
<eicwyl>	1
<eicwâl>	1
<iqwal>	1

{fetching}

<ffetshing>	25
<ffetching>	19
<fetching>	8
<ffetsing>	7
<ffetsin>	6
<fetshing>	3
<ffetshin>	3
<feching>	1
<fejing>	1
<fetchin>	1
<fetsing>	1
<ffeching>	1
<ffejing>	1
<ffetjung>	1
<ffetsiyn>	1
<ffettshing>	1

{forward}

<fforward>	31
<fforwyrđ>	15
<fforwyd>	8
<forward>	6
<fforwad>	4
<ffôwyd>	3
<ffoward>	2
<ffowyrđ>	2
<fforwod>	1
<fforwâd>	1
<fforwÿrd>	1
<ffowyd>	1
<ffwrward>	1
<ffôrwad>	1
<ffôward>	1
<forwerđ>	1
<forword>	1
<forwyd>	1
<forwyrđ>	1
<fôward>	1

{goal}

<gôl>	64
<gowl>	8
<gol>	3
<goal>	1
<gowal>	1
<gowyl>	1

	<goyl>	1
	<gowl>	1
{goldfish}		
	<goldffish>	31
	<goldfish>	19
	<gowldffish>	13
	<gôldffish>	5
	<gowldfish>	3
	<gôldfish>	2
	<goldffisi>	1
	<goldffush>	1
	<goldfish>	1
	<golffish>	1
	<gowldfsh>	1
	<gowldfush>	1
	<gôlfish>	1
{goodness}		
	<gwdnes>	46
	<gwdnys>	14
	<gwdness>	6
	<gôwdnes>	6
	<goodness>	5
	<gwdnas>	1
	<gwdnus>	1
	<gwness>	1
	<gôwdnis>	1
	<gôwdnys>	1
{hatching}		
	<hatching>	24
	<hatshing>	21
	<hatshin>	6
	<hatsing>	6
	<hatsin>	4
	<haching>	3
	<hachin>	2
	<hajing>	1
	<hatchinh>	1
	<hatjing>	1
	<hatjung>	1
	<hattshing>	1
	<âtshen>	1

{hemisphere}

<hemisffer>	17
<hemisffir>	14
<hemisffêr>	13
<hemisffiy>	10
<hemisffiy>	6
<hemisffir>	3
<hemisphere>	3
<hemisffier>	2
<hemisffiy>	2
<hemesffer>	1
<hemesffir>	1
<hemesfiyr>	1
<hemesphêr>	1
<hemisffiar>	1
<hemisffiy>	1
<hemisfir>	1
<hemissfir>	1
<hemysffiy>	1
<hemysffir>	1
<hémusffiy>	1
<hêmisphere>	1

{issue}

<ishw>	19
<ishiw>	18
<isiw>	18
<issue>	11
<ishw>	4
<isiw>	2
<ushw>	2
<esiw>	1
<ishiew>	1
<ishue>	1
<issw>	1
<issîw>	1
<isyw>	1
<isw>	1
<usw>	1

{ivy league}

<aifi lig>	29
<eifi lig>	16
<ivy league>	12
<ivy lig>	9
<ifi lig>	3

<aifi lig>	2
<aifi lig>	1
<aifi lêg>	1
<eivî lig>	1
<eyfi lig>	1
<eyfi lÿg>	1
<ify league>	1
<ify lig>	1
<ivi lige>	1
<ivy lig>	1
<yifi lig>	1

{journey}

<jyrni>	46
<journey>	7
<siwrne>	7
<jyrnî>	3
<jwrne>	2
<jwrni>	2
<siwrna>	2
<siwrni>	2
<gournê>	1
<jernu>	1
<jurney>	1
<jyrnei>	1
<jyrney>	1
<jyyni>	1
<jw̄rnei>	1
<jÿni>	1
<jÿrni>	1
<jÿrnî>	1
<shwrna>	1
<siournei>	1

{karma}

<carma>	57
<karma>	17
<carmā>	2
<câma>	2
<kârma>	2
<camy>	1
<carrmar>	1
<câmyh>	1

{lounge}

<lownj>	45

<lawnj>	24
<lounge>	7
<lywnj>	2
<laung>	1
<liwnj>	1
<lounj>	1
<lôwnj>	1

{major}

<major>	16
<mejyr>	14
<meijyr>	12
<mêjyr>	8
<mejor>	6
<meijor>	4
<majyr>	3
<meijy>	3
<maijor>	2
<mêjor>	2
<majer>	1
<mayjyr>	1
<meijwr>	1
<meijÿr>	1
<meja>	1
<mejo>	1
<mêjar>	1
<mêjwr>	1
<mêjy>	1
<mêsior>	1

{materials}

<materials>	29
<matirials>	16
<matiriyls>	8
<matîrials>	6
<matîriyls>	5
<materiyls>	4
<matêrials>	2
<materiels>	1
<materiwls>	1
<matirîwls>	1
<maturials>	1
<matîriylz>	1
<matîriÿls>	1
<meterials>	1
<myterials>	1
<mytiriyls>	1
<mytîriyls>	1

{nasty}

<nasti>	61
<nasdi>	8
<nasty>	6
<nastí>	4
<nâsti>	2
<nastie>	1
<nastÿ>	1

{onion}

<yniyn>	38
<onion>	17
<ynion>	9
<wnion>	2
<wniwn>	2
<yniwn>	2
<yniÿn>	2
<oniwn>	1
<oniyn>	1
<onîon>	1
<unyîun>	1
<yniÿn>	1
<ynnion>	1
<ynîon>	1

{oops}

<wps>	54
<ŵps>	22
<wwps>	4
<oops>	2
<woops>	1

{ooze}

<ŵs>	38
<ŵz>	10
<ooze>	8
<wws>	6
<ŵws>	4
<ŵŵs>	3
<wwz>	2
<wz>	2
<ooz>	1
<wwss>	1
<wwws>	1
<wwze>	1
<wze>	1
<wzz>	1
<ôoze>	1

	<ŵse>	1
	<ŵss>	1
	<ŵwz>	1
{opera}		
	<opera>	39
	<opra>	35
	<opyra>	5
	<oprâ>	3
	<opryh>	1
{orange}		
	<orenj>	35
	<orinj>	27
	<orynj>	4
	<orange>	3
	<oranj>	3
	<orainj>	1
	<orenge>	1
	<oring>	1
	<oringe>	1
	<orins>	1
	<orunj>	1
	<orênj>	1
{original}		
	<orijinal>	37
	<original>	19
	<orijinył>	15
	<orijinâl>	2
	<originel>	1
	<orijinỳł>	1
	<orijional>	1
	<orijynal>	1
	<orutsionyl>	1
	<orújunyl>	1
	<wriginal>	1
	<ôoriginal>	1
{orphan}		
	<orffan>	55
	<orffyn>	8
	<orphan>	8
	<offan>	2
	<orffon>	2

<ôffyn>	2
<offyn>	1
<orffyn>	1
<ôffyn>	1
<ôrffyn>	1
<ôrphan>	1

{otherwise}

<yddyrwais>	16
<otherwise>	15
<ythyrwais>	5
<ydderwais>	4
<yddyrweis>	4
<yddywais>	4
<oddyrwais>	3
<odderweis>	2
<odderwis>	2
<otherweis>	2
<yddyrwaiz>	2
<ytherwise>	2
<ythyrweis>	2
<odderwais>	1
<odderwise>	1
<oddyrweis>	1
<oddyrwise>	1
<othyrweis>	1
<overwais>	1
<yddawais>	1
<ydderweis>	1
<ydderwise>	1
<ydderwys>	1
<ydderwys>	1
<yddyweis>	1
<yddyweiz>	1
<ytherwais>	1
<ytherwaisz>	1
<ytherwis>	1
<ytherwâis>	1
<ythrweis>	1
<ýddywaiz>	1
<ýddyrwais>	1

{ought to}

<ôt tw>	20
<ôt tŵ>	19
<ought to>	17

<ort t ^h >	4
<owt to>	3
<ot tw>	2
<ot t ^h >	2
<otw>	2
<owt tw>	2
<ôrt tw>	2
<ôtw>	2
<aut to>	1
<ort tw>	1
<ought tw>	1
<oyt t ^h >	1
<ôt to>	1
<ôt two>	1

{outfit}

<owtffit>	42
<awtffit>	16
<outfit>	6
<owtfit>	5
<outffit>	3
<awtffut>	2
<ywtffit>	2
<autfit>	1
<awtffyt>	1
<awtfit>	1
<iwtffit>	1
<ôwtfit>	1
<wtffit>	1

{overtaking}

<ofyrtêcing>	7
<ofertêcing>	6
<ofertêcin>	4
<ofertaking>	3
<ofyrtecing>	3
<ofyrteicing>	3
<overtaking>	3
<ofertacing>	2
<ofyrtêcin>	2
<owferteicing>	2
<owfyртеicing>	2
<ofartecing>	1
<ofartêcen>	1
<ofertaicing>	1
<ofertakin>	1
<ofertecin>	1

<ofertecing>	1
<oferteicing>	1
<ofetêcing>	1
<ofyrtecin>	1
<ofyrteicin>	1
<ofyrtêcyng>	1
<ofyrtêking>	1
<ofyteicing>	1
<overtêcing>	1
<owferteicin>	1
<owferteicyn>	1
<owfyrteiking>	1
<owfytacing>	1
<owfyteicing>	1
<owfyteicung>	1
<ywfyteicing>	1
<ôfyrteicin>	1
<ôfytêcing>	1

{oyster}

<oistyr>	27
<oister>	23
<oyster>	12
<oisdyr>	6
<oistar>	3
<oistr>	3
<oisdy>	2
<oestr>	1
<oestyr>	1
<oisdyh>	1
<oisty>	1
<oystyr>	1
<oïstyr>	1
<wyster>	1

{painting}

<peinting>	48
<painting>	11
<peintin>	11
<pênting>	4
<paenting>	2
<penting>	2
<paunting>	1
<peintyng>	1
<peynting>	1
<pêntîn>	1

{reservoir}

<reservoir>	13
<resyfwar>	9
<reserfwar>	8
<resefwar>	7
<resefwâr>	6
<resyfwâr>	6
<resyrfwar>	6
<resyrfwâr>	4
<reserfoir>	3
<resyfwa>	3
<resefoir>	2
<reserfwâr>	2
<resyfoir>	2
<raesyfwar>	1
<resefwa>	1
<resefwâr>	1
<resevoir>	1
<resofwar>	1
<resrfwa>	1
<resyrfoir>	1
<resyrfwa>	1
<rezefoir>	1
<résyfwar>	1
<rézyfwa>	1
<rêserfwâr>	1

{rough}

<ryff>	68
<rwff>	6
<ruff>	5
<rough>	1
<rûff>	1
<rÿff>	1
<rÿff>	1

{sawdust}

<sôdyst>	31
<sawdust>	13
<sodyst>	11
<sawdyst>	8
<sodust>	3
<sôdust>	3
<sawdwst>	2
<sordyst>	2

	<sowdyst>	2
	<sôdysd>	2
	<sodysd>	1
	<sodÿst>	1
	<sordust>	1
	<sôdÿst>	1
	<sôrdyst>	1
{semaphore}		
	<semaffôr>	38
	<semaffor>	28
	<semaphore>	5
	<semaphôr>	2
	<semoſfor>	2
	<semyffôr>	2
	<samaffor>	1
	<semaffwr>	1
	<semaffô>	1
	<semafor>	1
	<sémyffo>	1
{shoot}		
	<shwt>	55
	<shwt>	9
	<siwt>	8
	<shoot>	6
	<shiwt>	1
	<shwwt>	1
	<shwwwt>	1
	<shôt>	1
	<siwt>	1
{survive}		
	<syrfaif>	25
	<syrfeif>	21
	<survive>	7
	<syfaif>	3
	<syrfif>	2
	<syrfif>	2
	<serfif>	1
	<serfife>	1
	<serveive>	1
	<surfeif>	1
	<surfife>	1
	<syfeif>	1

<syfif>	1
<syfáif>	1
<syrfife>	1
<syrfive>	1
<syrfâif>	1

{table}

<têbl>	24
<teibl>	12
<tabl>	10
<tebl>	10
<teibyl>	9
<table>	8
<tebyl>	3
<têbyl>	2
<tabêl>	1
<taibyl>	1
<teble>	1
<têibl>	1

{technology}

<tecnoloji>	31
<tecnologi>	14
<tecnolyji>	10
<technology>	6
<tecnologi>	3
<tecnoloji>	3
<tecnology>	2
<tecnoloji>	1
<tecnolyji>	1
<tecnolagu>	1
<tecnoleji>	1
<tecnologî>	1
<tecnolojy>	1
<tecnolygi>	1
<tecnôloji>	1
<tegnolyji>	1
<toecnoloji>	1

{themes}

<thîms>	59
<themes>	10
<thems>	2
<thims>	2

	<thêms>	2
	<thimes>	1
	<thêmes>	1
	<thîmz>	1
{themselves}		
	<ddemselfs>	45
	<themselves>	12
	<themselfs>	10
	<ddemselfes>	4
	<themselfes>	3
	<ddymselfs>	2
	<ddemsels>	1
	<ddemselvz>	1
	<ddemsélfz>	1
	<themselvs>	1
	<thymselfs>	1
{tourist}		
	<twrist>	40
	<tŵrist>	20
	<torist>	6
	<torrist>	2
	<tourist>	2
	<twryst>	2
	<tôrist>	2
	<tŵrisd>	2
	<twrisd>	1
	<twrust>	1
	<tôrisd>	1
	<tôrust>	1
{tractor}		
	<tractor>	61
	<tracdor>	9
	<tractyr>	5
	<tractwr>	2
	<tractôr>	2
	<tracdy>	1
	<tracdyh>	1
	<tracter>	1
	<tracty>	1
{untied}		
	<yntaid>	30

<untied>	11
<ynteid>	11
<untaid>	3
<untide>	2
<yntied>	2
<yntîd>	2
<unteid>	1
<untiêd>	1
<yntid>	1
<yntide>	1
<yntáid>	1

{vacuum}

<faciwm>	56
<faciŵm>	4
<vacuum>	4
<faciwym>	3
<facwm>	3
<facîwm>	3
<vaciwym>	2
<faciwem>	1
<faciwn>	1
<facum>	1
<facŵm>	1
<fáciwym>	1
<vaciwym>	1
<vacum>	1
<vacîwm>	1

{variation}

<variation>	10
<ferieishyn>	9
<ferieshyn>	6
<fariation>	4
<feriation>	4
<fêriêshyn>	4
<feriashon>	3
<feriesiyn>	3
<fêrieishyn>	3
<fariasion>	2
<fariatiwn>	2
<feriasiyn>	2
<ferieishiyn>	2
<ferieshiyn>	2
<ferieshon>	2
<feriesion>	2
<fêriêshon>	2

<fariashon>	1
<fariatshon>	1
<farieshon>	1
<feriashion>	1
<feriashyn>	1
<feriashyn>	1
<feriasion>	1
<ferieishon>	1
<ferieisiwn>	1
<ferieisiyn>	1
<ferieshiwn>	1
<ferietion>	1
<feriêtion>	1
<feriêtsion>	1
<fêrieisiyn>	1
<fêrieitsion>	1
<fêrieshiyn>	1
<fêriêsiwn>	1
<fêriêsiyn>	1
<veriâshyn>	1
<vêriêshyn>	1

{vicious}

<fishys>	31
<fisiys>	9
<fishiys>	8
<vicious>	6
<fisiws>	5
<ficious>	3
<fishws>	3
<fishios>	2
<vishys>	2
<fichiys>	1
<fishas>	1
<fishous>	1
<fishÿs>	1
<fishÿs>	1
<fisios>	1
<fisious>	1
<fushys>	1
<vishous>	1
<visiys>	1
<visuss>	1

{violent}

<fiolent>	20

<feiolent>	19
<failynt>	7
<faiolynt>	7
<violent>	7
<failent>	4
<faiolent>	4
<filent>	3
<faiylynt>	2
<fiolant>	2
<vilynt>	2
<faiylent>	1
<feilynt>	1
<fiolynt>	1
<faiolent>	1
<filynt>	1
<voilent>	1

{ wardrobe }

<wardrôb>	20
<wardrob>	15
<wordrob>	14
<wordrôb>	14
<wôdrôb>	4
<wardrobe>	3
<wodrôb>	2
<wadrob>	1
<wadrop>	1
<wardrowb>	1
<wodrob>	1
<wordrôp>	1
<wâdrob>	1
<wâdrôbe>	1
<wôdrowb>	1
<wôdrôb>	1
<wôdrob>	1

{ warfare }

<warffêr>	14
<wôrffêr>	13
<warfare>	10
<worfffer>	8
<warfffer>	6
<worffêr>	6
<warfêr>	3
<worfêr>	3
<warfer>	2

<wôffê>	2
<wôrffer>	2
<waffer>	1
<warfair>	1
<warfere>	1
<warffér>	1
<worfer>	1
<worffair>	1
<wârfer>	1
<wârffêr>	1
<wôffe>	1
<wôfê>	1
<wôrffare>	1
<wôrffor>	1
<wôrffêr>	1
<worfêr>	1

{yesterday}

<iestyrdei>	19
<iesterdei>	12
<yesterday>	9
<iestyrdê>	5
<iesdyrdê>	3
<iesterday>	3
<iesterdê>	3
<iestydei>	3
<iestyrde>	3
<iesdydei>	2
<iestyrdai>	2
<yesterdei>	2
<iesdyrdei>	1
<iestaday>	1
<iestardei>	1
<iesterde>	1
<iesterdêi>	1
<iestrdei>	1
<iestydê>	1
<iésdydei>	1
<yestaday>	1
<yesterdai>	1
<yesterde>	1
<yesterdê>	1
<yesturday>	1
<yestwrdeu>	1
<yestyrdêi>	1
<yestyrdê>	1

{zero}

<sero>	40
<zero>	12
<siro>	10
<ŝiro>	8
<zïro>	4
<zîro>	3
<serô>	1
<siryw>	1
<ŝirow>	1
<ŝïro>	1
<tsero>	1
<zirow>	1

Appendix F

Corpus of Welsh-language tweets

F.i Twitter: summary statistics

Table F.1: Twitter corpus and subcorpora: basic counts.

Corpus	Words	Users	Tweets	Words per tweet (avg.)
official	7,817,912	380	368,880	21.2
personal	7,913,712	380	515,786	15.3
ALL DATA	15,731,624	760	884,666	18.3

Table F.2: Twitter corpus and subcorpora: Number of original tweets (OT), quote tweets (QT), and replies, and average length in words of each type of tweet.

Corpus	OT	QT	Reply	OT length (avg. wds.)	QT length (avg. wds.)	Reply length (avg. wds.)
original	327,794	30,077	46,556	20.6	13.0	14.2
personal	262,162	36,597	311,494	13.0	11.5	13.1
ALL DATA	589,956	66,674	358,050	17.3	12.3	13.7

Table F.3

Number of variants	Headword count (official)	Headword count (personal)	Headword count (all)
1	69	21	13
2	89	60	61
3	43	41	46
4	20	36	31
5	11	29	28
6	6	27	30
7	4	13	14
8	0	7	10
9	0	2	2
10	0	3	6
11	0	4	3
12	0	2	2
13	0	3	4

Table F.4

Phoneme	Total	Total (official)	Total (personal)	Unique Welsh graphemes	Unq. Welsh graphemes (official)	Unq. Welsh graphemes (personal)	Unique English graphemes	Unq. English graphemes (official)	Unq. English graphemes (personal)
j	1247	432	815	2	2	2	2	2	2
i	43,041	23,075	19,966	8	2	8	2	2	2
ε	114,283	79,137	35,146	9	8	9	4	4	4
æ	24,126	10,616	13,510	4	4	3	2	2	2
ɔ	76,417	36,969	39,448	12	7	12	4	4	4
ʌ	16,483	2,941	13,542	9	6	9	4	4	4
ʊ	2,091	490	1,601	5	4	5	3	3	3
ii	57,654	20,767	36,887	13	7	13	7	7	7
ei	3,736	563	3,173	11	9	11	4	4	4
ai	22,684	5,833	16,851	9	6	9	4	4	3
oi	1,859	763	1,096	2	2	2	1	1	1
au	11,149	5,256	5,893	5	3	5	1	1	1
əu	26,273	16,471	9,802	5	4	5	3	3	3
əu	5,534	3,443	2,091	11	9	10	5	5	5
w	11,827	2,602	9,225	6	6	6	3	3	3
jɥu	51,665	2,333	3,332	17	8	16	5	4	5
jɔə	115	44	71	9	4	9	3	3	3
ɪə	38	16	22	5	5	3	2	2	2
εə	1,331	709	622	6	4	6	3	2	3

Phoneme	Total	(off.)	(per.)	W. graph.	(off.)	(per.)	E. graph.	(off.)	(per.)
aə	82	60	22	3	3	3	2	2	2
ɑə	1,090	686	404	7	4	7	1	1	1
ʊə	366	106	260	8	5	8	3	3	3
ɑ:	11,094	7,138	3,956	4	3	4	2	2	2
o:	5,118	1,638	3,480	13	12	12	8	8	8
ɜ:	5,789	2,071	3,718	6	4	6	3	3	3
ə	116,473	66,000	50,473	16	11	16	8	7	8
ə	8,590	4,031	4,559	15	10	15	6	6	6
p	121,403	71,747	49,656	4	4	4	3	3	3
b	24,131	7,006	17,125	6	6	6	4	4	4
m	42,248	19,535	22,713	5	4	5	4	4	4
f	41,535	18,425	23,110	3	3	3	3	3	3
v	39,362	24,447	14,915	5	4	5	2	2	2
t	67,458	28,782	38,676	9	6	8	6	6	6
d	55,695	28,866	26,829	6	4	6	2	2	2
θ	1,342	688	654	2	2	2	1	1	1
ð	288	89	199	2	2	2	1	1	1
l	124,311	67,141	57,170	6	4	6	4	4	4
n	108,773	56,466	52,307	5	4	5	3	3	3
s	83,300	38,383	44,917	9	7	8	7	7	7
z	17,927	4,834	13,093	4	4	4	3	3	3
ʧ	7,841	2,619	5,222	10	6	10	3	3	3
ʤ	17,698	12,477	5,221	10	10	9	5	5	5

Phoneme	Total	(off.)	(per.)	W. graph.	(off.)	(per.)	E. graph.	(off.)	(per.)
f	47,113	32,800	14,313	8	6	8	7	6	7
ʒ	3,884	1,419	2,465	6	2	5	2	1	2
ɹ	96,105	47,596	48,509	6	4	6	4	4	4
k	56,991	31,927	25,064	9	8	9	7	7	7
g	2,773	1,190	1,583	3	3	3	3	3	3
ŋ	1,548	329	1,219	2	2	2	2	2	2
h	39,808	29,733	10,075	1	1	1	1	1	1

Table F.5: Twitter lexemes - number of lexemes (Welsh-origin ILUs and English-origin ELUs) selected for study in each semantic category.

Category	Origin	
colloquial	Welsh	40
	English	45
domestic	Welsh	32
	English	43
media	Welsh	37
	English	41
sociopolitical	Welsh	38
	English	41
technical	Welsh	33
	English	40
uncategorized	Welsh	39
	English	42

Table F.6: Twitter lexemes (English-origin ELUs only) - number of lexemes in each semantic category with a strong, weak, or absent Welsh-origin synonym.

Synonym status	Category	
no synonym	colloquial	4
	domestic	20
	media	24
	sociopolitical	13
	technical	18
	uncategorized	9
strong synonym	colloquial	11
	domestic	11
	media	9
	sociopolitical	15
	technical	10
	uncategorized	12
weak synonym	colloquial	30
	domestic	11
	media	8
	sociopolitical	13
	technical	12
	uncategorized	21

Top 20 emoji which are most characteristic of the personal subcorpus

Faces:



Hands:



Gestures:



Food and drink:



Top 20 emoji which are most characteristic of the official subcorpus

Hands:



Flags and signs:



Sport



Media



Celebration and evaluation



Top 20 'neutral' emoji, which are used equally frequently in both subcorpora

Travel and geography:



Animals:



Faces:



People:



Food and Drink:



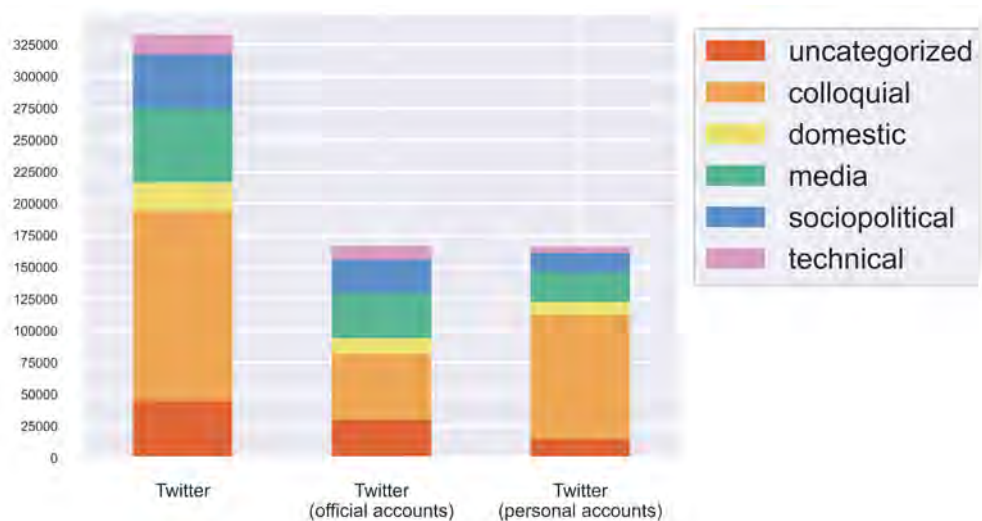
Flags and signs:



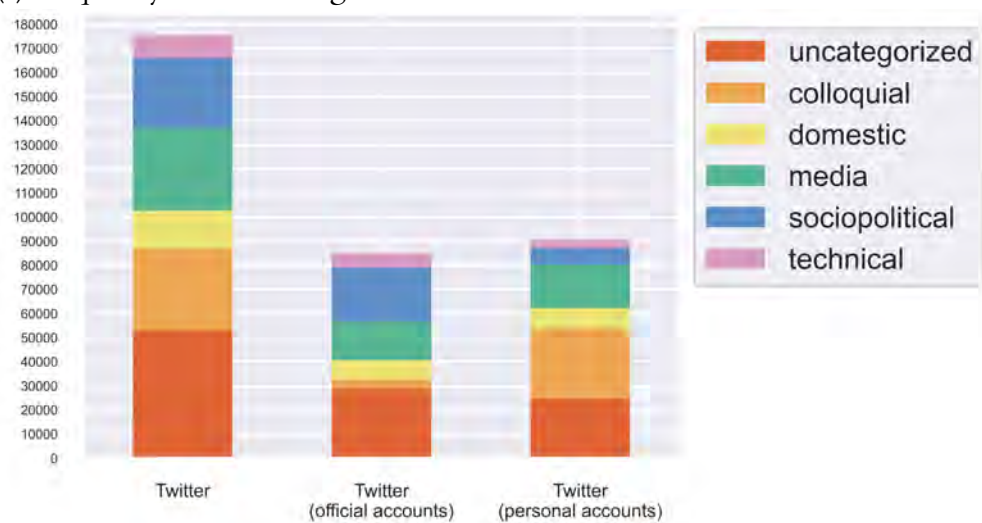
F.ii Twitter: lexeme frequency graphs

Notes on data representation and pre-processing

Proper nouns have been excluded from all graphs in this section.

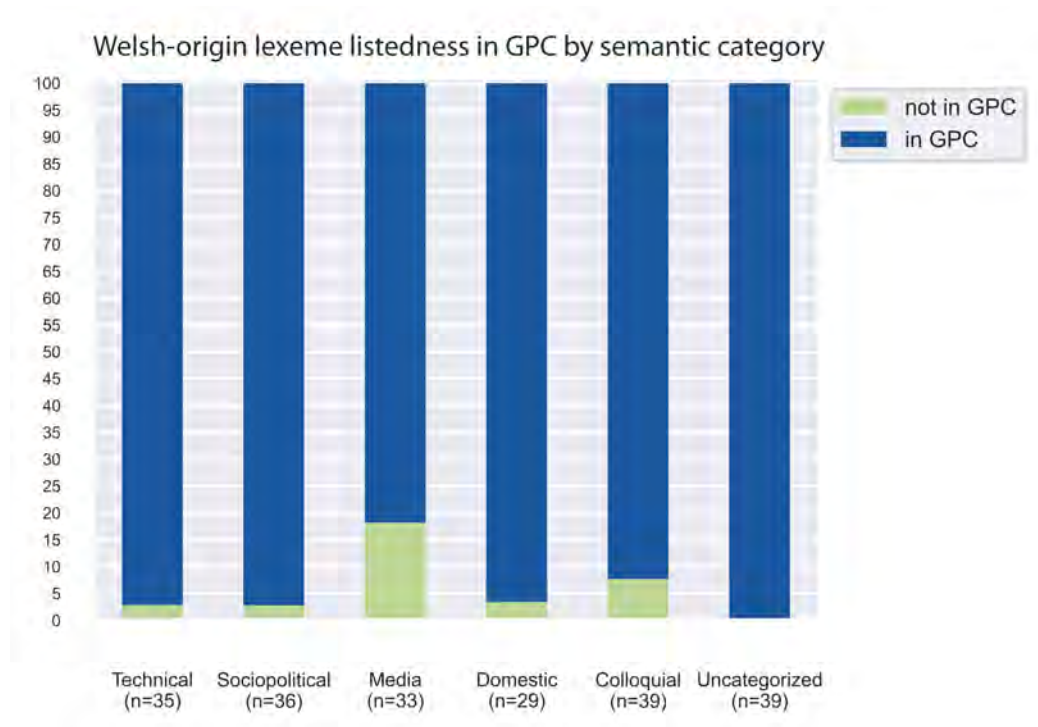


(a) Frequency of Welsh-origin ILUs

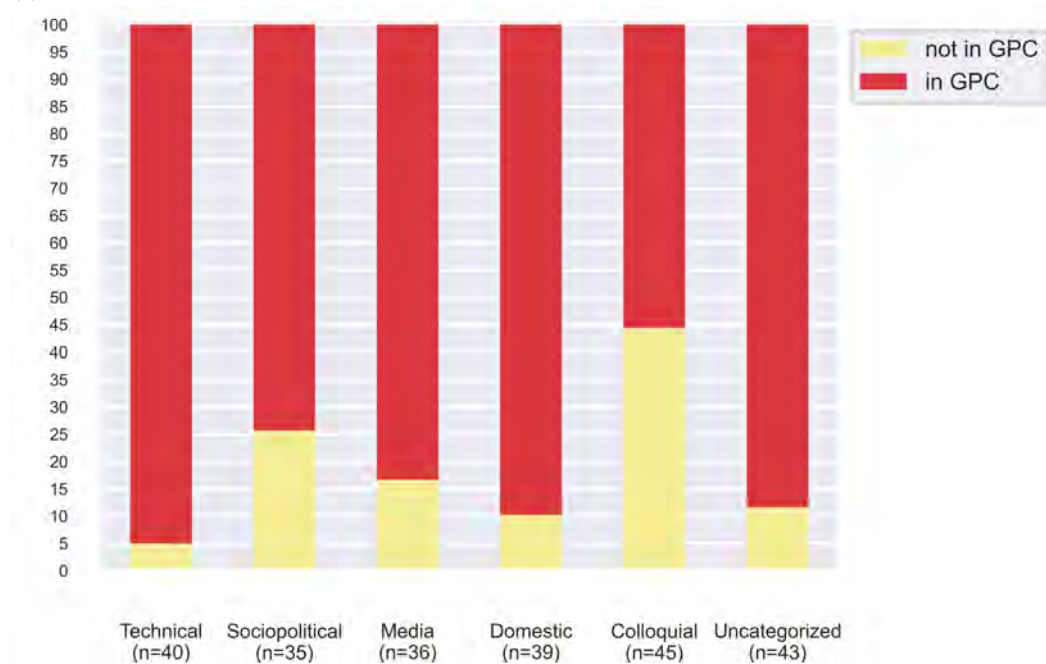


(b) Frequency of English-origin ELUs

Figure F.1: Lexeme frequency by semantic category in the Twitter corpus and subcorpora

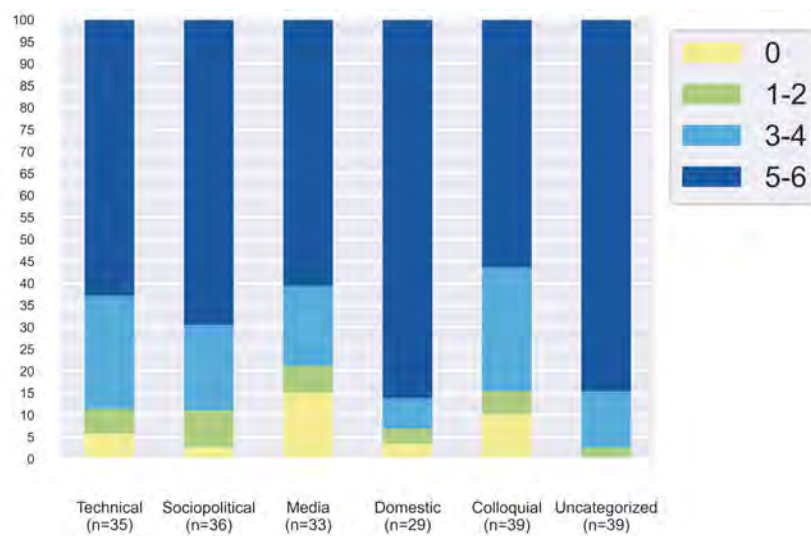


(a) Listedness of ILUs

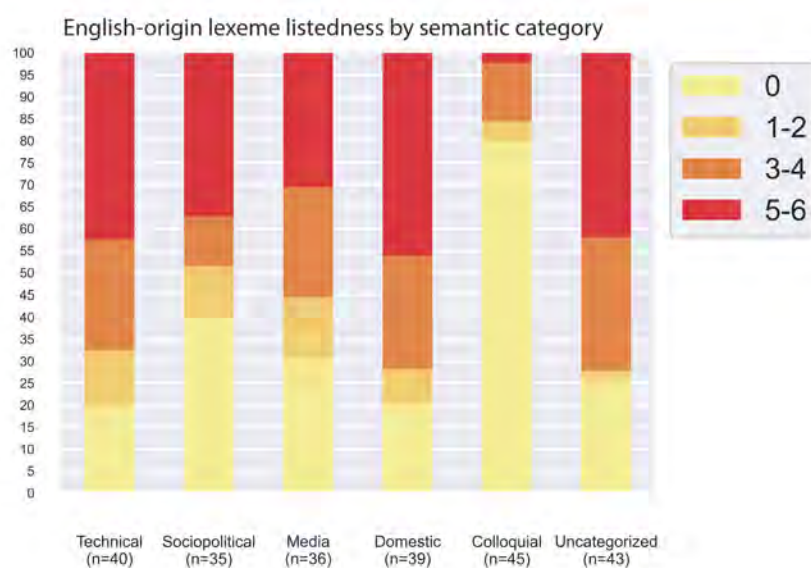


(b) Listedness of ELUs

Figure F.2: Lexeme listedness in GPC by semantic category in the Twitter corpus and subcorpora

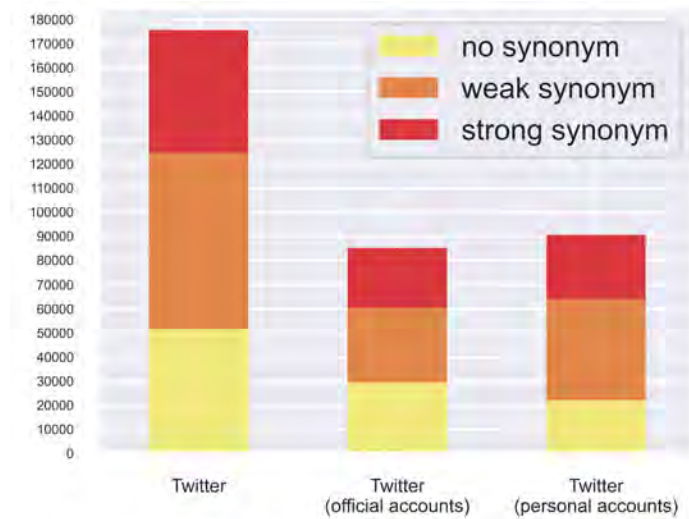


(a) Listedness of ILUs

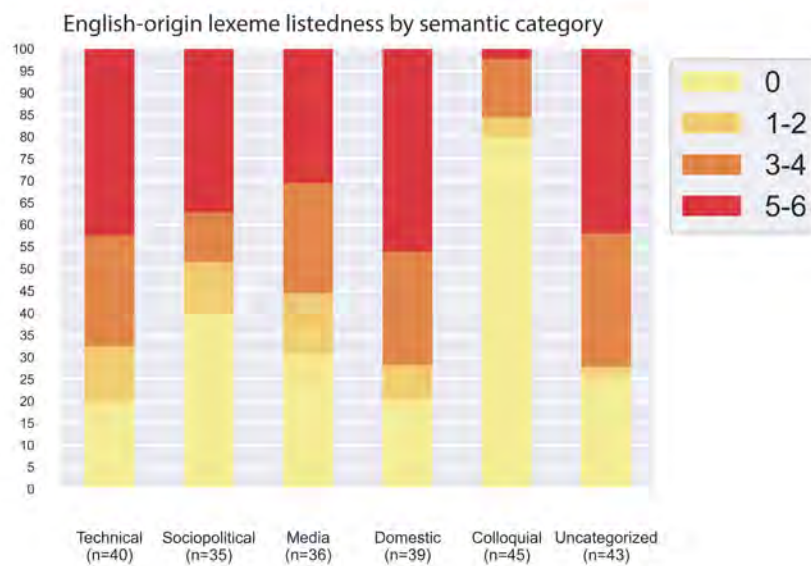


(b) Listedness of ELUs

Figure F.3: Lexeme listedness in six dictionaries by semantic category in the Twitter corpus and subcorpora



(a) ELU lexeme frequency by synonym status



(b) ELU lexeme listedness in six dictionaries by synonym status

Figure F.4: ELU frequency and listedness by synonym status.

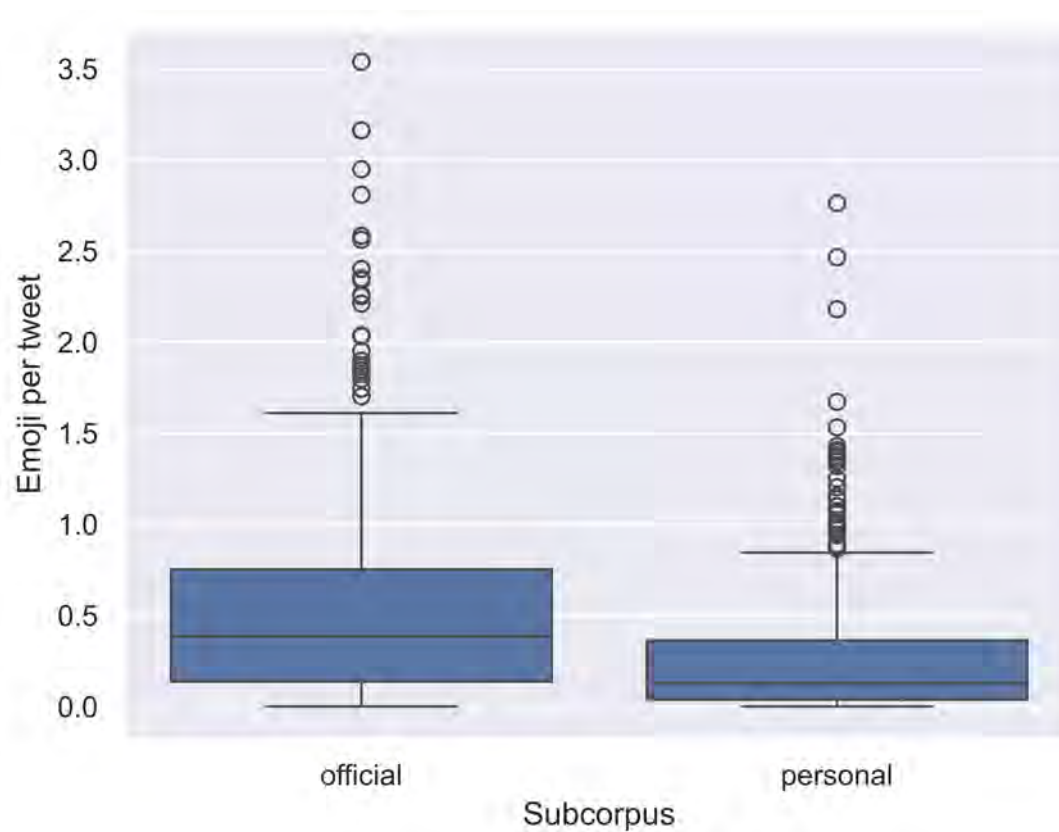
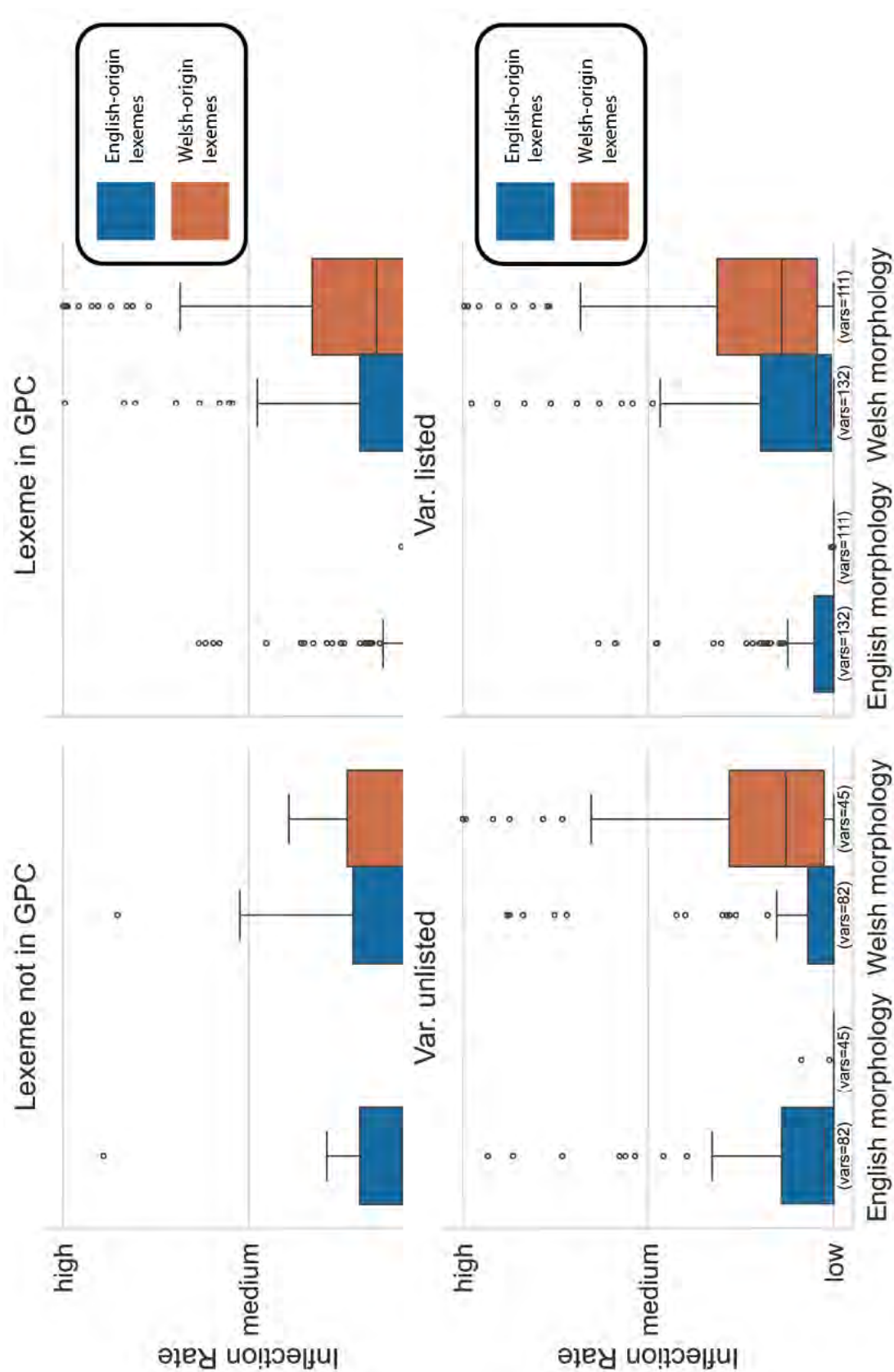


Figure F.5: Average number of emoji per tweet in the official and personal Twitter subcorpora.

F.iii Twitter: comparing linguistic features of ELUs and ILUs in the personal and official subcorpora



(a) ELU and ILU inflection rates by listedness in GPC

(b) ELU and ILU inflection rates by binary listedness status in six dictionaries

Figure F.6: Rates of English and Welsh inflectional morphemes on ELUs and ILUs by listedness status.

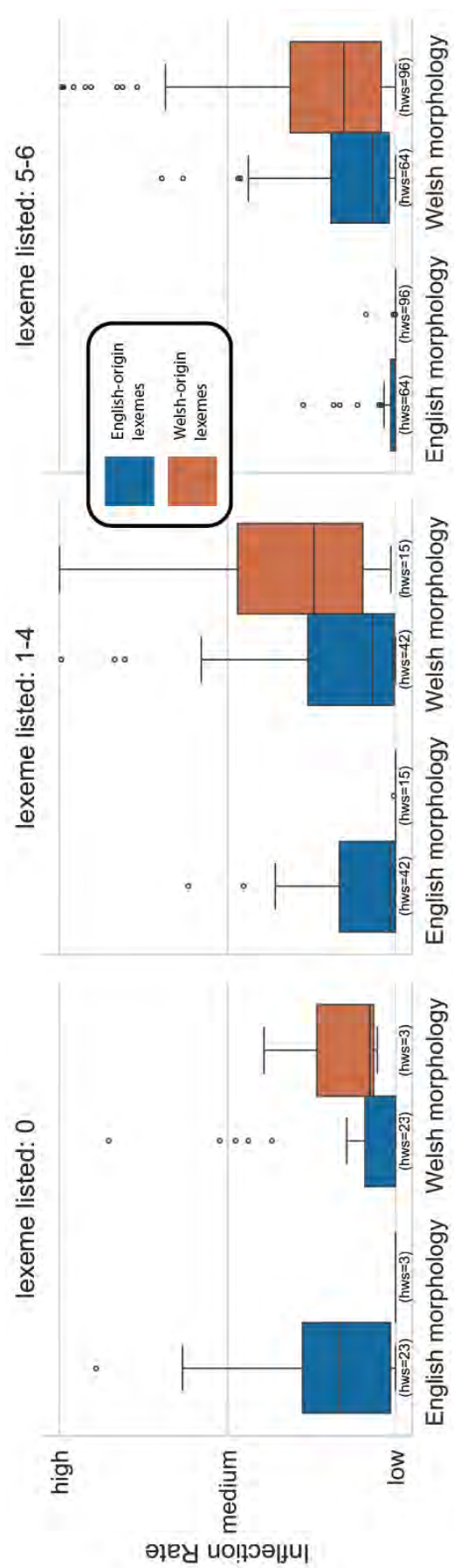


Figure F.7: Rates of inflection on lexemes by listedness status in six dictionaries.

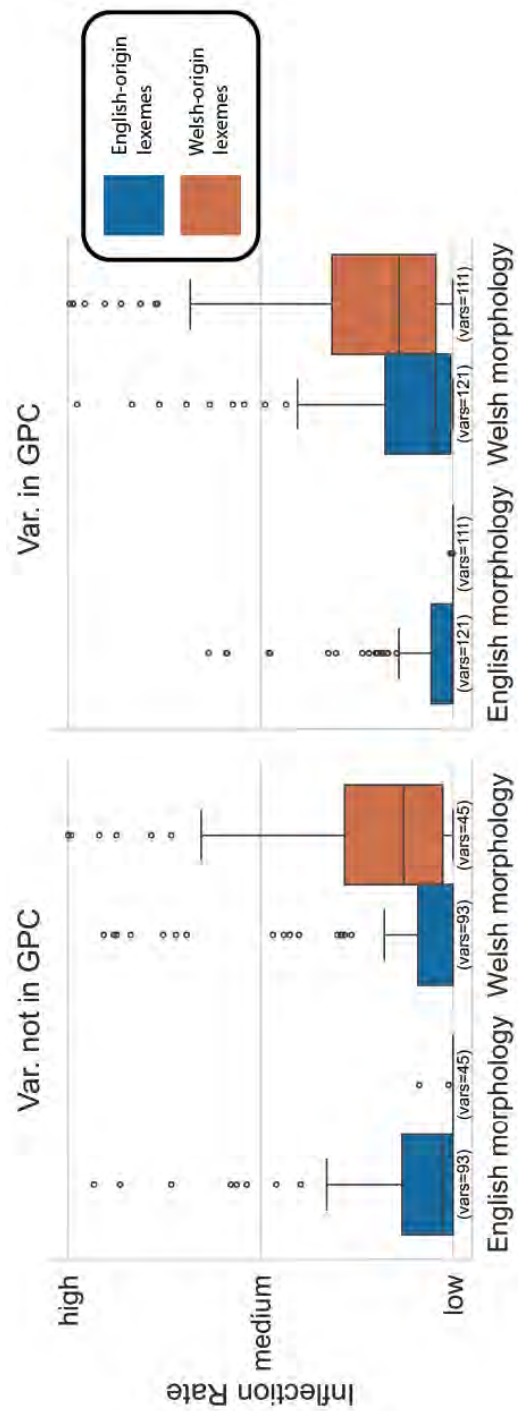


Figure F.8: Rates of inflection on variants by listedness in GPC.

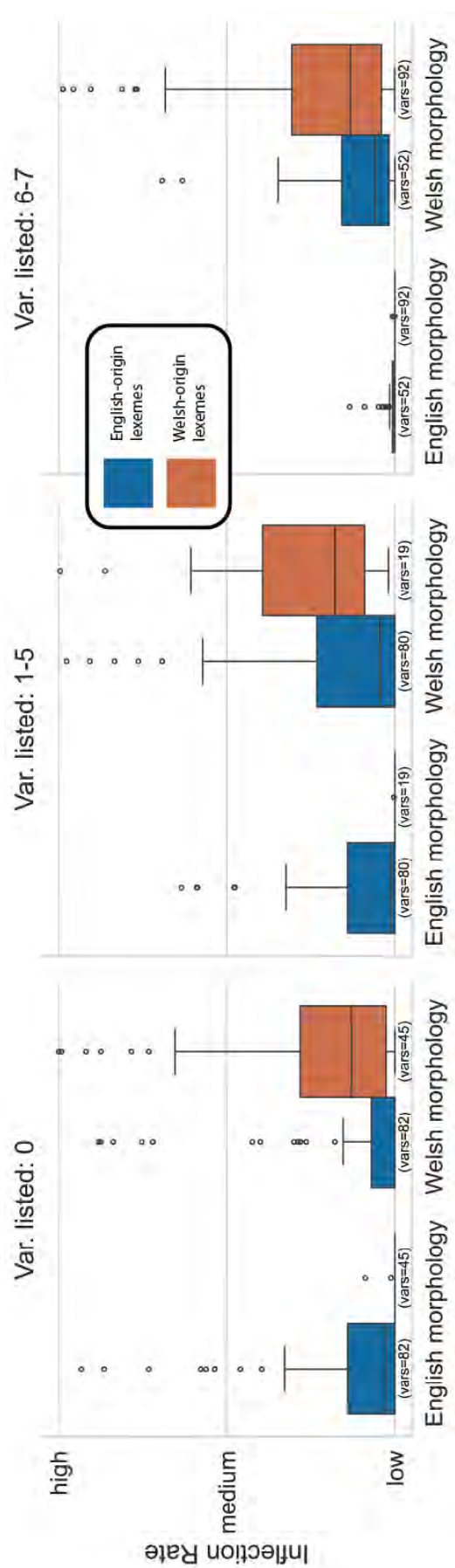
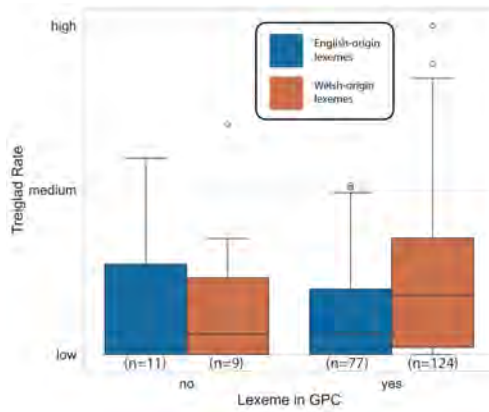
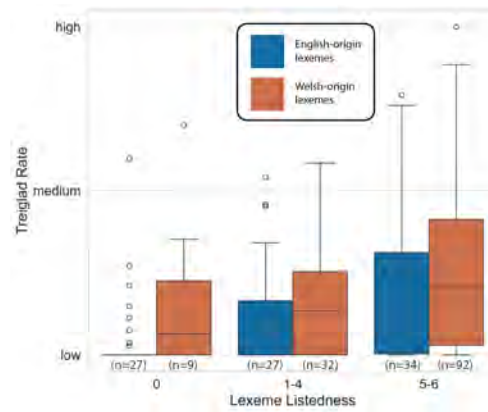


Figure F.9: Rates of inflection on variants by listedness in seven dictionaries.

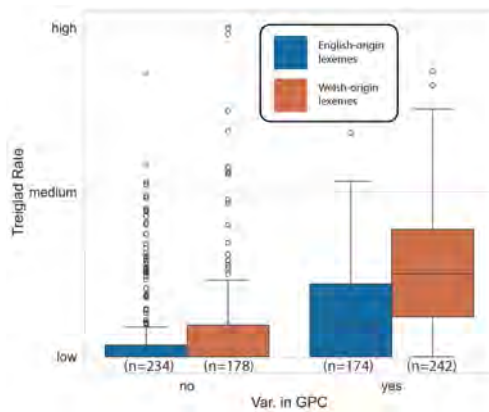


(a) ELU and ILU treiglad rates by listedness in GPC

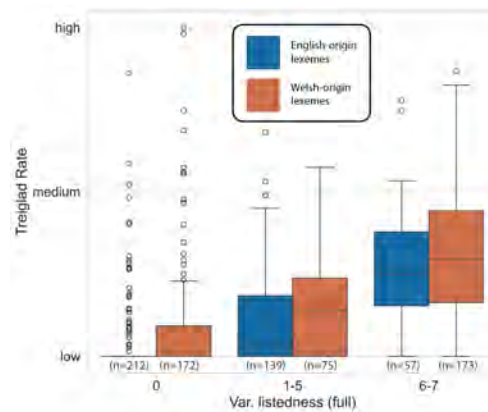


(b) ELU and ILU treiglad rates by listedness in seven dictionaries

Figure F.10: Rates of treiglad on lexemes by listedness status.

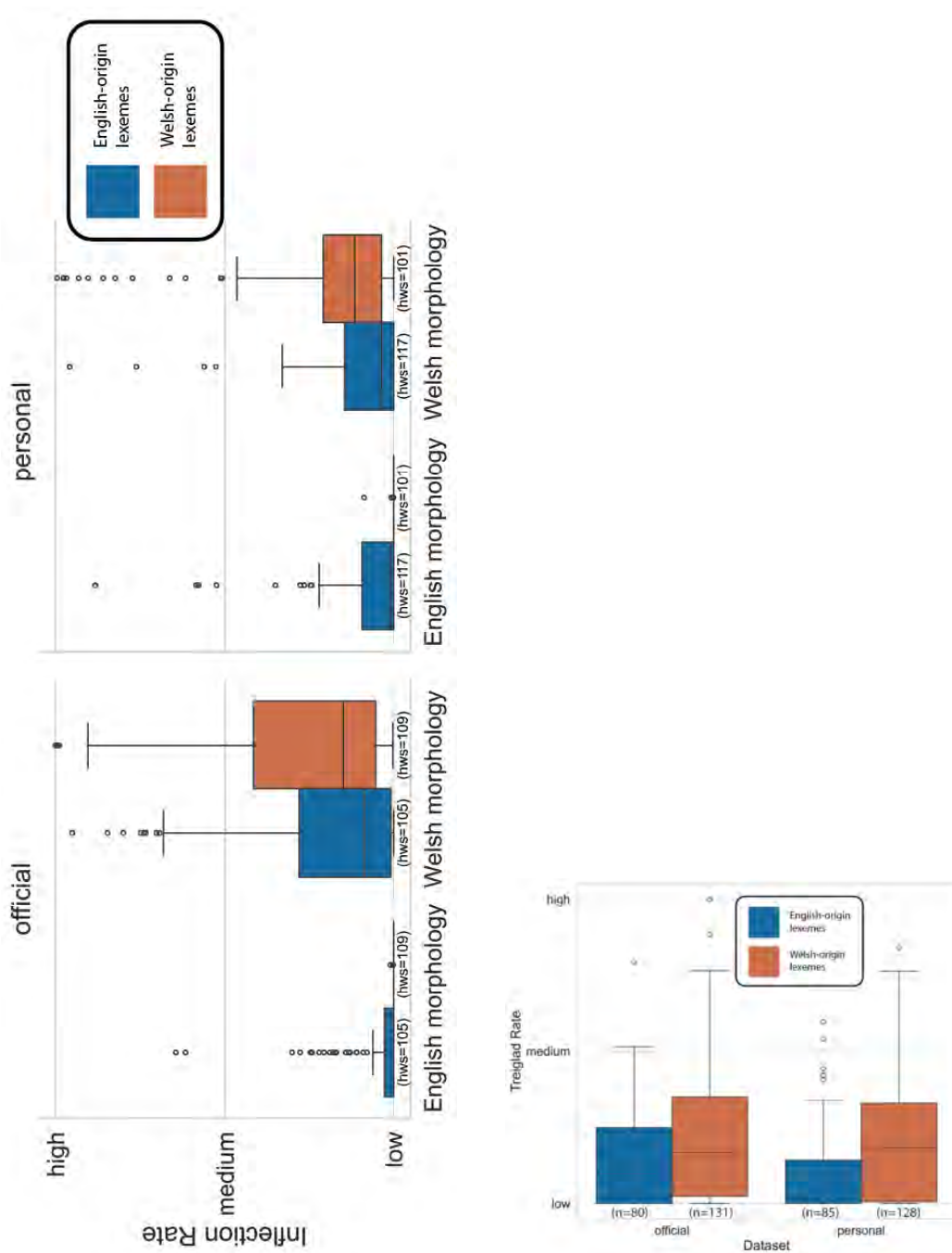


(a) ELU and ILU treiglad rates by listedness in GPC



(b) ELU and ILU treiglad rates by listedness in seven dictionaries

Figure F.11: Rates of treiglad on variants by listedness status.



(a) Inflection rates of English-origin and Welsh-origin lexemes.

(b) Treiglad rates of English-origin and Welsh-origin lexemes.

Figure F.12: Comparisons of inflection and treiglad of ELUs and ILUs, separated by Twitter subcorpus.

F.iv Twitter: list of lexemes selected for close study

Technical lexemes

ELUs

aeroplane	archive	atomic
autistic	cancer	demographic
depress	diagnose	dyslexia
example	fossil	fraction
genetic	immune	insulin
metamorphosis	minimum	molecule
mosquito	negative	nuclear
operation	paracetamol	parameter
potential	prescription	process
psychopath	quarantine	quarter
schizophrenia	square	syndrome
syringe	telescope	therapy
translate	uniform	virus

ILUs

adfail	afiechyd	anhawster
anhwylder	awyren	canran
chwistrell	cyfarwydddeb	cyfeithu
dehongli	demograffaid	diheinbrawf
enghraifft	genynnol	gorbryder
gweddnewidiad	gweithrediad	gwenyn
hunanladdiad	iselder	llawdriniaeth
lleiafswm	llyfrgell	meddyginiaeth
mymryn	negyddol	newidyn
petryal	rhagnodyn	triniaeth
unffurf	ysbienddrych	

Sociopolitical lexemes

ELUs

agenda	auction	barricade
bonus	brexit	canada
commission	conservative	controversy
draft	economy	exclusive
fascist	freelance	homophobia
insurance	lesbian	level
manifesto	pakistan	perspective
petition	photocopy	professional
project	protest	qualify
queer	recruit	referendum
russia	sabotage	session
sexist	stereotype	suffragette
ukip	ukraine ILUs	
ad-drefniant	arwerthiant	braslun

breindal
cyfyngedig
datganiad
ffasgydd
galwedigaethol
lesbiad
llungopïo
penodi
pleidleisio
rhaglen
safon
ymryson

ceidwadol
cymraes
deiseb
ffoadur
gwrthdystio
lesbiaidd
llywodraeth
plaid
pwyl
rhywiaethol
tanseilio
ystrydeb

cyflog
cywaith
eisteddiad
ffrinc
hiliol
llawrydd
mewnfudo
cymru
rhagfarn
safbwynt
ymgymhwyso
yswiriant

Media lexemes

ELUs

algorithm
camera
dance
facebook
format
icon
lottery
music
photoshop
screen
surf
tweet
version
watch
addasiad
broliant
chwilio
cyfrifiadur
cysill
efelychiad
gweplyfr
hapchwarae
lliwgar
podlediad
tafwyl
testun
twmpath

amateur
channel
electric
film
gadget
inbox
melody
novel
psychedelic
scroll
telephone
ukulele
video
youtube

ILUs

alaw
casgliad
clustffon
cyfrinair
cysylltu
ffurf
gwyllo
hunlun
mewnflwch
porwr
teclyn
trydan

blurb
computer
eurovision
football
google
jukebox
mention
opinion
rhythm
selfie
text
upload
violin

barddoniaeth
cerddoriaeth
crybwyll
cyngerdd
delwedd
gorsaf
gwyneblyfr
lawrlwytho
pobl y cwm
pêldroed
telyn
trydar

Domestic lexemes

ELUs

barbecue

bottle

bungalow

caravan	cardboard	catalogue
chandelier	chocolate	cinnamon
cocktail	cupboard	decorate
discount	fridge	hoover
hungry	jumper	laundrette
mortgage	outfit	parlour
picnic	puncture	sausage
scooter	screwdriver	sellotape
shop	shower	sideboard
smoke	smoothie	spatula
sponge	starve	sugar
toilet	tourist	travel
trousers	uncle	velvet
wardrobe ILUs		
addurno	brethyn	cadach
cadair	cegin	cerbyd
cinio	denu	dillad
esgid	ewythr	ffwrn
golchdy	gostyngiad	halen
llawlyfr	llwgu	melys
morthwyl	oergell	pabell
paned	rhwyg	selsig
sugnydd llwch	tafarn	teithio
tâp gludiog	tŷ bach	ymwelydd
ysbodol	ysmygu	

Colloquial lexemes

ELUs

actually	alright	amazeballs
anyway	apparently	awesome
awful	awkward	basically
bastard	beautiful	bloody
blooming	bollocks	bother
bugger	embarrass	fabulous
fantastic	flirt	fuck
genius	genuinely	gibberish
jealous	literally	lovely
nice	nonsense	okay
oops	pathetic	photobomb
please	ridiculous	right
sorry	spooky	superduper
surely	thankyou	wank
wanker	whatever	whatevs

ILUs

anferth	arswydus	athrylith
beth bynnag	braf	cebyst
cenfigennus	chwerthinllyd	chwithig

cyboli	cythraul	cywilydd
diawledig	dichon	diolch
dymunol	go iawn	iawn
lembo	lletchwith	llythrennol
malu awyr	malu cachu	mawredd
mewn gwirionedd	myn brain i	myn uffern i
o'r gorau	ofnadwy	os gwelwch yn dda
penigamp	rhagorol	rhyfeddol
sothach	ta beth	tebygol
truenus	twll tin	yffach
yn y bôn	yn ôl pob sôn	

Uncategorized lexemes

ELUs

behave	bounce	cancel
champion	describe	desperate
drive	ecstatic	emotional
exploit	flap	focus
friend	help	honest
hooligan	join	kaleidoscope
legend	massive	nervous
obscure	option	panic
power	private	problem
psychic	purpose	reality
reason	round	scribble
security	shrug	sound
sprint	stop	stupid
threaten	trampoline	whip

ILUs

atal	bwriad	bwrw
bygwth	canolbwyntio	chwedl
cyfaill	cymorth	cynorthwyo
dewisiad	diddymu	diffuant
dihiryn	diobaith	dirgel
dychryn	egluro	enfawr
esboniad	ffwdan	goruwchnaturiol
gwibio	gwirion	gwirionedd
gyrru	llesmeiriol	mabolgamp
neilltuedig	o gwmpas	pencampwr
pryderus	sbonc	teimladwy
traed brain	trafferth	ymddwyn
ymelwa	ymuno	ysgwyd

F.v Twitter: variant forms

{actually}

	total	personal	official
<actiwali>	74	73	1
<actiwli>	38	36	2
<actwali>	21	21	0
<acshiwali>	8	6	2
<acshwli>	4	4	0
<acshwali>	3	3	0
<acshiwli>	2	2	0
<actwili>	2	2	0
<acshli>	1	1	0
<acsiwali>	1	1	0
<actiwaly>	1	1	0
<actiwly>	1	1	0
<actsiwli>	1	1	0

{aeroplane}

	total	personal	official
<aeroplane>	11	10	1
<eroplen>	3	3	0
<eroplên>	2	2	0
<aeroplên>	1	0	1
<êroplên>	1	1	0

{agenda}

	total	personal	official
<agenda>	1462	170	1292

{algorithm}

	total	personal	official
<algorithm>	37	18	19
<algorythm>	3	2	1
<algorhythm>	1	0	1

{alright}

	total	personal	official
<olreit>	58	57	1
<oreit>	39	35	4
<alright>	25	23	2
<orait>	4	4	0

<allright>	2	2	0
<alreit>	1	1	0
<awright>	1	1	0
<olrait>	1	1	0
<olréit>	1	1	0

{amateur}

	total	personal	official
<amatur>	52	15	37
<amateur>	13	13	0
<amatyr>	2	2	0
<amature>	1	1	0

{amazeballs}

	total	personal	official
<amazeballs>	25	24	1
<amazeballz>	1	1	0
<amêsbols>	1	1	0
<ymezbôls>	1	1	0

{anyway}

	total	personal	official
<eniwe>	728	714	14
<anyway>	224	221	3
<eniwei>	27	26	1
<eniwê>	19	19	0
<aniwe>	13	13	0
<anywe>	6	3	3
<anywei>	1	1	0
<eniwé>	1	1	0

{apparently}

	total	personal	official
<apparently>	169	165	4
<aparyntli>	38	38	0
<aparantli>	14	14	0
<aparentli>	10	10	0
<apparantly>	3	3	0
<aparently>	1	1	0
<aparentyli>	1	1	0
<aparyntlî>	1	1	0
<apparantli>	1	1	0
<apparentli>	1	1	0
<apparyntli>	1	1	0

<apperntli>	1	1	0
{archive}			
	total	personal	official
<archif>	1237	245	992
<archive>	21	9	12
{atomic}			
	total	personal	official
<atomig>	8	2	6
<atomic>	4	4	0
<attomic>	1	1	0
{auction}			
	total	personal	official
<ocsiwn>	161	57	104
<auction>	7	4	3
<ocshiwn>	2	2	0
<ocshwn>	1	1	0
<octsiwn>	1	1	0
{autistic}			
	total	personal	official
<awtistig>	56	10	46
<autistic>	3	3	0
{awesome}			
	total	personal	official
<awesome>	242	222	20
<osym>	102	99	3
<ôsym>	54	44	10
<awsome>	37	36	1
<oswm>	19	19	0
<orswm>	12	12	0
<ossum>	2	2	0
<awswm>	1	1	0
<awsym>	1	1	0
<orsym>	1	1	0
<ôsum>	1	1	0
{awful}			
	total	personal	official

<awful>	123	120	3
<offwl>	7	7	0
<offyl>	7	7	0
<awfful>	1	1	0
<awfyl>	1	1	0
<owfyl>	1	0	1
<ôffyl>	1	1	0

{awkward}

	total	personal	official
<awkward>	117	117	0
<ocwyrđ>	7	6	1
<ôcwyrđ>	2	1	1
<owcwyrđ>	1	1	0

{barbecue}

	total	personal	official
<barbeciw>	124	26	98
<barbyciw>	8	5	3
<barbiciw>	7	5	2
<barbecue>	3	2	1
<barbecîw>	2	0	2

{barricade}

	total	personal	official
<baricêđ>	1	1	0
<barricade>	1	0	1

{basically}

	total	personal	official
<basically>	136	131	5
<beisicli>	7	7	0
<besicli>	5	5	0
<basicali>	3	3	0
<basicyli>	1	1	0
<bêsicli>	1	1	0

{bastard}

	total	personal	official
<basdad>	255	253	2
<bastard>	65	65	0
<basdard>	3	3	0

<basterd>	3	3	0
<basdyd>	2	2	0
<bastart>	1	1	0
<bastyrd>	1	1	0

{beautiful}

	total	personal	official
<beautiful>	79	59	20
<biwtiffwl>	77	72	5
<biwtiffyl>	21	21	0
<biwriffwl>	14	14	0
<biwtifwl>	6	6	0
<biwriffyl>	2	2	0
<bootiful>	2	2	0
<biwtiful>	1	1	0

{behave}

	total	personal	official
<bihaf>	162	133	29
<byhaf>	52	46	6
<bahaf>	11	11	0
<behave>	5	2	3
<bihef>	1	1	0
<buhaf>	1	1	0

{bloody}

	total	personal	official
<blydi>	2323	2264	59
<bloody>	126	121	5
<blwdi>	45	45	0
<bludi>	5	5	0

{blooming}

	total	personal	official
<blwmin>	113	91	22
<bloomin>	15	15	0
<blymin>	12	9	3
<blooming>	9	7	2
<blwming>	6	5	1
<blwmin>	1	1	0

{blurb}

	total	personal	official
<blyrb>	4	2	2
<blurb>	1	0	1

{bollocks}

	total	personal	official
<bolycs>	84	80	4
<bollocks>	60	59	1
<bolocs>	5	5	0
<bollocs>	2	2	0
<bolloks>	1	1	0
<bollucks>	1	1	0

{bonus}

	total	personal	official
<bonws>	233	111	122
<bonus>	59	52	7
<bonys>	2	2	0

{bother}

	total	personal	official
<boddr>	70	68	2
<bother>	35	35	0
<bothr>	9	8	1
<bodder>	6	6	0
<boddyr>	5	5	0
<bothyr>	2	2	0
<bothar>	1	1	0

{bottle}

	total	personal	official
<potel>	334	258	76
<botel>	280	224	56
<botal>	113	111	2
<bottle>	36	26	10
<botl>	5	5	0

{bounce}

	total	personal	official
<bowns>	110	78	32

<bounce>	22	11	11
{brexit}			
	total	personal	official
<brexit>	2296	1138	1158
<brecsut>	12	12	0
<bregsit>	6	6	0
<brecsit>	5	4	1
<brecsùt>	1	1	0
{bugger}			
	total	personal	official
<bugger>	59	58	1
<bygr>	30	30	0
<bwger>	4	4	0
<bygggar>	1	1	0
{bungalow}			
	total	personal	official
<byngalo>	14	9	5
<bungalow>	5	4	1
<bynglo>	3	2	1
<byngalow>	2	2	0
{camera}			
	total	personal	official
<camera>	1033	649	384
<camra>	42	37	5
<camaera>	1	1	0
{canada}			
	total	personal	official
<canada>	311	153	158
{cancel}			
	total	personal	official
<cansl>	772	360	412
<cancel>	14	11	3

{cancer}

	total	personal	official
<canser>	735	57	678
<cancr>	147	70	77
<cansyr>	69	63	6
<cancer>	57	29	28
<cansar>	9	7	2
<cansr>	4	3	1

{caravan}

	total	personal	official
<carafan>	498	284	214
<caravan>	49	40	9
<carafán>	30	7	23

{cardboard}

	total	personal	official
<cardfwrdd>	43	21	22
<cardbord>	23	5	18
<cardboard>	7	6	1
<carbord>	3	3	0
<carboard>	1	0	1
<cardbod>	1	1	0
<cardbwrrdd>	1	1	0
<carfwrdd>	1	1	0

{catalogue}

	total	personal	official
<catalog>	309	85	224
<catalogue>	30	23	7

{champion}

	total	personal	official
<champion>	392	326	66
<jampion>	2	2	0
<siampion>	2	2	0
<tshiampion>	1	1	0
<tsiampion>	1	1	0

{chandelier}

	total	personal	official
<chandelier>	5	4	1

<chanderlier>	1	1	0
<siandelier>	1	0	1

{channel}

	total	personal	official
<sianel>	1450	620	830
<channel>	256	188	68
<siannel>	2	0	2

{chocolate}

	total	personal	official
<siocled>	672	412	260
<chocolate>	160	134	26
<chocled>	47	47	0
<choclet>	22	22	0
<jocled>	9	9	0
<choclit>	2	2	0
<joclet>	2	2	0
<choclat>	1	1	0
<chocolet>	1	1	0
<tshoclet>	1	1	0
<tsiocled>	1	1	0

{cinnamon}

	total	personal	official
<sinamon>	16	10	6
<cinnamon>	12	12	0
<sinamwn>	6	0	6
<cinamwn>	1	1	0
<synamon>	1	1	0

{cocktail}

	total	personal	official
<coctel>	96	39	57
<coctêl>	66	43	23
<cocktail>	64	60	4
<cocteil>	7	5	2
<coctail>	2	2	0
<coctël>	1	1	0

{commission}

	total	personal	official
<comisiwn>	659	224	435

<comisiyn>	426	119	307
<commission>	39	37	2
<comisyn>	9	9	0
<comiswn>	5	5	0
<commision>	2	2	0
<comishwn>	1	1	0
<comission>	1	1	0

{computer}

	total	personal	official
<computer>	56	52	4
<compiwtar>	15	14	1
<compiwter>	1	1	0
<computer>	1	1	0

{conservative}

	total	personal	official
<conservative>	49	25	24
<consyrfatif>	1	1	0

{controversy}

	total	personal	official
<controfersi>	2	1	1
<controfyrsi>	1	1	0

{cupboard}

	total	personal	official
<cwprwdd>	289	163	126
<cupboard>	5	3	2
<cwbwrt>	2	2	0
<cwprwrt>	2	2	0
<cwprwrth>	2	2	0

{dance}

	total	personal	official
<dawns>	2071	707	1364
<dance>	270	129	141
<downs>	2	2	0

{decorate}

total	personal	official

<decorate>	13	6	7
<decorat>	6	6	0
<decyret>	1	1	0

{demographic}

	total	personal	official
<demograffig>	12	8	4
<demographic>	3	3	0
<demograffic>	1	1	0

{depress}

	total	personal	official
<depress>	127	124	3
<depres>	13	13	0
<dipres>	11	11	0
<dipress>	4	4	0

{describe}

	total	personal	official
<disgrif>	707	326	381
<describe>	8	2	6
<discrif>	2	2	0
<descrif>	1	1	0
<discraib>	1	1	0
<disgreif>	1	1	0

{desperate}

	total	personal	official
<desperate>	57	53	4
<despret>	32	30	2
<desbret>	10	9	1
<despryt>	7	7	0
<desprate>	5	5	0
<desparêt>	3	3	0
<desperet>	3	2	1
<desprêt>	3	3	0
<desparet>	2	2	0
<desbrêt>	1	1	0
<desparate>	1	1	0
<desperêt>	1	1	0
<desprat>	1	1	0

{diagnose}

	total	personal	official
<diagnos>	20	2	15

{discount}

	total	personal	official
<disgownt>	78	33	45
<discount>	37	27	10
<discownt>	17	17	0

{draft}

	total	personal	official
<drafft>	952	176	776
<draft>	41	23	18

{drive}

	total	personal	official
<dreif>	516	494	22
<drive>	170	117	53
<dref>	9	9	0
<draif>	3	3	0
<drief>	1	1	0

{dyslexia}

	total	personal	official
<dyslecsia>	35	8	27
<dyslexia>	5	4	1

{economy}

	total	personal	official
<economi>	1435	407	1028
<economy>	39	21	18

{ecstatic}

	total	personal	official
<ecstatic>	9	9	0
<ecstatig>	2	1	1
<egstatig>	1	1	0
<extatic>	1	1	0

{electric}

	total	personal	official
<electric>	29	23	6
<letrig>	11	11	0
<electrig>	6	3	3
<lectrig>	3	3	0

{embarrass}

	total	personal	official
<embaras>	97	64	33
<embarrass>	59	56	3

{emotional}

	total	personal	official
<emosiynol>	702	323	379
<emotional>	52	40	12
<emosiynnol>	19	13	6
<emotiynol>	4	3	1
<emosiyniol>	2	2	0
<emoshional>	1	0	1
<emoshionyl>	1	1	0
<emoshiynyl>	1	1	0
<emoshynol>	1	1	0
<emoshynyl>	1	1	0
<emosionol>	1	1	0
<emosiynal>	1	1	0
<emosynol>	1	1	0

{eurovision}

	total	personal	official
<eurovision>	152	110	42
<ewrofision>	5	5	0
<eurofision>	2	2	0
<iwrofishyn>	2	2	0
<ewrofishion>	1	1	0
<ewrofishon>	1	1	0
<ewrofisiyn>	1	1	0
<ewrovision>	1	1	0
<iwrofision>	1	1	0
<iwrofisiwn>	1	1	0
<iwrofisiyn>	1	1	0

{example}

	total	personal	official
<esiempl>	498	334	164
<example>	26	16	10
<eisempl>	2	1	1
<ecsiempl>	1	0	1
<egsampler>	1	1	0
<eisiempl>	1	1	0

{exclusive}

	total	personal	official
<ecsgliwsif>	102	7	95
<exclusive>	33	18	15
<ecsglwsif>	12	3	9
<ecslwsif>	11	3	8
<egslwsif>	10	4	6
<egsglwsif>	10	3	7
<ecslwsif>	1	1	0
<egsglwsif>	1	0	1

{exploit}

	total	personal	official
<ecsbloet>	35	14	21
<exploit>	3	2	1
<ecsploit>	1	1	0

{fabulous}

	total	personal	official
<fabulous>	66	56	10
<ffabiwllys>	34	29	5
<ffabiwlws>	16	16	0
<ffabwlyss>	1	1	0

{facebook}

	total	personal	official
<facebook>	2538	725	1813
<ffesbwc>	19	18	1
<ffèsbwc>	8	7	1
<fèsbwc>	5	5	0
<ffeisbwc>	4	4	0
<ffesbwc>	2	2	0
<ffèsbwc>	1	1	0

{fantastic}

	total	personal	official
<ffantastig>	790	479	311
<fantastic>	366	263	103
<ffantastic>	359	293	66
<fantastig>	17	13	4
<ffantasdeg>	3	3	0
<ffantasdic>	1	1	0
<ffantasteg>	1	1	0

{fascist}

	total	personal	official
<fascist>	11	11	0
<ffashist>	4	4	0

{film}

	total	personal	official
<ffilm>	6123	3033	3090
<film>	478	367	111
<ffulm>	1	1	0

{flap}

	total	personal	official
<fflap>	10	8	2
<flap>	4	4	0

{flirt}

	total	personal	official
<fflyrt>	19	14	5
<fflirt>	1	1	0
<flirt>	1	1	0
<flyrt>	1	1	0

{focus}

	total	personal	official
<ffocws>	376	116	260
<focus>	75	42	33
<ffocys>	73	30	43
<focws>	3	1	2
<ffocus>	2	1	1
<focys>	1	1	0

{football}

	total	personal	official
<football>	426	313	113
<ffwtbol>	359	347	12
<ffwtbôl>	3	3	0
<ffwtbal>	1	1	0

{format}

	total	personal	official
<ffformat>	120	73	47
<format>	17	17	0

{fossil}

	total	personal	official
<ffossil>	39	7	32
<fossil>	6	3	3

{fraction}

	total	personal	official
<ffracsiwn>	46	4	42
<fraction>	1	0	1

{freelance}

	total	personal	official
<ffrilans>	4	4	0
<freelance>	4	3	1
<ffrilans>	2	2	0

{fridge}

	total	personal	official
<ffrij>	80	72	8
<fridge>	62	59	3
<fri>	20	20	0
<ffridj>	5	5	0
<fridj>	1	1	0

{friend}

	total	personal	official
<ffrind>	5196	2744	2452

<friend>	269	179	90
<ffrend>	8	8	0
<frind>	8	6	2
<ffrynd>	1	1	0

{fuck}

	total	personal	official
<ffwc>	1436	1433	3
<ffyc>	404	395	9
<fuck>	298	298	0
<ffuc>	3	3	0
<ffwck>	2	2	0
<ffuck>	1	1	0
<ffwc>	1	1	0

{gadget}

	total	personal	official
<gadget>	16	15	1
<gajet>	14	2	12
<gadjet>	7	3	4

{genetic}

	total	personal	official
<genetig>	30	3	27
<genetic>	1	1	0
<ginetig>	1	0	1

{genius}

	total	personal	official
<genius>	180	173	7
<jiniys>	38	35	3
<jîniys>	6	6	0
<genious>	3	3	0
<jeniws>	2	2	0
<jeanius>	1	1	0

{genuinely}

	total	personal	official
<genuinely>	67	63	4
<geniwinli>	2	2	0
<genuienly>	2	2	0
<genuenly>	1	1	0

<jeniwinli>	1	1	0
<jeniwinli>	1	1	0

{gibberish}

	total	personal	official
<gibberish>	5	4	1
<siprys>	4	1	3
<jibberish>	3	3	0
<jiberish>	1	1	0
<jibyrish>	1	1	0

{google}

	total	personal	official
<google>	686	507	179
<gwgl>	104	96	8
<gŵgl>	27	27	0
<googel>	1	1	0
<gwgal>	1	1	0
<gwgył>	1	1	0
<gŵgył>	1	1	0

{help}

	total	personal	official
<help>	19496	4755	14741

{homophobia}

	total	personal	official
<homoffobia>	93	11	82
<homophobia>	5	3	2

{honest}

	total	personal	official
<onest>	779	611	168
<gonest>	156	98	58
<onast>	107	105	2
<honest>	49	46	3
<gonast>	6	5	1
<honast>	5	5	0
<onesd>	1	1	0

{hooligan}

	total	personal	official
<hooligan>	10	10	0
<hwligan>	9	8	1
<hŵligan>	2	2	0

{hoover}

	total	personal	official
<hwfr>	48	47	1
<hoover>	36	32	4
<hwfer>	13	10	3
<hŵfr>	6	6	0
<hŵfyr>	4	4	0
<hŵfer>	1	1	0

{hungry}

	total	personal	official
<hungry>	11	9	2
<hyngri>	1	1	0

{icon}

	total	personal	official
<eicon>	68	27	41
<icon>	41	38	3

{immune}

	total	personal	official
<imiwn>	4	0	4
<immune>	4	4	0
<imune>	1	1	0

{inbox}

	total	personal	official
<inbox>	25	20	5
<inbocs>	2	2	0

{insulin}

	total	personal	official
<inswlin>	10	0	10
<insulin>	2	2	0

{insurance}

	total	personal	official
<insurance>	28	24	4
<insiwrans>	9	9	0
<siwrans>	8	8	0
<inshiwrans>	2	1	1
<inshwrans>	1	1	0

{jealous}

	total	personal	official
<jelys>	179	178	1
<jealous>	102	99	3
<jelws>	4	4	0
<jelas>	2	2	0
<jelos>	1	1	0

{join}

	total	personal	official
<join>	433	238	195

{jukebox}

	total	personal	official
<jiwcbocs>	28	9	19
<jukebox>	21	16	5
<jiwcbbox>	1	1	0
<jŵcbocs>	1	1	0

{jumper}

	total	personal	official
<siwmp̄er>	249	117	132
<jumper>	54	42	12
<jympyr>	18	17	1
<jwmp̄er>	7	7	0
<jymp̄er>	6	4	2
<jymp̄r>	2	2	0
<siwmp̄wr>	1	0	1

{kaleidoscope}

	total	personal	official
<caleidosgop>	2	0	2

{laundrette}

	total	personal	official
<laundrette>	3	3	0
<londret>	1	1	0
<lôndret>	1	1	0

{legend}

	total	personal	official
<legend>	231	201	30
<lejynd>	65	59	6
<lejand>	36	32	4
<lejend>	16	14	2
<ledgend>	2	2	0
<ledjend>	1	1	0
<legynd>	1	0	1

{lesbian}

	total	personal	official
<lesbian>	32	28	4
<lezbian>	1	0	1

{level}

	total	personal	official
<lefel>	3104	750	2354
<level>	175	102	73
<lefal>	12	11	1
<lefl>	2	2	0

{literally}

	total	personal	official
<literally>	160	155	5
<literali>	5	5	0
<litryli>	2	2	0
<literaly>	1	1	0
<literraly>	1	1	0
<literyli>	1	1	0
<litrali>	1	1	0
<litrely>	1	1	0
<litrli>	1	1	0
<litteraly>	1	1	0
<lityrali>	1	1	0

{lottery}

	total	personal	official
<loteri>	291	113	178
<lottery>	16	11	5
<lotri>	1	0	1
<lotterry>	1	1	0
<lotwri>	1	1	0
<lotyri>	1	1	0

{lovely}

	total	personal	official
<lovely>	412	297	115
<lyfli>	15	15	0
<lufli>	5	5	0
<lwfli>	3	3	0
<lyfly>	3	3	0
<llyfli>	2	1	1
<lovley>	2	2	0
<lovelly>	1	0	1
<lovli>	1	1	0
<lwfle>	1	1	0
<lwflee>	1	1	0
<lyflu>	1	1	0

{manifesto}

	total	personal	official
<maniffesto>	306	129	177
<manifesto>	18	12	6

{massive}

	total	personal	official
<massive>	259	242	17
<masif>	128	123	5
<masuf>	18	18	0
<masive>	2	2	0
<massiv>	1	1	0

{melody}

	total	personal	official
<melody>	22	13	9
<melodi>	11	2	9

{mention}

	total	personal	official
<mention>	75	65	10
<menshyn>	3	3	0

{metamorphosis}

	total	personal	official
<metamorffosis>	1	0	1
<metamorphosis>	1	1	0

{minimum}

	total	personal	official
<minimum>	14	12	2
<minimwm>	2	1	1
<minumum>	1	1	0

{molecule}

	total	personal	official
<moleciwl>	4	0	4
<molecule>	1	1	0

{mortgage}

	total	personal	official
<morgais>	57	28	29
<morgej>	3	3	0
<mortgage>	3	3	0
<morgage>	1	1	0
<morgij>	1	1	0
<mortgais>	1	1	0

{mosquito}

	total	personal	official
<mosgito>	8	7	1
<mosquito>	3	3	0
<mosgitô>	1	1	0

{music}

	total	personal	official
<miwsig>	606	290	316
<music>	559	334	225
<miwsic>	110	91	19

<mwsig>	4	4	0
<mywsic>	1	1	0
<mywsig>	1	1	0

{negative}

	total	personal	official
<negatif>	131	41	90
<negative>	25	19	6
<negetif>	1	1	0

{nervous}

	total	personal	official
<nerfus>	217	169	48
<nervous>	51	51	0
<nyrfys>	45	42	3
<nerfys>	24	24	0
<nerfis>	1	1	0
<nurfys>	1	0	1

{nice}

	total	personal	official
<neis>	4235	3901	334
<nice>	448	409	39
<nais>	59	58	1
<neus>	15	14	1

{nonsense}

	total	personal	official
<nonsens>	238	213	25
<nonsense>	22	22	0
<nonsans>	11	11	0

{novel}

	total	personal	official
<nofel>	2381	836	1545
<novel>	27	16	11
<nofal>	6	6	0

{nuclear}

	total	personal	official
<niwclear>	268	98	170

<niwcliar>	79	64	15
<nuclear>	14	10	4

{obscure}

	total	personal	official
<obscure>	10	9	1
<obsciwar>	3	2	1
<obsciwyr>	2	2	0
<obsciwrr>	1	1	0
<obsgiwarr>	1	1	0

{okay}

	total	personal	official
<ok>	1850	1739	111
<ocê>	82	62	20
<okay>	69	67	2
<owcei>	13	13	0
<oce>	10	10	0
<okei>	4	4	0
<okk>	3	3	0
<oecei>	1	1	0
<owcai>	1	1	0
<owkei>	1	1	0

{oops}

	total	personal	official
<wps>	506	431	75
<oops>	42	37	5
<woops>	7	7	0
<whoops>	6	6	0
<ŵps>	5	5	0
<wpss>	2	2	0
<wups>	1	0	1
<wŵps>	1	1	0

{operation}

	total	personal	official
<operation>	45	33	12
<operasiwn>	2	2	0
<opereshyn>	1	1	0

{opinion}

	total	personal	official

<pinewn>	155	99	56
<opinion>	47	42	5
<opiniwn>	3	3	0

{option}

	total	personal	official
<opsiwn>	1283	490	793
<option>	78	51	27
<opshiwn>	1	1	0
<opsiyn>	1	1	0
<opswn>	1	1	0
<optsiwn>	1	1	0

{outfit}

	total	personal	official
<outfit>	50	43	7
<owtffit>	8	8	0

{pakistan}

	total	personal	official
<pacistan>	79	6	73
<pakistan>	17	10	7

{panic}

	total	personal	official
<panic>	192	169	23
<panig>	87	38	49
<panick>	3	3	0

{paracetamol}

	total	personal	official
<paracetamol>	33	26	7
<paracetomol>	2	2	0
<paracetemol>	1	1	0
<paracetimol>	1	1	0
<paracetmol>	1	1	0
<paracetomal>	1	1	0
<parasetamol>	1	0	1

{parameter}

total	personal	official

<paramedr>	2	1	1
<paramater>	1	1	0
{parlour}			
	total	personal	official
<parlwr>	74	38	36
<parlour>	9	7	2
{pathetic}			
	total	personal	official
<pathetic>	92	89	3
<pathetig>	45	42	3
{perspective}			
	total	personal	official
<persbectif>	92	49	43
<perspectif>	52	35	17
<perspective>	18	14	4
{petition}			
	total	personal	official
<petition>	91	67	24
<petisiwn>	2	2	0
<petishiwn>	1	1	0
{photobomb}			
	total	personal	official
<photobomb>	11	9	2
<photobom>	3	3	0
<ffotobom>	2	2	0
<ffotobomb>	1	1	0
{photocopy}			
	total	personal	official
<ffotocopi>	7	5	2
<photocopy>	5	3	2
{photoshop}			
	total	personal	official

<photoshop>	51	45	6
<ffotosiop>	4	4	0
<ffotoshop>	3	2	1

{picnic}

	total	personal	official
<picnic>	262	117	145
<picnig>	1	0	1

{please}

	total	personal	official
<plis>	4880	3803	1077
<plís>	1139	750	389
<please>	358	218	140
<plys>	22	22	0
<plis>	8	8	0
<plís>	1	1	0
<pwease>	1	1	0
<pwis>	1	1	0

{potential}

	total	personal	official
<potensial>	457	196	261
<potential>	41	29	12
<potenshal>	2	2	0
<potenshial>	1	1	0

{power}

	total	personal	official
<pwer>	653	183	470
<pwêr>	236	51	185
<power>	104	85	19
<pwar>	5	5	0
<pwêr>	4	1	3
<pawyr>	1	1	0
<powyr>	1	1	0

{prescription}

	total	personal	official
<presgripsiwn>	129	19	110
<prescription>	8	6	2
<prescripsiwn>	4	2	2
<prescripsiwn>	1	0	1

<prescriptiwn>	1	1	0
<presgripshiwn>	1	1	0

{private}

	total	personal	official
<preifat>	910	271	639
<private>	52	41	11
<preifet>	4	2	2
<prifat>	3	1	2
<preufat>	1	1	0

{problem}

	total	personal	official
<problem>	5297	3066	2231
<problam>	69	69	0

{process}

	total	personal	official
<proses>	1852	439	1413
<process>	22	9	13

{professional}

	total	personal	official
<proffesiynol>	1127	240	887
<professional>	42	26	16
<proffesiynol>	20	16	4
<profesiynol>	6	2	4
<proffesional>	3	3	0
<proffessiynol>	2	0	2
<proffesyynol>	2	2	0
<profesiynol>	1	0	1
<proffeshynol>	1	1	0
<proffessiynol>	1	1	0

{project}

	total	personal	official
<prosiect>	5040	860	4180
<project>	319	108	211
<prosciect>	1	0	1
<prosiecd>	1	1	0

{protest}

	total	personal	official
<protest>	1017	531	486

{psychedelic}

	total	personal	official
<seicadelig>	14	5	9
<seicadelic>	12	7	5
<psychedelic>	6	2	4
<seicedelic>	2	2	0
<psychadelic>	1	1	0

{psychic}

	total	personal	official
<psychic>	16	13	3
<seicic>	11	10	1
<seicig>	5	3	2
<psycic>	2	2	0

{psychopath}

	total	personal	official
<psychopath>	7	7	0
<seicopath>	7	5	2

{puncture}

	total	personal	official
<puncture>	9	8	1
<pynctiwr>	4	4	0
<pynctjyr>	2	2	0
<pynctiar>	1	1	0
<pynctshyr>	1	0	1
<pynjar>	1	1	0

{purpose}

	total	personal	official
<pwrpas>	466	311	155
<purpose>	6	4	2
<pwrpws>	2	1	1
<pwrpais>	1	1	0
<pŵrpas>	1	1	0

{qualify}

	total	personal	official
<qualifi>	17	15	2
<qualify>	16	15	1
<cwaliifei>	3	3	0
<cwalifi>	1	1	0
<qualifai>	1	1	0

{quarantine}

	total	personal	official
<cwarantin>	31	9	22
<cwarantîn>	19	7	12
<quarantine>	9	8	1
<cworantîn>	1	0	1
<quarantîn>	1	1	0
<quarentine>	1	1	0

{quarter}

	total	personal	official
<chwarter>	705	323	382
<chwartar>	45	40	5
<cwarter>	43	20	23
<quarter>	32	23	9
<chwater>	4	2	2
<cwartar>	1	1	0

{queer}

	total	personal	official
<queer>	13	7	6
<cwiar>	8	5	3
<cwiyar>	1	0	1

{racist}

	total	personal	official
<racist>	42	40	2

{reality}

	total	personal	official
<realiti>	340	143	197
<reality>	64	49	15
<reailiti>	1	0	1

{reason}

	total	personal	official
<rheswm>	2364	1680	684
<reason>	41	29	12
<rhesym>	21	19	2
<rheswn>	15	13	2
<rhyswm>	1	0	1

{recruit}

	total	personal	official
<recrwt>	1554	106	1448
<ricrwt>	34	33	1
<recrwt>	31	1	30
<recrwt>	6	0	6
<recruit>	5	4	1
<recrwt>	4	2	2
<recrewt>	1	1	0
<recrut>	1	1	0
<recrwt>	1	1	0

{referendum}

	total	personal	official
<refferendwm>	891	522	369
<referendum>	27	22	5
<reffrendwm>	5	5	0
<reffarendwm>	1	1	0

{refugee}

	total	personal	official
<refugee>	19	10	9

{reshuffle}

	total	personal	official
<reshuffle>	7	5	2

{rhythm}

	total	personal	official
<rhythm>	34	16	18
<rythm>	9	5	4
<rhythm>	2	2	0
<rhyddam>	1	0	1
<riddim>	1	0	1

<riddm>	1	0	1
<ryddym>	1	1	0

{ridiculous}

	total	personal	official
<ridiculous>	51	48	3
<ridicilys>	28	28	0
<ridicilous>	27	27	0
<ridiciwlys>	3	3	0
<ridiclws>	2	2	0
<ridicylys>	1	1	0

{right}

	total	personal	official
<reit>	2783	2544	239
<right>	270	234	36
<rheit>	1	1	0

{round}

	total	personal	official
<rownd>	4553	2203	2350
<round>	210	160	50

{russia}

	total	personal	official
<rwsia>	488	278	210
<russia>	35	32	3
<rusia>	1	1	0

{sabotage}

	total	personal	official
<sabotage>	3	3	0
<sabotaj>	3	3	0

{sausage}

	total	personal	official
<sosej>	119	104	15
<sausage>	51	45	6
<sosij>	27	26	1
<sosěj>	2	2	0

{schizophrenia}

	total	personal	official
<sgitsoffrenia>	21	1	20
<schizophrenia>	1	0	1

{scooter}

	total	personal	official
<sgwter>	79	10	69
<scooter>	14	12	2

{screen}

	total	personal	official
<sgrin>	815	374	441
<sgrîn>	319	161	158
<screen>	128	93	35
<sgrün>	4	4	0
<scrîn>	2	2	0
<scrin>	1	1	0

{screwdriver}

	total	personal	official
<sgriwdreifar>	3	2	1
<sgriwdreifer>	2	0	2
<screwdriver>	1	1	0
<sgriwdreifyr>	1	1	0

{scribble}

	total	personal	official
<sgribl>	23	21	2
<scribble>	7	4	3

{scroll}

	total	personal	official
<sgrol>	44	15	29
<scroll>	11	9	2
<sgrôl>	5	1	4

{security}

	total	personal	official
<security>	26	21	5

<seciwriti>	3	3	0
<siciwriti>	1	1	0

{selfie}

	total	personal	official
<selfie>	164	125	39
<selffi>	40	22	18
<selffie>	3	3	0
<selfi>	3	3	0
<selffi>	1	0	1

{sellotape}

	total	personal	official
<sellotape>	7	7	0
<selotep>	5	5	0
<selotêp>	5	3	2

{session}

	total	personal	official
<sesiwn>	8454	1271	7183
<session>	210	59	151
<sessiwn>	72	56	16
<seshwn>	27	24	3
<seshiwn>	15	8	7
<sesiyn>	4	0	4
<seshan>	2	2	0
<seshawn>	1	1	0
<seshion>	1	0	1
<seshwn>	1	1	0

{sexist}

	total	personal	official
<sexist>	37	33	4
<secsist>	4	4	0

{shop}

	total	personal	official
<siop>	8617	3237	5380
<shop>	373	271	102
<siôp>	4	4	0

{shower}

	total	personal	official
<shower>	85	76	9
<shawyr>	1	1	0

{shrug}

	total	personal	official
<shrug>	4	1	3
<shryg>	2	2	0

{sideboard}

	total	personal	official
<seidbord>	2	0	2

{smoke}

	total	personal	official
<smoc>	107	100	7
<smoke>	23	19	4
<smôc>	15	13	2
<ysmoc>	1	1	0

{smoothie}

	total	personal	official
<smwddi>	20	3	17
<smoothie>	10	8	2
<smŵddi>	2	1	1
<smwddî>	1	0	1

{sorry}

	total	personal	official
<sori>	3908	3490	418
<sorri>	209	203	6
<sorry>	118	109	9
<sorei>	1	1	0

{sound}

	total	personal	official
<sownd>	301	207	94
<sound>	258	179	79
<sown>	3	3	0

<sawnd>	2	2	0
{spatula}			
	total	personal	official
<sbatiwla>	1	1	0
<spatula>	1	0	1
{sponge}			
	total	personal	official
<sponge>	19	16	3
<sbwng>	14	7	7
<sbwnj>	8	7	1
<sbynj>	8	5	3
<spwng>	7	6	1
<sbyng>	1	0	1
<spwnj>	1	1	0
<spynj>	1	1	0
{spooky}			
	total	personal	official
<sbwci>	39	26	13
<spooky>	34	26	8
<sbwci>	6	1	5
<spwci>	6	3	3
<spooki>	1	0	1
<spwci>	1	1	0
{sprint}			
	total	personal	official
<sprint>	41	35	6
<sbrint>	16	12	4
{square}			
	total	personal	official
<sgwâr>	334	86	248
<sgwar>	263	183	80
<square>	116	65	51
{starve}			
	total	personal	official
<starf>	39	30	9

<starve>	1	1	0
{stereotype}			
	total	personal	official
<stereoteip>	32	12	20
<stereotype>	9	8	1
<sterioteip>	2	1	1
<sterieoteip>	1	1	0
<stéréotype>	1	1	0
{stop}			
	total	personal	official
<stop>	2516	1866	650
<sdop>	15	13	2
{stupid}			
	total	personal	official
<stupid>	135	124	11
<stiwpid>	55	53	2
<sdiwpid>	15	15	0
<stwpid>	5	5	0
<stoopid>	3	3	0
<sdiwpud>	2	2	0
<sdiwpyd>	2	2	0
<shtoopid>	2	2	0
<stiwpud>	1	1	0
<stiwpyd>	1	1	0
{suffragette}			
	total	personal	official
<suffragette>	16	5	11
<swffraget>	16	2	14
<swffragét>	13	0	13
<syffrajét>	5	0	5
<syffrajét>	2	0	2
<swffrajét>	1	0	1
<syffrajét>	1	0	1
<syfrajét>	1	1	0
{sugar}			
	total	personal	official
<siwgr>	262	111	151

<siwgwr>	110	94	16
<sugar>	51	44	7
<shwgwr>	6	6	0
<shugar>	2	0	2
<shwgur>	2	2	0
<siwgir>	2	2	0
<shwgir>	1	1	0
<shwgyr>	1	1	0
<siwger>	1	1	0

{suicide}

	total	personal	official
<suicide>	22	17	5

{superduper}

	total	personal	official
<superduper>	13	11	2
<swperdwper>	1	1	0
<swpyrdwpyr>	1	0	1

{surely}

	total	personal	official
<surely>	63	55	8
<shwarli>	3	3	0
<shwrli>	3	1	2
<shwyrli>	1	1	0
<surley>	1	1	0

{surf}

	total	personal	official
<syrff>	152	58	94
<surf>	26	15	11

{syndrome}

	total	personal	official
<syndrom>	54	12	42
<syndrome>	29	24	5
<sundrome>	1	1	0
<syndrôm>	1	1	0

{syringe}

	total	personal	official
<srinj>	2	2	0
<syringe>	2	2	0
<syrinj>	1	1	0

{telephone}

	total	personal	official
<teleffon>	10	2	8
<telephone>	6	4	2
<teliffon>	4	3	1
<teliffôn>	1	1	0

{telescope}

	total	personal	official
<telesgop>	57	23	34
<telesgôp>	2	2	0
<telescôp>	1	1	0
<telesgob>	1	1	0
<telisgop>	1	1	0

{text}

	total	personal	official
<text>	326	299	27
<tecst>	301	135	166
<tecsd>	8	8	0

{thankyou}

	total	personal	official
<thankyou>	324	144	180
<thanciw>	63	62	1
<thenciw>	38	33	5
<ffanciw>	21	20	1
<thancew>	2	2	0
<thancyw>	1	1	0

{therapy}

	total	personal	official
<therapi>	281	43	238
<therapy>	22	16	6

{threaten}

	total	personal	official
<threaten>	3	3	0
<thretn>	1	1	0
<thretyn>	1	1	0

{toilet}

	total	personal	official
<toiled>	293	143	150
<toilet>	166	151	15

{tourist}

	total	personal	official
<twrist>	121	76	45
<tourist>	41	36	5

{trampoline}

	total	personal	official
<trampolin>	20	7	13
<trampolîn>	13	5	8
<trampoline>	6	5	1

{translate}

	total	personal	official
<translate>	159	131	28
<translêt>	10	9	1
<translat>	5	5	0
<translet>	3	3	0
<transleit>	1	1	0

{travel}

	total	personal	official
<trafeil>	102	84	18
<travel>	82	60	22
<trafael>	22	15	7
<trafel>	12	11	1
<trafal>	4	4	0
<trafyl>	4	4	0

{trousers}

	total	personal	official
<trowsus>	111	95	16
<trwsus>	65	57	8
<trouser>	10	10	0
<trwser>	10	10	0
<trousers>	5	3	2
<trywsus>	3	3	0

{tweet}

	total	personal	official
<tweet>	1792	1636	156
<twî<	561	475	86
<twit>	108	96	12
<twî<	4	2	2
<twî<	1	1	0
<twî<	1	1	0

{ukip}

	total	personal	official
<ukip>	668	451	217
<iwcip>	16	13	3

{ukraine}

	total	personal	official
<wcraïn>	79	52	27
<wcráïn>	51	16	35
<ukraine>	23	21	2
<iwcraïn>	16	14	2
<wcraen>	8	4	4
<iwcraen>	6	3	3
<wcráïn>	1	0	1
<iwcraïn>	1	1	0
<wcráïn>	1	1	0

{ukulele}

	total	personal	official
<ukulele>	80	10	70
<ukelele>	23	9	14
<iwcalili>	15	1	14
<iwcaleli>	4	1	3
<iwcalele>	3	0	3
<iwceleli>	3	0	3
<iwcaleili>	1	0	1

{uncle}

	total	personal	official
<yncl>	146	134	12
<wncwl>	86	71	15
<uncle>	31	28	3
<wncl>	12	9	3
<uncl>	4	4	0

{uniform}

	total	personal	official
<uniform>	47	21	26
<iwniffom>	19	7	12
<iwniffôrm>	1	0	1

{upload}

	total	personal	official
<upload>	54	29	25
<yplod>	1	1	0

{velvet}

	total	personal	official
<velvet>	21	13	8
<melfed>	13	11	2
<felfed>	2	1	1

{version}

	total	personal	official
<fersiwn>	1878	1179	699
<version>	99	79	20
<fersion>	2	2	0
<ferswn>	1	0	1
<fyrsion>	1	1	0

{video}

	total	personal	official
<fideo>	6115	1521	4594
<video>	566	320	246
<fidio>	282	159	123
<vidio>	1	1	0

{violin}

	total	personal	official
<feiolin>	18	5	13
<violin>	6	4	2
<fiolin>	3	2	1

{virus}

	total	personal	official
<feirws>	491	119	372
<firws>	150	30	120
<virus>	39	38	1
<feirys>	20	4	16
<veirws>	2	1	1
<firus>	1	1	0
<virws>	1	1	0

{wank}

	total	personal	official
<wanc>	28	27	1
<wank>	17	17	0

{wanker}

	total	personal	official
<wanker>	28	28	0
<wankar>	10	10	0
<wancyr>	5	5	0

{wardrobe}

	total	personal	official
<wardrobe>	44	40	4
<wardrob>	37	24	13
<wardrôb>	5	4	1
<wordrob>	5	5	0
<wordrobe>	1	1	0
<wordrôb>	1	1	0

{watch}

	total	personal	official
<watch>	1318	1070	248
<gwatch>	428	423	5
<gwach>	247	247	0

<wach>	153	153	0
<wats>	101	96	5
<gwats>	89	89	0
<wotch>	50	48	2
<watsh>	40	37	3
<wotsh>	8	6	2
<gwitsh>	2	2	0
<gwits>	1	1	0
<gwêts>	1	1	0
<waets>	1	1	0

{whatever}

	total	personal	official
<whatever>	30	29	1
<worefyr>	3	3	0
<worefy>	2	2	0
<wotefar>	2	2	0
<watever>	1	1	0
<wotever>	1	1	0
<wotevva>	1	1	0

{whatevs}

	total	personal	official
<whatevs>	2	2	0
<warevz>	1	1	0
<wotefs>	1	1	0

{whip}

	total	personal	official
<chwip>	328	190	138
<whip>	22	20	2

{youtube}

	total	personal	official
<youtube>	1246	393	853
<iwtiwb>	8	8	0
<iywtiwb>	1	1	0

E.vi Twitter: variants of interest after PCA

- {actually} – <acshiwali>
- {actually} – <acshiwli>
- {actually} – <acshli>
- {actually} – <acshwali>
- {actually} – <acshwli>
- {actually} – <acsiwali>
- {actually} – <actiwali>
- {actually} – <actiwaly>
- {actually} – <actiwli>
- {actually} – <actiwly>
- {actually} – <actsiwli>
- {actually} – <actwali>
- {actually} – <actwili>
- {algorithm} – <algorhythm>
- {algorithm} – <algorythm>
- {alright} – <allright>
- {alright} – <alreit>
- {alright} – <awright>
- {alright} – <olrait>
- {alright} – <olreit>
- {alright} – <olréit>
- {alright} – <orait>

- {alright} – <oreit>
- {amazeballs} – <amazeballz>
- {amazeballs} – <amêsbols>
- {amazeballs} – <ymezbôls>
- {anyway} – <aniwe>
- {anyway} – <anywe>
- {anyway} – <anywei>
- {anyway} – <eniwe>
- {anyway} – <eniwei>
- {anyway} – <eniwé>
- {anyway} – <eniwê>
- {apparently} – <aparentli>
- {apparently} – <aparentli>
- {apparently} – <aparently>
- {apparently} – <aparentyli>
- {apparently} – <aparyntli>
- {apparently} – <aparyntli>
- {apparently} – <apparantli>
- {apparently} – <apparantly>
- {apparently} – <apparentli>
- {apparently} – <apparyntli>
- {apparently} – <apperntli>
- {archive} – <archif>
- {awesome} – <awsome>

- {awesome} – <awswm>
- {awesome} – <awsym>
- {awesome} – <orswm>
- {awesome} – <orsym>
- {awesome} – <ossum>
- {awesome} – <oswm>
- {awesome} – <osym>
- {awesome} – <ôsum>
- {awesome} – <ôsym>
- {awful} – <awfful>
- {awful} – <awfyl>
- {awful} – <offwl>
- {awful} – <offyl>
- {awful} – <owfyl>
- {awful} – <ôffyl>
- {awkward} – <ocwyrđ>
- {awkward} – <owcwyrđ>
- {awkward} – <ôcwyrđ>
- {barricade} – <baricêđ>
- {basically} – <basicali>
- {basically} – <basiclly>
- {basically} – <basicyli>
- {basically} – <beisicli>
- {basically} – <besicli>

- {basically} – <bêsicli>
- {beautiful} – <biwriffwl>
- {beautiful} – <biwriffyl>
- {beautiful} – <biwtiffwl>
- {beautiful} – <biwtiffyl>
- {beautiful} – <biwtiful>
- {beautiful} – <biwtifwl>
- {beautiful} – <bootiful>
- {behave} – <bahaf>
- {behave} – <bihaf>
- {behave} – <bihef>
- {behave} – <buhaf>
- {behave} – <byhaf>
- {bloody} – <bludi>
- {bloody} – <blwdi>
- {bloody} – <blydi>
- {blooming} – <bloomin>
- {blooming} – <blwmin>
- {blooming} – <blwming>
- {blooming} – <blymin>
- {blooming} – <blwmin>
- {blurb} – <blyrb>
- {bollocks} – <bollocs>
- {bollocks} – <bollucks>

- {bollocks} – <bolocs>
- {bollocks} – <bolycs>
- {bottle} – <botal>
- {bottle} – <botel>
- {bottle} – <potel>
- {brexit} – <brecsit>
- {brexit} – <brecsut>
- {brexit} – <brecsùt>
- {brexit} – <bregsit>
- {bugger} – <bwger>
- {bugger} – <bygggar>
- {bungalow} – <byngalo>
- {bungalow} – <byngalow>
- {bungalow} – <bynglo>
- {camera} – <camra>
- {cancel} – <cansl>
- {cardboard} – <carboard>
- {cardboard} – <carbord>
- {cardboard} – <cardbod>
- {cardboard} – <cardbord>
- {cardboard} – <cardbwrdd>
- {cardboard} – <cardfwrdd>
- {cardboard} – <carfwrdd>
- {catalogue} – <catalog>

- {champion} – <jampion>
- {champion} – <siampion>
- {champion} – <tshiampion>
- {champion} – <tsiampion>
- {chandelier} – <chanderlier>
- {chandelier} – <siandelïer>
- {channel} – <sianel>
- {channel} – <siannel>
- {commission} – <comishwn>
- {commission} – <comisiwn>
- {commission} – <comisiyn>
- {commission} – <comiswn>
- {commission} – <comisyn>
- {commission} – <commision>
- {computer} – <compiwtar>
- {computer} – <compiwter>
- {conservative} – <consyrfatif>
- {controversy} – <controfersi>
- {controversy} – <controfyrsi>
- {cupboard} – <cwbwrt>
- {decorate} – <decyret>
- {demographic} – <demograffic>
- {demographic} – <demograffig>
- {describe} – <discraib>

- {describe} – <disgreif>
- {draft} – <drafft>
- {dyslexia} – <dyslecsia>
- {electric} – <electrig>
- {electric} – <lectrig>
- {electric} – <letrig>
- {embarrass} – <embaras>
- {eurovision} – <eurofision>
- {eurovision} – <ewrofishion>
- {eurovision} – <ewrofishon>
- {eurovision} – <ewrofision>
- {eurovision} – <ewrofsiyn>
- {eurovision} – <ewrovision>
- {eurovision} – <iwrofishyn>
- {eurovision} – <iwrofision>
- {eurovision} – <iwrofsiwn>
- {eurovision} – <iwrofsiyn>
- {example} – <ecsiampl>
- {example} – <egsampyl>
- {example} – <eisampl>
- {example} – <eisiampl>
- {example} – <esiampyl>
- {exclusive} – <ecscliwsif>
- {exclusive} – <ecsclwsif>

- {exclusive} – <ecsgliwsif>
- {exclusive} – <ecsglwisif>
- {exclusive} – <ecsglwsif>
- {exclusive} – <ecslisif>
- {exclusive} – <egscliwsif>
- {exclusive} – <egsgliwsif>
- {exclusive} – <egsglwsif>
- {exclusive} – <egsliwsif>
- {exploit} – <ecsbloet>
- {exploit} – <ecsploit>
- {fabulous} – <ffabiwlws>
- {fabulous} – <ffabiwllys>
- {fabulous} – <ffabwlyss>
- {facebook} – <ffeisbwc>
- {facebook} – <ffesbwc>
- {facebook} – <ffesbŵc>
- {facebook} – <ffésbwc>
- {facebook} – <ffêsbwc>
- {facebook} – <fêsbwc>
- {fantastic} – <fantastig>
- {fantastic} – <ffantasdeg>
- {fantastic} – <ffantasdic>
- {fantastic} – <ffantasteg>
- {fantastic} – <ffantastic>

- {fantastic} – <ffantastig>
- {film} – <ffilm>
- {film} – <ffulm>
- {focus} – <ffocus>
- {focus} – <ffocws>
- {focus} – <ffocys>
- {focus} – <focws>
- {focus} – <focys>
- {format} – <fformat>
- {fraction} – <ffracsiwn>
- {freelance} – <ffrilans>
- {freelance} – <ffrîlans>
- {fridge} – <ffridj>
- {fridge} – <ffrij>
- {fridge} – <fridj>
- {fridge} – <frij>
- {friend} – <ffrend>
- {friend} – <ffrind>
- {friend} – <ffrynd>
- {friend} – <frind>
- {fuck} – <ffuc>
- {fuck} – <ffuck>
- {fuck} – <ffwc>
- {fuck} – <ffwck>

- {fuck} – <ffyc>
- {fuck} – <ffŵc>
- {gadget} – <gadjet>
- {gadget} – <gajet>
- {genuinely} – <geniwinli>
- {genuinely} – <genuenly>
- {genuinely} – <genuienly>
- {genuinely} – <jeniwainli>
- {genuinely} – <jeniwini>
- {google} – <googel>
- {google} – <gwgäl>
- {google} – <gwgl>
- {google} – <gwgył>
- {google} – <gŵgl>
- {google} – <gŵgył>
- {homophobia} – <homoffobia>
- {honest} – <gonast>
- {honest} – <gonest>
- {honest} – <honast>
- {honest} – <onast>
- {honest} – <onesd>
- {honest} – <onest>
- {hungry} – <hyngri>
- {icon} – <eicon>

- {inbox} – <inbocs>
- {insulin} – <inswlin>
- {insurance} – <inshiwrans>
- {insurance} – <inshwrans>
- {insurance} – <insiwrans>
- {insurance} – <siwrans>
- {jealous} – <jelas>
- {jealous} – <jelos>
- {jealous} – <jelws>
- {jealous} – <jelys>
- {kaleidoscope} – <caleidosgop>
- {level} – <lefal>
- {level} – <lefeal>
- {level} – <lefel>
- {level} – <lefl>
- {literally} – <literali>
- {literally} – <literaly>
- {literally} – <literraly>
- {literally} – <literyli>
- {literally} – <litrali>
- {literally} – <litrely>
- {literally} – <litrli>
- {literally} – <litryli>
- {literally} – <litteraly>

- {literally} – <lityrali>
- {lovely} – <llyfli>
- {lovely} – <lovelly>
- {lovely} – <lovley>
- {lovely} – <lovli>
- {lovely} – <lufli>
- {lovely} – <lwfle>
- {lovely} – <lwfflee>
- {lovely} – <lwfli>
- {lovely} – <lyfli>
- {lovely} – <lyflu>
- {lovely} – <lyfly>
- {manifesto} – <maniffesto>
- {massive} – <masif>
- {massive} – <masive>
- {massive} – <massiv>
- {massive} – <masuf>
- {melody} – <melodi>
- {mention} – <menshyn>
- {metamorphosis} – <metamorffosis>
- {molecule} – <moleciwl>
- {nervous} – <nerfus>
- {nervous} – <nerfys>
- {nervous} – <nurfys>

- {nervous} – <nyrfys>
- {nice} – <nais>
- {nice} – <neis>
- {nice} – <neus>
- {nonsense} – <nonsans>
- {nonsense} – <nonsens>
- {nuclear} – <niwclear>
- {nuclear} – <niwcliar>
- {obscure} – <obsciar>
- {okay} – <oce>
- {okay} – <ocê>
- {okay} – <ok>
- {okay} – <okei>
- {okay} – <owcai>
- {okay} – <owcei>
- {okay} – <owkei>
- {oops} – <whoops>
- {oops} – <woops>
- {oops} – <wps>
- {oops} – <wups>
- {oops} – <ŵps>
- {operation} – <operasiwn>
- {operation} – <opereshyn>
- {opinion} – <opiniwn>

- {**opinion**} – <pinewn>
- {**option**} – <opshewn>
- {**option**} – <opsiwn>
- {**option**} – <opsiyn>
- {**option**} – <opswen>
- {**option**} – <optsiwn>
- {**pakistan**} – <pacistan>
- {**parameter**} – <paramater>
- {**parameter**} – <paramedr>
- {**pathetic**} – <pathetig>
- {**perspective**} – <persbectif>
- {**perspective**} – <perspectif>
- {**petition**} – <petishiwn>
- {**petition**} – <petisiwn>
- {**photobomb**} – <ffotobom>
- {**photobomb**} – <ffotobomb>
- {**photobomb**} – <photobom>
- {**photocopy**} – <ffotocopi>
- {**photoshop**} – <ffotoshop>
- {**photoshop**} – <ffotosiop>
- {**picnic**} – <picnig>
- {**please**} – <plis>
- {**please**} – <plys>
- {**please**} – <plís>

- {please} – <plis>
- {please} – <plis>
- {please} – <pwease>
- {please} – <pwis>
- {potential} – <potenshal>
- {potential} – <potenshial>
- {potential} – <potensial>
- {power} – <pawyr>
- {power} – <powyr>
- {power} – <pwar>
- {power} – <pwer>
- {power} – <pwêr>
- {power} – <pŵer>
- {problem} – <problam>
- {process} – <proses>
- {project} – <prosciect>
- {project} – <prosiecd>
- {project} – <prosiect>
- {psychopath} – <seicopath>
- {purpose} – <pwrpais>
- {purpose} – <pwrpas>
- {purpose} – <pwrpws>
- {purpose} – <pŵrpas>
- {queer} – <cwiar>

- {queer} – <cwyr>
- {reality} – <reailiti>
- {reality} – <realiti>
- {reason} – <rheswm>
- {reason} – <rheswn>
- {reason} – <rhesym>
- {reason} – <rhyswm>
- {ridiculous} – <ridicilous>
- {ridiculous} – <ridicilys>
- {ridiculous} – <ridiciwlys>
- {ridiculous} – <ridiclws>
- {ridiculous} – <ridicylys>
- {right} – <reit>
- {right} – <rheit>
- {round} – <rownd>
- {russia} – <rusia>
- {russia} – <rwsia>
- {sausage} – <sosej>
- {sausage} – <sosij>
- {sausage} – <sosêj>
- {scooter} – <sgwter>
- {screen} – <scrin>
- {screen} – <scrîn>
- {screen} – <sgrin>

- {screen} – <sgrîn>
- {screen} – <sgrîn>
- {scribble} – <sgribl>
- {scroll} – <sgrol>
- {scroll} – <sgrôl>
- {security} – <seciwriti>
- {security} – <siciwriti>
- {selfie} – <selffi>
- {selfie} – <selffie>
- {selfie} – <selffi>
- {selfie} – <selfi>
- {session} – <seshan>
- {session} – <seshawn>
- {session} – <seshion>
- {session} – <seshiwn>
- {session} – <seshwn>
- {session} – <seshw̃n>
- {session} – <sesiwn>
- {session} – <sesiyn>
- {session} – <sessiwn>
- {shower} – <shawyr>
- {shrug} – <shryg>
- {smoothie} – <smwddi>
- {smoothie} – <smwddî>

- {smoothie} – <smôddi>
- {sorry} – <sori>
- {sorry} – <sorri>
- {spooky} – <sbwci>
- {spooky} – <sbôci>
- {spooky} – <spooki>
- {spooky} – <spwci>
- {spooky} – <spôci>
- {sprint} – <sbrint>
- {square} – <sgwar>
- {square} – <sgwâr>
- {starve} – <starf>
- {stereotype} – <stereoteip>
- {stereotype} – <sterieoteip>
- {stereotype} – <sterioteip>
- {stereotype} – <stéréotype>
- {stop} – <sdop>
- {stupid} – <sdiwpid>
- {stupid} – <sdiwpud>
- {stupid} – <sdiwpyd>
- {stupid} – <shtoopid>
- {stupid} – <stiwpid>
- {stupid} – <stiwpuđ>
- {stupid} – <stiwpyđ>

- {stupid} – <stoopid>
- {stupid} – <stwpid>
- {superduper} – <swperdwper>
- {superduper} – <swpyrdwpyr>
- {surely} – <shiwarli>
- {surely} – <shwrli>
- {surely} – <shŵyrli>
- {surely} – <surley>
- {surf} – <syrff>
- {syndrome} – <sundrome>
- {syndrome} – <syndrom>
- {syndrome} – <syndrôm>
- {text} – <tecsd>
- {text} – <tecst>
- {thank you} – <ffanciw>
- {thank you} – <thancew>
- {thank you} – <thanciw>
- {thank you} – <thancyw>
- {thank you} – <thenciw>
- {tweet} – <twit>
- {tweet} – <twît>
- {tweet} – <twït>
- {tweet} – <twîit>
- {tweet} – <twîit>

- {ukip} – <iwcip>
- {ukraine} – <iwcraen>
- {ukraine} – <iwcraín>
- {ukraine} – <wcraen>
- {ukraine} – <wcraín>
- {ukraine} – <wcraín>
- {ukraine} – <îwcraín>
- {ukraine} – <ŵwcraín>
- {ukulele} – <iwcaleili>
- {ukulele} – <iwcalele>
- {ukulele} – <iwcaleli>
- {ukulele} – <iwcalili>
- {ukulele} – <iwceleli>
- {ukulele} – <ukelele>
- {uniform} – <iwnifform>
- {uniform} – <iwniffôr^hm>
- {upload} – <yplod>
- {video} – <fideo>
- {video} – <fdio>
- {video} – <vidio>
- {virus} – <feirws>
- {virus} – <feirys>
- {virus} – <firus>

- {virus} – <firws>
- {virus} – <veirws>
- {virus} – <virws>
- {wanker} – <wancyr>
- {wanker} – <wankar>
- {watch} – <gwach>
- {watch} – <gwatch>
- {watch} – <gwats>
- {watch} – <gwits>
- {watch} – <gwitsh>
- {watch} – <gwêts>
- {watch} – <wach>
- {watch} – <waets>
- {watch} – <wats>
- {watch} – <watsh>
- {watch} – <wotch>
- {watch} – <wotsh>
- {whatever} – <watever>
- {whatever} – <worefy>
- {whatever} – <worefyr>
- {whatever} – <wotefar>
- {whatever} – <wotever>
- {whatever} – <wotevva>
- {whatevs} – <warevz>

- {**whatevs**} – <wotefs>
- {**whip**} – <chwip>
- {**youtube**} – <iwtiwb>
- {**youtube**} – <iywtiwb>