Applied Corpus Linguistics 3 (2023) 100052

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Corpus to curriculum: Developing word lists for adult learners of Welsh

Dawn Knight[a,*], Tess Fitzpatrick[b], Steve Morris[a], Bethan Tovey-Walsh[b], Helen Prosser[c], Emyr Davies[d]

[a] Cardiff University, Cardiff, Wales, United Kingdom
[b] Swansea University, Swansea, Wales, United Kingdom
[c] National Centre for Learning Wales, United Kingdom
[d] Welsh Joint Education Committee, Wales, United Kingdom

A B S T R A C T

The launch of a language's first comprehensive general corpus promises a sea-change in teaching and learning resources. Effective transition from corpus to classroom is not necessarily straightforward, though; expert and end-user input is essential for the potential of the corpus resource to be realised. This paper outlines the process by which fit-for-purpose vocabulary lists were derived from the new National Corpus of Contemporary Welsh (*Corpws Cenedlaethol Cymraeg Cyfoes* – CorCenCC). The immediate purpose in this case was to inform the revision of A1 and A2 level course materials for adult learners. A longer-term aim was to put in place a method by which vocabulary lists for more advanced level learners and learners of different ages could be extracted and developed from the corpus. The new corpus means that for the first time, the Welsh language curriculum is able to use word frequency information; teaching and assessment materials in major languages have been informed by word frequencies for several decades. Raw frequency lists, though, include troublesome content, and can exclude items with high relevance to learners. This paper demonstrates how, by working in partnership, Welsh language curriculum writers, assessors, language experts and corpus linguists can effectively manipulate corpus data into curriculum content. The methods and approaches reported here are replicable for use in other language contexts.

1. Introduction

The compilation of a general language corpus is indisputably necessary for the construction of frequency-based pedagogical word lists. However, if those lists are to be fit for purpose then corpus creation is merely the first step of an iterative, collaborative and sometimes complex process. Collaborative, complex processes are typically challenging to replicate, or to report in a replicable frame, and in this paper we take on this challenge, charting a user-driven approach to building pedagogical word lists for Welsh, that capitalises on expert-led scrutiny and reflection, along with dynamic, full-team engagement with a new corpus resource. The approach is markedly fitting for minoritised or non-major languages with distinct community and policy infrastructure: it uses a resource which is openly accessible, community-informed, and free to use; it connects closely with language policy direction; it draws on the expertise and motivation of the principal national Welsh language learning and language assessment providers, and it seeks to address the needs they identify. This contrasts sharply with parallel endeavours in major and well-resourced languages such as English, where corpora are more plentiful, but often designed for expert access, and where teaching and assessment provision is spread across a variety of providers, with diverse curricular, commercial and/or policy objectives.

The arrival on the scene of a new resource – in this case the National Corpus of Contemporary Welsh (*Corpws Cenedlaethol Cymraeg Cyfoes* – CorCenCC, Knight et al., 2020a) – can be seen as heralding a new phase of materials creation and curriculum design. However, it is important to recognise context: in reality the many languages without corpus resource are likely already to be working with sophisticated, well-informed curricula that are subject to a programme of scheduled critical reviews and revision. So, the real-world challenge when faced with a significant new resource, after validity and relevance evaluations, is to integrate and embed it within an existing curriculum while simultaneously i) minimising disruption to current cohorts of learners and teachers, including with regard to assessment; ii) operating within the resource available for material and curriculum design; iii) retaining clarity about the principles and methods used in order to facilitate iterative review and revision in the future. In order to meet this challenge, the project reported here is shaped around three key objectives:

a) to work iteratively in collaboration with curriculum and assessment strategists and designers towards the most effective use of corpus data in pedagogical word lists,
b) to converge with ongoing curriculum development by prioritising word list development for the current stage of the development cycle (in this case the national resource materials for A2 adult learners of Welsh), and
c) to review and revise elements of the corpus infrastructure in order to expedite the efficient and effective development of word lists for other levels of learning (i.e. other than A2), and for other learners (i.e. young learners) – in other words to future-proof the methodology.

* Corresponding author.
E-mail address: knightd5@cardiff.ac.uk (D. Knight).

https://doi.org/10.1016/j.acorp.2023.100052
Received 25 January 2023; Accepted 10 March 2023
Available online 16 March 2023

The approach presented here can be replicated and adapted for other language communities; for easy identification of parallels and differences between those contexts and the Welsh language situation, we open by outlining key features of Welsh language status and language education in Wales. The project is characterised by collaboration, and we go on to examine the roles and agendas of project partners – the National Centre for Learning Welsh, the WJEC/CBAC (Wales' largest qualification awarding body) - and the CorCenCC team of corpus linguists and applied linguists. There is a rich history of pedagogical word list development in major languages, and methods and approaches are critically reviewed and evaluated in light of the Welsh language context and stakeholders, and the background to the CorCenCC corpus. The paper then reports the methods used to extract an initial set of frequency lists (*Yr Amliadur*[1]) from the CorCenCC corpus, and the subsequent sets of consultation and refinement that eventually generated i) a fit-for-purpose pedagogical word list to inform the A2 curriculum (the *Geirfan*[2] word list), and ii) revisions to the corpus tagger that will facilitate creation of word lists for different levels or learner profiles in the future. Word lists based solely on statistical data (frequency, range, similarity across sub-corpora) are relatively straightforward to replicate, but lack pedagogical nuance; our aim in this paper is to report the creation of a word list informed by both corpus and pedagogical expertise, in such a way that it too can be replicated.

2. Context

2.1. Language status and language education in Wales

Welsh has held official status as a language in Wales since the passing of the Welsh Language (Wales) Measure in 2011. The 2021 census recorded 538,300 speakers of Welsh (17.8 per cent of the total population of Wales in that census – 3018,171 – StatsWales, 2022), but annual population surveys consistently indicate a higher number of speakers, with the December 2021 survey suggesting a figure of 892,000 (ONS, 2021). There is a significant diaspora of Welsh speakers (and learners) beyond Wales, and an established Welsh speaking community in Patagonia, Argentina. Welsh speakers in Wales are not distributed evenly across the country; percentages of speakers vary from over 50 per cent in the north-west to just under 10 per cent in parts of the south-east. Nevertheless, all public – as well as some private – bodies in Wales offer bilingual services and the demand for a bilingual workforce to provide these services is growing. Welsh language education is available to all children in Wales, with 110,142 (23%) attending Welsh medium schools (Welsh Government figures for April 2021 (2021a)). Welsh is a compulsory school subject for all 3–16 year olds. In addition, the Welsh language is, for many, closely linked with Welsh identity and these two instrumental and integrative drivers form the main motivations for the adults who enrol on the National Centre for Learning Welsh courses every year (Baker et al., 2011). In 2018 the Welsh Government announced their target to increase the number of Welsh speakers to 1 million by 2050, and initiatives associated with this, along with current cultural and socio-political swings have seen a steady increase in learner numbers (further detail below).

Education is a devolved power in Wales, governed in terms of legislature by the Senedd (Welsh Parliament) and executively by Welsh Government. A new schools Curriculum for Wales (Hwb, 2021a) was launched in 2021, with Welsh language education embedded within "Languages, Literacy and Communication" – one of the six "Areas of Learning and Experience" that constitute the curriculum. The roll-out of the new curriculum has been scheduled to take place in stages between 2021 and 2027 (Hwb, 2021b), mapping onto a cyclical process of curriculum revision and improvement. Easy access to fit-for-purpose resources, including facilities for creating and developing pedagogical word lists, will assist educators in this challenging process of reform. Objective (c) of our project, noted above, relates to this. The second of our objectives (b) listed above, though, relates to a different group of learners: adult learners of Welsh; 11% of Welsh speakers are reported to have acquired the language as adults (Welsh Government, 2021b). While many adult learners use commercial apps such as Duolingo or the innovative SaySomethingInWelsh (https://en.saysomethingin.com/) to progress or supplement their learning, the provision of live Welsh classes for adults, and associated resources, is almost entirely the domain of the National Centre for Learning Welsh, one of the two principal industry partners in this project.

Table 1
Welsh for adults list of courses.

Qualification	English translation	CEFR level
Mynediad	Entry	A1
Sylfaen	Foundation	A2
Canolradd	Intermediate	B1
Uwch	Advanced	B2

2.2. Key players in the project

The National Centre for Learning Welsh (NCLW) is key to the project reported in this paper. It is funded by Welsh Government and is responsible for all aspects of the provision for adults to learn the language, including the development of "a high quality, appropriate and engaging national curriculum" and the production of a "wide range of resources suitable for a range of learners" (NCLW, 2021). The Centre and its learning providers are subject to inspection by Estyn, the education and training inspectorate for Wales, and Welsh Government commissions regular reviews of the Centre's provision. The 'Learn Welsh' courses provided by the NCLW via regional providers, are offered at four levels (Table 1) and are typically taken as weekly or twice-weekly classes across an academic year, though intensive courses are also available, along with a suite of 'Work Welsh' courses. The NCLW provision also includes specific resources for parents and families, targeted especially at those with children in Welsh-medium or bilingual schools, and in this way intersects with the school provision described above. Since 2020 the majority of classes have shifted to online or blended provision, and have started to attract learners from outside Wales as well as those resident in Wales. 86% of learners are of working age (2020–21 figures, Learn Welsh, 2022). The number of individual learners registered on NCLW courses per year rose from 12,680 in 2017–18, to 17,505 in 2019–20.

WJEC–CBAC (Welsh Joint Education Committee), the other industry partner on this project, is Wales' main qualification provider for GCSEs, A Levels and vocational awards, and is a full member of ALTE (Association of Language Testers in Europe). Qualifications are offered bilingually, in line with the education system in Wales. Importantly for this project, the WJEC runs the suite of Welsh for Adults qualifications that sit alongside the Learn Welsh courses offered by the NCLW, described above. Qualifications are offered at Entry, Foundation, Intermediate and Advanced levels (WJEC, 2020), corresponding respectively to levels A1, A2, B1 and B2 of the Common European Framework of Reference (Council of Europe, 2001). The WJEC Welsh for Adults team had previously collaborated on and commissioned work on pedagogical word lists resulting in the *Geirfa Graidd* lists at A1 and A2/B2, before a corpus was available (Morris, 2011), and were also stakeholder representatives on the CorCenCC project team.

The release of CorCenCC v.1.0.0, the first comprehensive corpus of contemporary Welsh, offered an opportunity for the NCLW and WJEC

[1] Modelled on other language related coinage in Welsh where the affix -adur is added to the root word e.g. gair [word] + -adur = geiriadur [dictionary] thus aml [often, frequent] + -adur = amliadur [frequency list]. Yr = the definite article. *Yr Amliadur* is a name coined by the team.

[2] *Geirfan* is a name coined by the team, combining gair [word] and man [site or place].

project partners to revise, using frequency information, the pedagogical word lists that underpin their course materials and assessment instruments. In the first instance, slotting into the established cycle of materials revision, the priority was to inform the creation by NCLW of a new edition of the coursebook for Foundation (A2) level, and a revision of the course material at Entry level (A1). Thereafter, lists are to be developed for Intermediate and Advanced level (B1 and B2) materials. By working alongside each other to inform the process of word list creation, WJEC and NCLW can ensure that instruction and assessment instruments are developed in a coordinated and consistent way. WJEC also deliver Welsh language qualifications for schools; a further benefit of their ongoing involvement in list creation is the potential for future knowledge transfer to school contexts.

The CorCenCC corpus, like other major corpora, is tagged for part-of-speech (POS) and also, more unusually, by semantic field; the taggers were built as part of the CorCenCC project, and the POS tagger in particular is instrumental to the accuracy of frequency data extracted from the corpus. The construction of taggers is rarely, perhaps never, perfectly fitted to the multiple possible uses of a corpus. To maximise the suitability of the corpus for future applications, the CorCenCC project had incorporated knowledge and information from multiple stakeholders into the construction of the corpus and its infrastructure. The project reported here took this to the next logical step, by offering members of the corpus team an opportunity to scrutinise the fitness for purpose of elements of the corpus infrastructure – and in this case the frequency lists it generated – in an operational setting, and to begin the inevitable process of revision and adjustment to improve the usability of the resource. Below we detail the steps taken as part of this revision and adjustment, but first we consider some of the decisions and deliberations entailed in creating pedagogical word lists.

3. Corpus-informed word lists

3.1. Availability, utility and general principles

Over recent decades the use of corpus-informed word lists has become commonplace in the teaching and assessment of English and, to a lesser extent, of other major and/or corpus-resourced languages (evidenced by the fast-growing Routledge series of Frequency Dictionaries, at the time of writing available for Portuguese, Czech, French, Japanese, Dutch, Russian, Arabic, Mandarin Chinese, Turkish, Korean, Persian, Spanish, German - see https://www.routledge.com/Routledge-Frequency-Dictionaries/book-series/RFD). While frequency lists that are derived from major English language corpora such as the British National Corpus (BNC) World Lists (Leech et al., 2001), the BNC database and frequency lists (Kilgariff, 1998), the Corpus of Contemporary American frequency list (COCA, Davies, 2008), and the new-General Service list (Brezina and Gablasova, 2015) differ slightly in terms of the ranking of items, there is general agreement regarding which words occur most often. In pedagogical word lists, these 'high frequency' items are typically (though not always) represented as word families. Learners are encouraged to prioritise the learning of these items, so as to maximise the proportion of words they will know in any given text, i.e. the lexical coverage, as early as possible in their learning. 'High frequency' items variously number between 1000 (Engels, 1968; Dang and Webb, 2016) and 3000 (Schmitt and Schmitt, 2014), but typically, for English, the first 2000 words have been labelled 'high frequency' (see Nation, 2016). 'Low frequency' items are suggested to be those less frequent than the 9000 most frequent word families (Schmitt and Schmitt, 2014). Of course some members of a word family will be much more frequent than others; for English, systematic teaching of common affixes has supported learners in extrapolating from acquisition of one item to knowledge of other word family members (Bauer and Nation's 1993 graded taxonomy of affixes is a notable resource in this area).

Word frequency is generally regarded as the most important factor in creating pedagogical word lists, but it is not the only factor to consider; see for example Nation's "subjective criteria" (2016: 10) and Ishikawa's "pedagogical adjustments" (2019: 2). Consideration of other criteria for inclusion in pedagogical word lists entails a shift away from the exclusively quantitative domain of ranked frequencies of corpus items, and necessitates more qualitative, context-specific decision-making. Other criteria might include:

- the "learnability" or "learning burden" of a vocabulary item – for example, the transparency of its orthography, the typicality of its grammatical patterning, etc. (West, 1953; Nation, 2001),
- relevance to specialised, or syllabus-defined topics, modality or register,
- necessity relating to context/environment (e.g. classroom language),
- inclusion or exclusion of proper nouns, numbers, etc., and
- consideration of the L1 of the learners (e.g. the JACET lists were constructed specifically for Japanese L1 learners of English (JACET, 2016; Ishikawa, 2019)).

These criteria are not necessarily independent of frequency or of each other: for example, frequency interacts with contextual constraints, dispersion across registers and domains, collocation pairings and, in English at least, word length (Schmitt, 2010: 64). The application of criteria beyond corpus frequency is most usefully supplied by expert practitioners, and as such are captured by the concept of 'indigenous criteria', which is considered later in this paper.

As noted above, the items in pedagogical word lists often represent word families, so that one entry encompasses all inflectional and derivational forms. Other approaches use lemmas (head word and inflectional forms within same part-of-speech) (Dang and Webb, 2016) or even types – the 'raw' word forms found in the corpus (Zeno et al., 1995). Deciding whether to use word families, lemmas, flemmas (head word and inflectional forms across different parts of speech) or types in pedagogical word lists entails assumptions about learners' capacity to apply morphological – specifically inflectional and derivational – knowledge, and it is important to note that the ways in which this operates will differ from language to language, and will depend on learner level. Decisions are also necessary about the pedagogical expedience of listing single word items only, or the inclusion of multi-word units.

From the points noted above it is evident that the conversion of raw frequency rankings of word forms in a corpus into fit-for-purpose pedagogical word lists, entails robust and informed decision-making on multiple factors. Furthermore, the automaticity with which appropriate frequency data can be extracted from a corpus depends on how reliable the corpus annotation is and on what rule system the corpus taggers are based. Below we address the operationalisation of this within our project, but first a word about the principles underlying the CorCenCC corpus itself.

3.2. The CorCenCC corpus

The quality and nature of a frequency-based word list are fundamentally dependant on the corpus from which it derives. CorCenCC (*Corpws Cenedlaethol Cymraeg Cyfoes* - the National Corpus of Contemporary Welsh) was launched in November 2020 (see Knight et al., 2020a and www.corcencc.org). CorCenCC was constructed using a principled sampling frame and is the first large-scale corpus of Welsh designed to capture language use across communication types (spoken, written and e-language), genres, language varieties (regional and social) and contexts, with contributors representative of the 538,300 Welsh speakers in the UK. The CorCenCC v1.0.0 dataset extends to over 11.16 million words of contemporary Welsh language usage (across 11,432 texts). As seen in Table 2, this includes data from a range of contexts, genres and modes of communication, from spoken broadcast texts to written magazine articles to SMS messages.

The CorCenCC project team included corpus linguists, applied linguists, computational linguists, software engineers, Welsh language experts, and an advisory group of stakeholder representatives compris-

Table 2
The size and composition of CorCenCC v1.0.0 (from Knight et al., 2020a: 61).

Mode		No. of texts	No. of words	Total
Spoken	broadcast	564	750,078	
	professional	80	477,983	1331 texts
	educational	136	296,709	2864,974 words
	transactional	191	204,758	
	public or institutional	137	433,361	
	social	131	456,487	
	private	92	245,598	
Written	academic_journal	9	272,831	
	book	137	1928,582	704 texts
	essays_coursework_and_exams	31	26,047	3895,115 words
	leaflet_document_announcement	339	792,679	
	letter	53	12,873	
	magazine	80	329,203	
	miscellaneous	5	8251	
	newsletter	33	78,803	
	papurau_bro	13	117,334	
	thesis	4	328,512	
E-language	blog	48	2345,909	
	email	781	141,554	9397 texts
	SMS (inc. instant messages)	8487	93,541	4402,003 words
	website	81	1820,999	
		11,432	**11,162,092**	

ing Welsh Government (including the Translation and Reporting Service), National Assembly for Wales (now the Senedd), BBC, S4C, Gwasg y Lolfa publishers, SaySomethinginWelsh language learning software, University of Wales Dictionary of the Welsh Language, and one of the partners on the current word lists project, WJEC. The diverse interests of the advisory group members was intended to ensure that the corpus construction accommodated, as far as possible, future applications of the resource. A part-of-speech (POS) tagset for Welsh, *CyTag* (Neale et al., 2018), was developed, informed by the Bangor Autoglosser (Donnelly and Deuchar, 2011), but adapted and refined for application to spoken and e-language texts. A semantic tagger, *CySemTag*, was also developed (Piao et al., 2018). This entailed the adaptation and extension of the existing UCREL Semantic Analysis System (USAS – Rayson et al., 2004) tagset to accommodate the special characteristics of Welsh (bringing the number of languages covered by USAS to 12).

In addition to the CorCenCC v1.0.0 dataset and taggers, a key output from the CorCenCC project was the production of *Yr Amliadur* (Knight et al., 2020b), which included the initial set of corpus-derived word lists, and a starting point for the lists developed in the current project.

3.3. *Word list version 1:* Yr Amliadur

Yr Amliadur (Knight et al., 2020b, and available at: www.corcencc.org/download) is a set of word forms and lemmas from the CorCenCC corpus, ranked by frequency, and presented in a series of lists extracted from the whole corpus, from each of the sub-corpora, for mode of communication (written, spoken, e-language), and by part-of-speech.

To create *Yr Amliadur*, some pervasive tagging errors and ambiguous part-of-speech tags in the corpus data were manually corrected. In addition, the data file was refined in order to remove a few unwanted passages (e.g. of JavaScript code in the e-language samples) which had not already been deleted during the manual or automatic processing of the data. *Yr Amliadur* therefore differs slightly from any similar lists generated directly from the first public release of the CorCenCC v1.0.0 dataset. Specifically, *Yr Amliadur* frequency lists provide:

- the top 1000 lemmas in CorCenCC, sorted both by rank and alphabetically by lemma,
- the top 1000 word forms in CorCenCC, sorted both by rank and alphabetically by word,
- banded word lists of 5000 of the most common open-class words in CorCenCC, split into lists of bands containing 500 words each (including lists for the top 500 nouns, verbs, adjectives),
- the top 50 adverbs and interjections in CorCenCC, and,
- the top 100 open-class words in each of CorCenCC's written, spoken and e-language sub-corpora.

In the production of *Yr Amliadur*, only alphabetic tokens were counted towards the total number of tokens in a (sub)corpus, with the following items removed from the list (based on Knight et al., 2020b: 6–7):

- anonymised data: e.g. where *enwb* [fem. noun] has replaced the name of a female mentioned in the text or *cyfenw* [surname] replaced the surname of an individual,
- non-lexical features of speech: e.g. coughing, laughing or yawning that have been annotated by transcribers of the spoken content,
- non-Welsh words, predominantly those from English (which are not uncommon in the bilingual context of contemporary Welsh), and
- proper nouns: e.g. where *gwyn* is both a proper noun (a person's name) and an adjective [white].

Although removed, these items were still counted within the total number of tokens included in the (sub)corpus, with frequency counts of all items in the corpus calculated at their rate per million words, as is a common standard in the field of corpus linguistics.

The extraction of *Yr Amliadur* lists was conducted by the CorCenCC team; at this stage potential end users of the lists were not involved. This was for two reasons: i) the construction of the lists enabled an evaluation of the efficacy of the corpus and the frequency extraction tools, and guided any necessary adjustments; and ii) frequency lists can be used by a range of different end users (teachers, learners, material writers, publishers, authors, translators, broadcasters…), each of whom will use and adapt the lists in slightly different ways. Notwithstanding this, *Yr Amliadur* was immediately taken up for use by the NCLW, who were eager to ensure that the top 100 most frequent words of the Welsh language were included in their Entry level (A1) course books for beginners. *Yr Amliadur* allowed the NCLW curriculum writers, for the first time, to consult corpus based frequency lists in the preparation of teaching materials.

However, in order for the frequency lists deriving from the CorCenCC corpus to be a targeted and effective teaching, learning and assessment resource, their content and usability would need further development.

This would necessitate reconciling the purely corpus-based initial iteration of the lists with the kinds of pedagogical word list criteria noted earlier in this paper. In order to do this, expert insights were sought from our language education (NCLW) and assessment (WJEC) partners; their experience and their knowledge of the specific contexts and situations in which the lists would be used were essential to the fitness for purpose of the resource. These "insights from domain experts" with "unique understandings of the context of interest" (Elder and McNamara, 2016: 153, 154) constitute the *indigenous criteria* that lends authenticity and applicability to a resource or output (see also Pill, 2016). In order for pedagogical word lists to be developed from the 'raw', statistically derived *Yr Amliadur*, it was necessary to apply indigenous criteria from the educators and assessors who would be putting the lists to use. The remainder of this paper, then, reports on the work undertaken to critically review the information available via *Yr Amliadur*, and to make the adjustments and revisions necessary to create new pedagogically orientated word lists, using the data from CorCenCC v1.0.0, along with expert input from representatives of the NCLW, the WJEC and the CorCenCC team.

3.4. Revising the corpus infrastructure

3.4.1. Revising the CyTag tagger

The identification and annotation of large corpora according to individual word and lemma forms, for the development of word lists, depends on the availability of POS taggers. This process of identification and classification would be impossible to undertake manually, particularly when using an extensive dataset such as CorCenCC to create a word list. In major language contexts the accuracy and robustness of POS taggers have increased alongside the availability of extensive corpora. For English, The TreeTagger (Schmid, 1994), for example, reportedly tags with an accuracy of 96%, while the Stanford Log-Linear Part-of-Speech Tagger (Toutanova et al., 2003) has an accuracy of over 97%. These are both probabilistic taggers (i.e. they estimate the probability of a given token to be relevant to a specific category/tag) that require extensive hand-coded training corpora to train and optimise available resources. In under-resourced and minoritised language contexts such as Welsh, there may be a lack of pre-annotated data available (and sometimes simply a lack of data itself) to train probabilistic taggers, so the development of rule-based taggers is often used as an alternative approach. Under the rule-based approach, tags are manually (and laboriously) assigned using pre-defined rules regarding which syntactic categories can typically be co-located in a language. *CyTag*, the tagger built by the CorCenCC team, uses a rule-based approach.

The original version of *CyTag* had a high level of tagging precision, with recall scores noted at well over 95% in the initial releases of the tagger (Neale et al., 2018). As with any tagger, however, some items were difficult for the tagger to disambiguate, leading to potential mis-tagging. This limitation was noted in the first release of *Yr Amliadur*, with the caveat "this should be borne in mind when interpreting frequencies of words or lemmas which are susceptible to mis-tagging" (Knight et al., 2020b: 20). An acknowledgement of this limitation, and of the ongoing constraints it might place on extracting corpus data, motivated the team to enhance the tagger's accuracy in tagging high frequency words in the first instance (as these were of primary concern to the external project partners), as well as making *CyTag's* code easier to debug and maintain in the long term.

The revision of the POS tagger was undertaken with three main objectives:

i) to further improve the general functionality of the tagger for both immediate and long-term utilisation, and to 'future-proof' it for extended potential applications,
ii) to facilitate automatic recognition of multiword units (MWUs), which when tagged as individual words could skew frequency counts, and
iii) to facilitate disambiguation of items in cases of identical word forms (homonymy).

The process of improving and augmenting the tagger was iterative: our review team of Welsh language experts from NCLW, WJEC and the CorCenCC team manually reviewed frequency lists at multiple stages as *CyTag* was being improved, and their feedback was used to isolate words which were (still) being mis-tagged, for further attention. The steps taken to achieve these objectives, in order to build the resulting *Geirfan* word list (v2.0) for pedagogical purposes (in contrast to the general purpose *Yr Amliadur*), are reported next.

3.4.2. Re-writing the CyTag code base

First, a decision was taken to rewrite the code base of *CyTag* in order to enhance its readability and maintenance. An OOP (object-orientated) programming paradigm, using Python, was selected as having a number of advantages relevant to *CyTag* and to future applications of the corpus: it facilitates the construction of hierarchies (e.g. of subgroups within the larger class 'nouns'); code is broken into small units, making the structure more visible to subsequent maintainers of the code; it relies on modular units (classes and methods) which are infinitely reusable, thus expediting repetition of similar operations; it is relatively intuitive, navigable, and simple to manipulate. These features help to future proof this tagging functionality by making the code easier to extend and expand in due course, for longer term project goals. The revised code/version of *CyTag* is available here: https://github.com/CorCenCC/CyTag.

3.4.3. Extending the CyTag lexicon

CyTag uses a Welsh lexicon, *Eurfa* (Donnelly, 2013) to look up words from the corpus. The results are stored as possible readings for the word, and these are fed to the constraint grammar (see Section 3.4.4), which chooses between readings based on a series of rules. Because *CyTag* relies on the lexicon to identify the possibilities for tagging a word, it is important that the lexicon has very broad coverage. In reviewing *Yr Amliadur*, the project team identified that the original version of the *CyTag* lexicon (i.e. that used when tagging the CorCenCC v1.0.0 and in the production of *Yr Amliadur*) was missing words in a number of key areas, which were subsequently added to the lexicon. These areas include:

- some irregular verb paradigms,
- terminology for more recent technology and science (e.g. the verbs *e-bostio* [to email] and *dad-ffrindio* [to unfriend],
- variant spellings of words with prefixes (e.g. the verb meaning 'to cooperate', which may be written *cyd-weithredu, cydweithredu*, or *cyd weithredu*),
- numerals (both ordinal and cardinal), especially longer numerals in the traditional Welsh counting system (e.g. *dau-ar-bymtheg-ar-hugain* [thirty-seven - literally two-on-fifteen-on-twenty], *pedwerydd-ar-bymtheg* [nineteenth - literally fourth-on-fifteen]), and,
- some common colloquialisms/regionalisms (e.g. *glei* ([indeed] - in some dialects of south-western Wales), *twlu* (variant of *taflu* [to throw] in southern Wales)).

3.4.4. Improving the grammar

Third, *CyTag* uses the Constraint Grammar formalism to disambiguate words with more than one potential reading. Constraint Grammar refers to a 'grammar' used in natural language processing (NLP), typically consisting of rules which are organised into groups and assigned tags. These rules are run iteratively, group by group, until no more disambiguating changes can be made by the grammar. The output can still contain ambiguous readings, if the grammar's rules were not sufficient to choose a preferred reading. Words which cannot be tagged (because they are unknown to the lexicon, or because their ambiguities cannot be resolved) cause knock-on effects in the grammar, because it uses a word's context to help determine its part-of-speech.

Improving the grammar's ability to resolve ambiguity and recognise words without readings from the lexicon was, thus, an important step

to take, to increase the tagger accuracy. In addition, a set of rules was introduced that merge words which have been separated during tokenisation because they contain apostrophes. Further was the introduction of a grammar rule which replaces all remaining unknown parts of speech with the tag 'E p' (*Enw priod* [proper noun]) if the word is capitalised, or 'E gb u' (*Enw gwrywaidd/benywaidd unigol* [masculine/feminine singular noun]) if not capitalised. Finally, a rule that removes all English readings where a Welsh reading is possible was added.

3.4.5. Adding additional support for multi-word units (MWUs)

Fourth, it was noted that the original version of *CyTag* was not able to apply contextual rules effectively when multi-word units (MWUs) were involved (i.e. a group of word that commonly co-occur and attribute a specific meaning that is not necessarily the sum of its parts). The solution to this problem was to add MWUs (mostly prepositions and adverbs, with some pronouns and conjunctions) to the *CyTag* lexicon, and to modify the tagger code to take account of these additions and treat them as single units. MWUs were first identified in Welsh language dictionaries (*Geiriadur Prifysgol Cymru* [Dictionary of the Welsh Language] and in King, 2007) and added to the lexicon. The code was then reworked to better identify the MWUs and rules were added to the grammar to identify these units. Future developments might therefore consider adding more of these ambiguous MWUs using grammar rules.

4. Creating the Geirfan word list: expert and end-user consultation

Once the work on the refinement of the tagger, as reported above, was complete, the entire CorCenCC v.1.0.0 dataset was re-tagged (CorCenCC v2.0.0) and frequency lists were generated from this dataset in preparation for the NCLW and WJEC partners to review them and apply the 'indigenous criteria' mentioned earlier in this paper.

First, though, it was necessary to confirm the industry partners' specific needs in relation to the word lists, and consultation with them identified the following immediate and longer-term goals:

- There was an immediate need, identified by the NCLW, to provide a list of the most essential words for inclusion in the A2 coursebook materials currently under development; their inclusion in the A1 level materials would be checked (or they would be added) at the point in the scrutiny cycle when the A1 materials were revised.
- In the medium term, once that list was established, a 'dictionary' of the items should be produced, providing details on useful collocations, conjugations and mutation patterns (i.e. the modification of initial consonants in certain grammatical/morphological environments) for each item, as informed by the content of the CorCenCC corpus. Hitherto, word lists in course books have been Welsh-English translations (as with many minoritised languages, the L1 of most learners is the dominant language, English in the case of Welsh), and this added information would provide a richer vocabulary resource.
- In the longer term, there would be a need for pedagogical word lists to supplement the initial A1/A2 list, and to provide for learners at B1, B2 and C1 levels; there would also likely be demand for pedagogical word lists for young learners. Any work on the initial list should, as far as possible, future-proof the CorCenCC infrastructure (the work on taggers, described above, contributes to this) and establish a methodology for the development of further lists.

It was necessary also to determine the target size of the initial essential word list. The NCLW partners considered that 500 words would be an appropriate size for the initial lists. While this seems relatively small compared to 'high frequency' lists compiled for English, which usually contain 1000 or 2000 words, there are a number of reasons why a figure closer to 500 is more fitting for the purpose here. First, studies indicate that it can take learners between two and nine years to master the 1000 most frequent words (see Webb and Nation, 2017: 46–48 for an overview); our purpose here was to compile a word list for courses typically completed within one or two years. Second, Zipf's law dictates an inverse relationship between the frequency of a word and its rank in a frequency list, such that the most frequent word in a list occurs twice as often as the word ranked second, and three times as often as the word ranked third, and so on. This means that the relative gain in terms of text coverage decreases for every word acquired by a learner working their way down a frequency list: Nation (2016: 162) demonstrates that in a frequency list of 1000 headwords, each 100 words learned beyond an 800-word cut-off point gains less than 1% extra text coverage. We note here that these two examples are based on research on acquisition of English; the lack (until now) of a suitable corpus means that it has not been possible to derive equivalent figures for Welsh. A third argument for focusing on just 500 items is that once these words are acquired, they will be encountered in different contexts (high frequency words are more likely to be polysemous, and to be used across a range of contexts), enriching their usefulness for learners, and supporting learners in inferring and beginning the acquisition process for new words, including those found in collocations with the known items. Lastly, knowledge of any lemmas or word family members in the list would provide access to other inflectional and derivational forms, especially as learners' command of morphology grows.

Having established our immediate aim as creating a list of the (approximately) 500 most essential vocabulary items for Welsh learners, then, the process of developing the initial version of the *Geirfan* word list began. The process drew on three main information sources: first, the statistical information derived from the newly-tagged CorCenCC corpus; second, our project team of four Welsh language/industry specialists (the Director of Teaching and Learning at the NCLW, the Welsh for Adults Examinations Officer at the WJEC, the Welsh language lead from the CorCenCC team (also an experienced applied linguist, teacher and examiner of Welsh), and a Welsh speaking linguist and researcher (PhD student on the CorCenCC project); and third, a group of experienced NCLW tutors. Together, the key experts, end-users and practitioners would ensure that indigenous criteria were embedded in decisions around the content of *Geirfan*.

In order to track the word list creation process, and as a framework to inform decisions about inclusion or exclusion in the *Geirfan*, a spreadsheet was used to record information about each candidate item. The starting list, based on frequency of occurrence in CorCenCC, had to contain more than the 500 word target identified by NCLW, in order to accommodate the possibility of items being considered unsuitable for that initial list, and therefore excluded. The process therefore began with a list of the highest frequency 750 lemmas from the CorCenCC corpus.

Each item from the frequency list was entered into a spreadsheet as a lemma, along with the following information:

- its frequency rank across the CorCenCC corpus,
- its part-of-speech,
- whether the item appeared in the existing pedagogical lists used by NCLW (the *Geirfa Graidd* lists - Morris, 2011), and
- its 'similarity rating' – this indicated the evenness (or otherwise) of distribution of the item across the three sub-corpora constituting CorCenCC: the spoken, written, and e-language sub-corpora.

There were then columns in which each of the four experts could enter comments about each item. Comments were varied, but always focused on the pedagogical usefulness of including words, and how they might be included, in the lists. Examples of points raised in the comments include:

- "include as a single entry" for *byw*, which can be verb or adjective [to live | living, alive], and *dechrau* (verb and noun) [to begin | beginning],
- "syntactically complex" (e.g. *sef* [thus]),
- "probably artificially highly ranked" because of disproportionate representation in the corpus e.g. due to repetition in television listings (*cyflwyno* [to present]); due to use in placenames (*cwm* [val-

ley]); due to presence in website metatext (*dychwelyd* [return]; *sylw* [comment]),
- for items with (e.g. regional) variants, inclusion of/cross reference to both variants (e.g. *efo* and *gyda* [with]),
- "polysemous (across parts of speech), so separate entries for different POS not necessary",
- for some items with multiple meanings, a note was made that only the most frequent meaning should be included in the pedagogical lists; for others with multiple relatively frequent meanings (e.g. *rhannu* [share/divide]), a note was made to include both meanings),
- where a plural form is higher on the ranking list, for some items (e.g. *plant* [children]) a note was made to also include the singular form (*plentyn* [child]); for others (e.g. *manylion* [details]) a note was made to only include the plural form, and
- "of limited use to learners"; primarily used in formal written texts (as indicated by low similarity score) – e.g. *cofnod* [written record, minute]; *llywodraeth* [government].

Further queries about the POS tagging were also raised in some comments. Interjections, fillers, and particles were marked as such with a view to excluding them from the final list.

An additional column headed "additions?" invited the experts to enter any items prompted by the ranked entries, that had not made the 750 item cut-off. Entries here included:

- antonyms – e.g. *diflas* [miserable] prompted by *braf* [fine] and *araf* [slow] prompted by *cyflym* [quick],
- completion of closed lexical sets or lexical coordinates - e.g. *chwith* [left] prompted by *de* [right]; colour words prompted by *coch* [red]; *llysiau* and *ffrwythau* [vegetables | fruit] prompted by *cig* [meat], and
- items motivated by inclusivity and diversity considerations – e.g. *mosg* [mosque], *teml* [temple] and *synagog* [synagog] prompted by *eglwys* [church].

As mentioned above, in order to ensure front-line practitioner input into the *Geirfan* lists, fifty experienced Learning Welsh practitioners – tutors - were invited to categorise candidate items according to whether they considered them most appropriate to A1, A2, B1 or B2+ levels of the CEFR (Common European Framework of Reference for Languages). Tutors were given the following descriptors to support their decision making (Council of Europe, 2001: 112):

- A1: Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.
- A2: (a) Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs. (b) Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.
- B1: Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel and current events.
- B2: Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.

Feedback from the tutors suggested that they viewed the task as a fairly onerous one. Nevertheless, 27 tutors provided data, and their responses were logged for each item, so that the percentage of tutors assigning the item to each of the four categories was clear. Some items were assigned to A1 level unanimously (e.g. *newydd, pobl, bach* [new, people, small]) or with a large majority (e.g. *rhoi, papur, iaith* [to give, paper, language]); some (e.g. *taliad, trafodaeth, casgliad* [payment, discussion, collection]) were unanimously assigned outside A1 level; for most, opinion was split, though usually with a clear majority view.

The final stage of the item selection process took the form of a meeting of the group of experts, and was informed by the categorisation data from tutors, along with the spreadsheet containing item information (see above) and individual experts' comments and suggested additions (indigenous criteria). While many decisions were based on the characteristics of individual items, some general principles emerged from the discussion regarding exclusion criteria; it was decided that the following should be excluded from the list:

- corpus tagging errors/repetitions,
- word forms that represent both verbs and nouns: it was agreed these should be one item rather than having separate entries for noun and verb,
- items which are high in the frequency lists only because they are part of a collocational unit,
- fillers,
- interjections/exclamations/hesitations,
- ordinals beyond *cyntaf, ail*[first | second],
- items appearing almost exclusively in formal written contexts, and
- multiple dialect variants which share the same etymology; these were merged into one entry.

Where the etymology is not shared but the frequency high e.g. *gyda/efo* [with] or *allan/mas* [out], items were to be listed separately, otherwise clear pointers to other dialect forms were to be included as additional information under the main entry. In cases where a word form has two distinct meanings (e.g. *de* [south/right]), both were to be listed separately.

The process reported above generated an essential word list – the *Geirfan* (v2.0) – of 618 items, exceeding the original target of 500. The fact that the word list was slightly larger than originally anticipated was considered by the industry partners to be unproblematic: it was agreed that compelling inclusion criteria, such as the ecological validity of including full lexical sets, and socially inclusive equivalents of relevant items, far outweighed the neatness of a canonical list length. The full *Geirfan* list (v2.0) can be found at https://corcencc.org/download/#ger.

5. Reflections on the word list creation process

This paper reports a collaborative project to create the first frequency-informed pedagogical wordlists for Welsh. Here we reflect on the ways in which expert practitioner input shaped both the process of this work, and its output (the *Geirfan* v2.0 list). We also note some language-specific features of the wordlist, and consider the next steps in materials creation that have been enabled by this project.

A useful reference point for reflection is Nation's comprehensive overview of word list creation *Making and using word lists for language learning and testing* (2016). It is important to flag immediately that the focus of that book is English language learning and testing, but also that there is a dearth of parallel volumes for other languages; hence our careful documentation in this paper regarding word list creation for Welsh. The practitioner-led approach used in this project, for the creation of the *Geirfan,* independently reached many of the same conclusions set out as "Recommendations" by Nation (2016, Section II). Points of commonality include:

- treatment of homonymous items (separate entries) (Nation, 2016: 53),
- treatment of polysemes (one entry) (Nation, 2016: 53),
- exclusion of proper nouns from the main list (Nation, 2016: 63),
- exclusion of interjections/exclamations/hesitations from the main list (Nation calls these 'marginal words' - Nation2016: 83), and
- caution was exercised regarding multiword units (MWUs); the only ones included were grammatically fixed and lexically invariable, and both the MWU and at least one component independently were included in the top 500 frequency band.[3]

[3] With the exception of MWUs created with the predicate *yn* (e.g. *yn barod* [ready]).

Decisions made about closed lexical sets also have substantial overlap with Nation's criteria for making a core word list (2016: 122–123): he suggests the inclusion of days of the week, months, seasons, compass points, family members and key numbers (one to twenty, round numbers to 100, plus thousand, million and billion), on the basis that one or more of each set fell within his initial frequency cut-off. We should note that in contrast to the 500 word working target list for this project, he was using a frequency list of the top 2000 words; there was less scope for items representing closed lexical sets to fall within the list of 500 used here. His-approach was mirrored, though, regarding sets of days of the week and numbers: *Iau* [Thursday], *Llun* [Monday] and *Gwener* [Friday] were within the 500 word frequency list, and a decision was taken to include the other four days of the week to complete the lexical set. Some numbers (*dau, deg,* etc. [two, ten]) fell within the 500 word band, and it was decided that others should be added to make a complete set of words for 1–10, 20, 100, and 1000. All four seasons and compass points were also included on the basis that at least one of each set fell within the 500 word band.However, because no months appeared within that high frequency band, none was included in the *Geirfan*.

Nation stresses that decisions around the unit of counting (types, lemmas, flemmas, word families) used in word lists should be informed by an understanding of acquisition processes, and in particular the capacity for learners to extrapolate their knowledge of one member of a word family to others – "Does each new word require new learning?" (2016: 23). Nevertheless, there is an implication that the unit of counting should be decided a priori, and then applied consistently, so that all entries would represent lemmas, for example (or flemmas, or types, or word families). The default unit of counting in the *Geirfan* was the lemma, with entries inclusive of inflectional forms and covering one part-of-speech; hence, separate entries for *gwaith* (N) [work] and *gweithio* (V) [to work], but not for *gweithio* [to work] and *gweithiodd* (3rd person singular past tense) [he/she/it worked]. However, it was not appropriate to rigidly adhere to this in all cases: decisions took into account a) the likelihood of learners knowing the specific inflectional or derivational patterning involved, and b) the relative importance of receptive and productive use: if a learner encounters a hitherto unknown inflected/derived version, will they be able to work out meaning from their knowledge of the head word, and (more challenging) if they need to produce the word will they be able to work out the appropriate morphological patterning. The application of 'indigenous criteria' in such cases generates a tension with the scholarly compulsion to design word lists within a consistent, uniform frame for entry types. It also challenges the notion of replicability: in this paper the focus on process rather than product is deliberate, and while we have attempted to present the process in a way that is replicable, we recognise that outcomes of the process will likely differ depending on context, expert representation, and indeed the target language.

In order to avoid language-specific biases (i.e. decisions made on the basis of English language structures, morphology etc.), and because of its specific purpose, the *Geirfan* was generated without reference to the principles underlying existing wordlists. Rather, as reported above, it derived wholly from the frequency list derived from the CorCenCC corpus, and from the consultation with industry experts and practitioners. In light of this, it is encouraging to note that when the *Geirfan* is subjected to Nation's taxonomy of "questions for critiquing a word list" (2016: 131–132), on the whole it stands up well to scrutiny, with all but a few of his 26 questions addressed. Nation's questions about justification for "criteria for inclusion" and "subjective criteria" are difficult to address in a generalised way, in the context of the word by word scrutiny applied in this project; as noted above, while some general principles did emerge from the consultations, there were a large number of decisions for individual words, based on the industry experience of the experts and practitioners. The question "Were the lists checked against competing lists?" is barely relevant for the *Geirfan*; it is the first core pedagogical wordlist of its kind for Welsh, so has no competing lists (*Yr Amliadur*, which did not have a pedagogical focus, and the non-corpus based *Geirfa Graidd,* both played a role in developing the *Geirfan*).

Research on word list creation has been dominated by work on the English language. Building word lists in languages other than English necessitates teasing apart the principles and practices that are applicable to all languages, from those which derive from the specific morphology and structure of English, and in turn attending to relevant features of the target language. For Welsh, these include formation of plurals, conjugated prepositions, masculine and feminine forms of numbers, and the three kinds of initial mutation. Another specific language feature of Welsh, the counting system, means that knowledge of numbers 1–10 and 20, can enable a learner to form all numbers to 100, so unlike English, it is not necessary to include 11–19 in a list of core numbers. Appropriate accommodation of these language specific features into the word lists was possible because of the collaborative nature of this project: CorCenCC was tagged for Welsh by corpus linguists and tags were refined in the course of the project; in the process of extracting frequency data from the corpus, (applied) linguists generated language-related questions for the experts and practitioners to consider, and they in turn brought their understanding of learner experience to decisions about item inclusion/exclusion.

As set out in Section 4 above, the project industry partners had identified three main requirements relating to CorCenCC and the frequency information it offers. The immediate need was for a list of the most essential words for A1 and A2 materials. This need has been met within this project, through the creation of the *Geirfan*, which is already being used by NCLW in the revision of their 2023 A2-Level coursebook. The project has also established a foundation for achieving the second, medium term need: a 'dictionary' of the essential items, with details on useful collocations, conjugations and mutation patterns for each item, as informed by the content of the CorCenCC corpus. Work on a dictionary is already underway, using corpus n-grams to identify idioms, collocations and phrases to be included in the entries and using corpus data as a basis for the examples used in the word list. The process of building both the word list and the dictionary has established a methodology and framework that can be replicated and extended for learners at different levels or with different requirements. The work undertaken on the corpus infrastructure as part of this project (the work on taggers, for example), has future-proofed CorCenCC for the development of further lists, thus addressing the third need identified at the outset of the project.

The collaboration on this project represents an innovative symbiosis of corpus-based methods and expert-led regulation, with the CorCenCC team adjusting elements of the corpus infrastructure to enable extraction of frequency data of maximum usefulness, and the industry partners bringing expertise and experience to inform decision making on inclusion criteria other than frequency. This synergy between the data and the team - the ability to take the raw data and shape it intelligently and appropriately – not only positions the *Geirfan* as a fit for purpose resource for adult learners of Welsh, but has also established a robust and versatile framework for future word list creation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research on which the *Geirfan* word list was based was funded by the UK Economic and Social Research Council (ESRC) Impact Acceleration Account (IAA) at Cardiff University. The CorCenCC dataset (which underpins the research presented here), and the *Geirfan* word list data (v2.0), can be found in the Cardiff University

data catalogue at https://doi.org/10.17035/d.2020.0119878310 and https://doi.org/10.17035/d.2022.0234583226, respectively.

References

Baker, C., Andrews, H., Gruffydd, I., Lewis, G., 2011. Adult language learning: a survey of Welsh for Adults in the context of language planning. Eval. Res. Education 24 (1), 41–59.

Bauer, L., Nation, P., 1993. Word families. Int. J. Lexicogr. 6 (4), 253–279.

Brezina, V., Gablasova, D., 2015. Is there a core general vocabulary? Introducing the new general service list. Appl. Linguist. 36 (1), 1–22.

Council of Europe, 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment [online]. Retrieved from.

Dang, T.N.Y., Webb, S., 2016. Making an essential word list for beginners. In: Nation, I.S.P. (Ed.), Making and Using Word Lists for Language Learning and Testing. Amsterdam, John Benjamins, pp. 153–167.

Davies, M., 2008. Word Frequency Data from The Corpus of Contemporary American English (COCA) Data available at.

Donnelly, K., Deuchar, M., 2011. Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text. In: Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop, Riga, Latvia, pp. 17–25.

Donnelly, K., 2013. Eurfa v3.0 - Free (GPL) Dictionary (Incorporating Konjugator and Rhymer) [online]. Retrieved from:.

Elder, C., McNamara, T., 2016. The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. Lang. Test. 33 (2), 153–174.

Engels, L.K., 1968. The fallacy of word counts. Int. Rev. Appl. Linguist. Lang. Teach. 6 (3), 213–231 Geiriadur Prifysgol Cymru: A Dictionary of the Welsh Language (1950 -). Cardiff, University of Wales Press.

Hwb. 2021a. Curriculum for Wales [online]. Retrieved from: https://hwb.gov.wales/curriculum-for-wales/ [Accessed 09/11/2022].

Hwb. 2021b. Curriculum for Wales: the journey to curriculum roll-out [online]. Retrieved from: https://hwb.gov.wales/curriculum-for-wales/curriculum-for-wales-the-journey-to-curriculum-roll-out#shared-expectations-for-curriculum-roll-out [Accessed 09/11/2022].

Ishikawa, S., 2019. A reconsideration of the construct of "a vocabulary for Japanese learners of English": a critical comparison of the JACET wordlists and new general service lists. Vocab. Learn. Instr. 8 (1), 1–7.

JACET Kihongo Kaitei Tokubetsu Iinkai (JACET, Special Committee forRevision of the JACET Wordlist), 2016. The New JACET List of 8000 basicwords. Tokyo: Kirihara Shoten.

Kilgariff, A. 1998. BNC database and world frequency lists [online]. Retreived from: https://www.kilgarriff.co.uk/bnc-readme.html [Accessed 09/11/2022].

King, G., 2007. Modern Welsh Dictionary: A Guide to the Living Language. OUP, Oxford.

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., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M., Scannell, K., 2020a. CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes—the National Corpus of Contemporary Welsh. Cardiff University See:.

Knight, D., Morris, S., Tovey-Walsh, B., Fitzpatrick, T., Anthony, L., 2020b. Yr Amliadur: Frequency Lists For Contemporary Welsh. Cardiff University See:.

Learn Welsh. 2022. Learn Welsh Statistics 2020-21 [online]. Retreived form: https://learnwelsh.cymru/about-us/statistics/statistics-2020-21/ [Accessed 09/11/2022].

Leech, G., Rayson, P., Wilson, A., 2001. Word Frequencies in Written and Spoken English: Based On the British National Corpus. Longman, London.

Morris, S., 2011. Geirfa Graidd i'r Gymraeg: creating an A1 & A2 core vocabulary for adult learners of Welsh—a Celtic template? J. Celt. Lang. Learn. 15/16, 27–43.

Nation, 2016. Making and Using Word Lists For Language Learning and Testing. John Benjamins Publishing Company.

Nation, I.S.P., 2001. Learning Vocabulary in Another Language. Cambridge, Cambridge.

NCLW. 2021. Adroddiad Blynyddol - Annual Report [online]. Retrieved from: adroddiad-blynyddol-2020-2021-ar-lein.pdf (dysgucymraeg.cymru) [Accessed 04.04.22]

Neale, S., Donnelly, K., Watkins, G., Knight, D, 2018. Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. In: Proceedings of LREC 2018, Miyazaki, Japan, pp. 3946–3954.

ONS. 2021. Welsh language data from the Annual Population Survey: 2021 [online]. Retrieved from: https://gov.wales/welsh-language-data-annual-population-survey [Accessed 09/11/2022].

Piao, S., Rayson, P., Knight, D., Watkins, G., 2018. Towards a welsh semantic annotation system. In: Proceedings of LREC 2018, Miyazaki, Japan, pp. 980–985.

Pill, J., 2016. Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: health professionals interacting with patients. Lang. Test. 33 (2), 175–193.

Rayson, P., Archer, D., Piao, S., McEnery, T., 2004. The UCREL semantic analysis system. In: Proceedings of the LREC-04 workshop, beyond named entity recognition semantic labelling for NLP tasks, Lisbon, Portugal, pp. 7–12.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, pp. 44–49.

Schmitt, N., Schmitt, D., 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Lang. Teach. 47 (4), 484–503.

Schmitt, N., 2010. Researching Vocabulary: A Vocabulary Research Manual. Palgrave Macmillan, London.

StatsWales. 2022. Ability of people aged here or older to speak Welsh by local authority and single year of age, 2011 and 2021 [online]. Retrieved from: https://statswales.gov.wales/Catalogue/Welsh-Language/Census-Welsh-Language/abilityofwelshpeopleaged3oroldertospeakwelsh-by-localauthority-singleyearage-censusyear [Accessed 06/12/2022].

Toutanova, K., Klein, D., Manning, C., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLTNAACL 2003, Edmonton, Canada, pp. 173–180.

Webb, S., Nation, I.S.P., 2017. How vocabularly is learned. Int. J. Appl. Linguist. 169 (2) 321-227.

Welsh Government. (2021a). School's census results: April 2021 [online]. Retrieved from: https://gov.wales/schools-census-results-april-2021-html [Accessed 09/11/2022].

Welsh Government. (2021b). Welsh language use in Wales (initial findings): July 2019 to March 2020 (revised) [online]. Retrieved from: https://gov.wales/welsh-language-use-wales-initial-findings-july-2019-march-2020-revised-html#section-79590 [Accessed 09/11/2022].

West, M., 1953. A General Service List of English words. Longman, London.

WJEC. 2020. Looking for a qualification? [online]. Retrieved from: https://www.wjec.co.uk/qualifications/ [Accessed 09/11/2022].

Zeno, S.M., Ivens, S.H., Millard, R.T., Duvvuri, R., 1995. The educator's word frequency guide. Touchstone Appl. Sci. Assoc..