Comparative Evaluation of Translation Memory (TM) and Machine Translation (MT) Systems in Translation between Arabic and English

Khaled Mamer Ben Milad

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

College of Arts and Humanities: Department of Modern Languages, Translation and Interpreting

2021

DECLARATIONS AND STATEMENTS

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 ....khaled............................ (candidate)

Date ....22-07-2021.....................................

STATEMENT 1

I, Khaled Mamer Ben milad, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

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

appended.

Signed .... khaled............................ (candidate)

Date .... 22-07-2021...................................

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available to outside organisations.

Signed .... khaled................................ (candidate)

Date .... 22-07-2021...............................

ABSTRACT

In general, advances in translation technology tools have enhanced translation quality significantly. Unfortunately, however, it seems that this is not the case for all language pairs. A concern arises when the users of translation tools want to work between different language families such as Arabic and English. The main problems facing Arabic<>English translation tools lie in Arabic's characteristic free word order, richness of word inflection – including orthographic ambiguity – and optionality of diacritics, in addition to a lack of data resources. The aim of this study is to compare the performance of translation memory (TM) and machine translation (MT) systems in translating between Arabic and English.

The research evaluates the two systems based on specific criteria relating to needs and expected results. The first part of the thesis evaluates the performance of a set of well-known TM systems when retrieving a segment of text that includes an Arabic linguistic feature. As it is widely known that TM matching metrics are based solely on the use of edit distance string measurements, it was expected that the aforementioned issues would lead to a low match percentage. The second part of the thesis evaluates multiple MT systems that use the mainstream neural machine translation (NMT) approach to translation quality. Due to a lack of training data resources and its rich morphology, it was anticipated that Arabic features would reduce the translation quality of this corpus-based approach. The systems' output was evaluated using both automatic evaluation metrics including BLEU and hLEPOR, and TAUS human quality ranking criteria for adequacy and fluency.

The study employed a black-box testing methodology to experimentally examine the TM systems through a test suite instrument and also to translate Arabic English sentences to collect the MT systems' output. A translation threshold was used to evaluate the fuzzy matches of TM systems, while an online survey was used to collect participants' responses to the quality of MT system's output. The experiments' input of both systems was extracted from Arabic<>English corpora, which was examined by means of quantitative data analysis.

The results show that, when retrieving translations, the current TM matching metrics are unable to recognise Arabic features and score them appropriately. In terms of automatic translation, MT produced good results for adequacy, especially when translating from Arabic to English, but the systems' output appeared to need post-editing for fluency. Moreover, when retrieving from Arabic, it was found that short sentences were handled much better by MT than by TM. The findings may be given as recommendations to software developers.

Arabic abstract (الخلاصة)

ساعدت أدوات الترجمة الحديث في تحسين جودة الترجمة بشكل كبير. ولكن، يبدو أن هذا ليس هو الحال بالنسبة لجميع اللغات، فاستعمال تلك الأدوات لترجمة بين لغتين مختلفتين في الانتماء مثل العربية والإنجليزية يواجه صعوبات. مثل هذه الصعوبات الاختلافات في ترتيب الكلمات في جمل اللغة العربية والإنجليزية (differences in word-order between English and Arabic)، الغنى النحوي للكلمة (richness of word inflection) - مثل الغموض الإملائي (orthographic ambiguity) وأدوات التشكيل "تكون اختيارية" (optionality of diacritics)، بالإضافة إلى النقص في موارد اللغوية (lack of data resources). فهذه الدراسة تهدف إلى تقيم أداء أدوات الترجمة كنظامي ذاكرة الترجمة (Translation Memory) والترجمة الآلية (Machine Translation) في ترجمتها بين العربية والإنجليزية.

هذه الدراسة تقيم النظامين بناءً على معايير محددة تتعلق بالاحتياجات والنتائج المتوقعة لكل نظام: الجزء الأول من الدراسة يقيِّم اداء لمجموعة من تطبيقات ذاكرة الترجمة في خاصية استرجاع نصوص عربية تتضمن بها فروق لغوية طفيفة عن النص المراد ترجمته. تقيم نظام ذاكرة الترجمة كان مبني على أنها تستخدم (Edit Distance metrics) لقياس التشابه بين النصين. لذلك الدراسة توقع أن نظام ذاكرة الترجمة - سيعرض نصوص ذاكرته التي بها أحد الفروق اللغوية كالمذكورة أعلاه - بنسبة مئوية منخفضة. تلك النسبة المنخفضة قد تؤدي منع استخدامه من جديد. الجزء الثاني من الدراسة يقيِّم أداء مجموعة من تطبيقات نظام الترجمة الآلية التي تستخدم (Neural Machine Translation Approach). الافتقار الى الموارد اللغوية مع الغنى اللغوي للعربية، المتوقع ان يسبب ذلك صعوبات لنظام الترجمة الآلية في أنتاج ترجمة جيدة. الترجمة المنتجة آليا من تم تقييمها يدويا وآليا: يدويا باستعمال طريقة تصنيف لـ (Adequacy and Fluency) وآليا باستعمال معياري (BLEU and hLEPTOR metrics).

الدراسة استخدمت طريقة تسمى (a Black Box method) لتقيم كلا من النظامين من خلال أجرى مجموعة من الاختبارات التجريبية: حيث تم استخدام اختبار يسمى (a Test Suite approach) ومن خلاله استخدمنا (a translation threshold) للحكم على مخرجات نظام ذاكرة الترجمة، بينما تم استخدام أستبان عبر الإنترنت (online survey) لجمع أجوبة المشاركين على جودة ترجمات نظام الترجمة الآلية، البيانات المتحصلة عليها تم فحصها وتحليلها كميا (quantitative analysis).

تظهر نتائج الدراسة ان نظام ذاكرة الترجمة غير قادر على التعامل بشكل ملائم على استرجاع نص العربية من الذاكرة به فوارق لغوية ضئيلة عن النص المراد ترجمته عند الترجمة من العربية الى الإنجليزية. أما عن نظام الترجمة الآلية فأنها رغم أنها أظهرت نتائج جيدة وخاصة في (Adequacy) في الترجمة من العربية إلى الإنجليزية، إلا أن يبدو الترجمة تحتاج الى مراجعة تحريرية (Fluency). علاوة على ذلك، وجد أن نظام الترجمة الآلية يتعامل مع الجمل القصيرة أفضل بكثير من نظام ذاكرة الترجمة.

ACKNOWLEDGEMENTS

In the name of ALLAH, Most Gracious, Most Merciful

At the very outset, all praises to ALLAH (Alhamdulillahirobbil'alammin) SWT, the source of wisdom and knowledge, for giving the researcher the strength, ability, and opportunity to complete this thesis in fulfilment of the requirements for the degree of PhD. And the best regards (Sholawat and Salam) always deliver to our Prophet Muhammad, peace be upon him, as the messenger of ALLAH.

The present thesis, resulting from four and half years of research, would not have been possible without the support of many people. With great pleasure, I would like to acknowledge those people:

First, I would like to express my deepest appreciation and sincere thanks to the primary supervisor of the thesis Professor Andrew Rothwell (Andy) and the secondary supervisor Dr Maria Fernandez Parra for their constructive feedback and judicious comments – their support throughout my doctoral candidacy from the inception of this research until completion. They were also willing to take me on as an MA teaching assistant in my second and third year of study. I was very lucky to have tremendous supervisors like this team. It was a real privilege and a great pleasure for me to learn from two of the pioneers of the Translation Technology.

I would also like to extend my special thanks to Tom Pritchard – an Application Support officer at College of Science, we did collaborative work regarding machine translation using API application, and also special thanks to COAH (College of Arts and Humanities) at Swansea university – for providing me facilities to access Labs and software of translations which have helped me become more confident with my research.

I very much appreciate the assistance of Prof Andy Way – who issued an invitation to visit ADAPT Centre[1] at the School of Computing at Dublin City University (DCU), Ireland. The visit to the ADAPT Centre provided me with important skills required to complete my research. Thanks also goes to Goudong Xie and Aaron Lifeng – Ph.D candidates at ADAPT Centre – who helped me to set up a script of BLEU and hLEPOR metrics.

I would also like to express my gratitude to Fran Cetti for her work in proofreading this research.

Last, but certainly not least I would like to express my deepest gratitude and most heartfelt thanks to my small family (wife and sons) and large family (parents and siblings) for their unwavering support throughout my research.

Funding and grants

This achievement would not have been possible without the funding sources. I gratefully acknowledge the financial sponsor – the government of my home country Libya – for awarding a full scholarship to pursue my postgraduate studies in the United Kingdom.

I also want to acknowledge LDC (Linguistic Data Consortium) for its grant "Arabic-English corpus LDC2004T18" through 'Autumn 2017 LDC data scholarship program'[2]. The corpus provided me with the database required to complete the experiments of machine translation systems evaluation.

I am also grateful to Santander Mobility who awarded me a travel grant which helped fund the trip to the ADAPT Centre at Dublin City University in August 2019.

[1] ADAPT: The Global Centre of Excellence for Digital Content and Media Innovation (adaptcentre.ie)

[2] Data Scholarships | Linguistic Data Consortium (upenn.edu)

Dedication (الإهداء)

بسم الله الرحمن الرحيم

نحمده الله وحده على إنجاز هذه البحث...

نصلي ونسلم على نبينا محمد وعلى آله وصحبه آجمعين ...

أهدي هذه العمل المتواضع إلى:

*** روح 'جدتي' الزكية الطاهرة- اللهم اغفر لها ورحمها – التي تربيت في حضنها وشجعتني على تحصيل العلم،**

*** الوالدين الكريمين حفظهما الله الذي رضائهم ودعائهم كان الأساس في هذه النجاح، وكذلك عمتي،**

*** زوجتي التي تحملت عبيء طوال مشوار الدراسة، وأبنيا الاثنين لانشغال بالدراسة عنهم،**

*** أخوتي وأخواتي الذين كانوا معي بالدعاء.**

CONFRENCE PRESENTATIONS DERIVED FROM THIS THESIS

(2020, November), ‘A Comparative Evaluation of the Performance of Translation Memory in the Retrieval of Arabic-English Segments Containing a Sub-segment Move’, Proceedings of Translating and the Computer Conference in London, AsLing, TC 42 https://asling.org/tc42/ .

(2020, November), ‘Comparative Evaluation of Quality Output from Neural Machine Translation Systems: Arabic<>English Translation’, Proceedings of Translating and the Computer Conference in London, AsLing, TC 42 https://asling.org/tc42/ .

(2021, March), ‘Comparative Evaluation of Translation Memory (TM) Retrieval of Arabic-to-English an Inflectional Affix Intervention’, 11th Annual Graduate Student Conference in Translation Studies, York University (Toronto), Canada. https://www.glendon.yorku.ca/transconf/

(2021. July) A Comparison of the Word Similarity Measurement in English-Arabic Translation Memory Segment Retrieval including an Inflectional Affix Interventio*n*, TRanslation and Interpreting Technology Online (TRITON 2021) http://triton-conference.org/accepted-papers/

Table of Contents

List of figures

List of tables

List of equations

Abbreviations

ALPAC	**Automatic Language Processing Advisory Committee**
ATB	**Penn Arabic TreeBank**
BLEU	**BiLingual Evaluation Understudy**
BPE	**Byte-Pair Encoding**
CAT	**Computer-Aided Translation**
CL	**Classical Language.**
CNN	**Convolution Neural Network**
DGT	**European Commission’s Directorate-General for Translation**
DP	**Dynamic programming**
EBMT	**Example-based machine translation**
hLEPOR	**harmonic mean of enhanced Length Penalty, Precision, n-gram Position difference Penalty and Recall**
HMT	**Hybrid Machine Translation**
IDF	**Inverse Document Frequency**
LDC	**Linguistic Data Consortium**
LMVR	**Linguistically Motivated Vocabulary Reduction**
LP	**length penalty**
LSTM	**Long short-term memory**
MSA	**Modern Standard Arabic**
MT	**Machine Translation**
NMT	**Neural Machine Translation**
NPD	**n-gram position difference**
PBSMT	**Phrasal Based Statistical Machine Translation**
POS	**Part of Speech**
RBMT	**Rule-based machine translation**
RNN	**Recurrent Neural Network**
SL	**Source Language**
SMT	**Statistical Machine Translation**
SVO	**subject-verb-object.**
TAUS	**Translation Automation User Society**
TL	**Target Language**

TM	**Translation Memory**
TMX	**Translation Memory eXchange**
TU	**Translation Unit**
VSO	**verb-subject-object**
WPM	**Weighted Percent Match**

CHAPTER ONE

INTRODUCTION

1.0 Introduction

Technological developments and increasing internet accessibility have encouraged an expansion in the use of computer-aided translation (CAT) tools and machine translation (MT) – it has been estimated that around 99% of translations worldwide are currently produced by machines (TAUS 2016: 74). One of the most significant features of computer-aided translation tools is the translation memory (TM). The initial idea for TM technology was proposed in the late 1970s and early 1980s (Arthern 1979; Kay 1980; Melby 1981). In his paper 'The Proper Place of Men and Machines in Language Translation' (which was not widely distributed until 1981), Kay proposed the use of a bilingual concordance, which laid the basis for the creation of TM. The function of a TM, which consists of a database of aligned pairs of source and target segments, is to retrieve a translation of the input segments by finding exact or close matches in the database. TM tools were designed to support translators in their work; it means they can reuse the translations of highly similar source segments during the translation process and not have to re-translate the same text twice. Meanwhile, neural machine translation (NMT) has become the mainstream architecture for machine (or automatic) translation, and its paradigm is recognised as enhancing translation quality (Bahdanau et al. 2014; Cho et al. 2014). NMT is a deep-learning-based approach: it uses an artificial neural network that has the ability to learn a statistical model for translating text from the source language into the target language(s). The key benefit of the neural approach is that a single model can be trained directly on source and target texts; it no longer requires the three-model systems used in the statistical MT approach.

The distinction between TM and MT systems is that CAT tools which use the TM system enable the translator to control the translation of the text while employing these tools to help increase their productivity (Bowker 2002), whereas with MT systems, the machine controls the translation, typically without human intervention. In the past, TM and MT were used separately in translation workflows. Later, however, the work of the two systems was combined, with the MT system being used as a back-up mechanism when the TM failed to retrieve an appropriate match (Federico et al., 2012). More recent developments have integrated the advantages of TMs into MT systems in order to improve their quality: one approach applies a fuzzy match-repair technique to repair the TM proposal (Ortega et al. 2016), while the NMT's data augmentation method augments the source sentence with fuzzy matches retrieved from the TM (Bulte and Tezcan 2019).

CAT tools seem to be designed to work well with European languages, however Arabic's unique characteristics, such as its richness of morphology and freedom of word order, complicate the functions of these tools. Its flexibility means that the sense of a sentence can be expressed in various ways using the same surface words but in a different order: for example, VSO (verb-subject-object) or SVO (Elming 2008). Arabic's morphological richness means that the language contains not only orthographic ambiguity and optionality of diacritics but also a large set of inflectional morphology, thus creating a huge number of surface forms (Habash 2010). This research aims to investigate how translation tools perform with the difficulties posed Arabic.

The study uses an experimental investigation (based on a quantitative methodology) and an evaluation technique that treats each translation tool – TM or MT – as a 'black-box' (Simard and Fujita 2012) to evaluate and compare the performance of a specific set of systems. It first investigates the performance of current TM matching metrics when retrieving Arabic source segments with complex linguistic features, including differences in word order, inflectional affixes and the omission of the Hamza marker, and sets out to determine the extent to which these features affect the fuzzy matching scores. It then uses the TAUS adequacy and fluency ranking to evaluate the performance of NMT systems compared with automatic evaluation metrics. This black-box evaluation is intended to demonstrate which translation tool performs better and whether any difficulties still exist.

1.1 Statement of the problem

The use of technology has greatly enhanced the translation industry; however, it seems that digital translation tools are not equally successful for all language pairs. They appear to face significant challenges in supporting the translation of Arabic content in particular, as its flexible word order and morphological richness render it difficult for a machine to understand. As a consequence, the shortcomings of the translations produced by the TM retrieval process and MT systems reveal deficiencies in many of these systems.

The core assumption behind this evaluation of the two types of system is that translators usually research whether the use of TM retrieval or MT systems best suits their language pairs.

The aforementioned problems highlight the need for a study that investigates the extent to which the development of translation tools has succeeded in overcoming the difficulties associated with handling the complexities of Arabic linguistics. As a response to this need, this research aims to evaluate the performance of translation tools (i.e. TM and MT systems) when translating from Arabic into English or vice versa, beginning with the design of a set of research questions to direct its investigation.

1.2 Motivation of the research

Arabic and English belong to different language families: Arabic, as a Semitic language, has unique linguistic characteristics such as a flexible word order and rich morphology, while English, an Indo-European language, has a fixed word order and simple morphology. This shows the importance of investigating how well translation technology systems handle these fundamental differences when translating between these language pairs.

The goal of the current study is to conduct an experimental evaluation of two types of translation tools (TM and MT) by undertaking a comparative analysis of the performance of different systems. Such research should benefit Arabic translators looking for a translation tool that meets their language requirements, as well as revealing the weaknesses of these systems to their developers. The study further hopes to contribute to the field of translation tool evaluation for other language pairs which, like Arabic and English, possess differing morphological features.

The following section provides a brief overview of Arabic's main linguistic characteristics that will be studied in the research, the features of TM and the development of MT.

1.3 Background to the research

This section establishes the context of the different aspects of the research. Section 1.4.1 explains why the selected Arabic linguistic features are important to the analysis of the performance of translation tools, while section 1.4.2 outlines the history and functionality of TM, including an explanation of why TM finds Arabic problematic. Section 1.4.3 gives a brief overview of the history of Arabic<>English MT up to the emergence of the neural paradigm, and details why difficulties arise in NMT with Arabic. Section 1.4.4 then describes the evaluation criteria used in the research.

1.3.1 Arabic linguistics

1.3.1.1 Historical introduction

Arabic, a Semitic language that developed in the Middle East, is one of the six official languages of the United Nations and the fourth most common language in the world. According to Internet World Stats,[3] it is the mother tongue of about 440 million people, distributed mainly across the Arab countries.

Historically speaking, the current standard form of Arabic became widely disseminated after the emergence of Islam with the Prophet Muhammad (Peace Be Upon Him) in Mecca and Medina in Saudi Arabia at the start of the seventh century. Arabic is therefore associated with Islam and the Holy Quran, although the language existed centuries before Islam – this pre-Quranic Arabic is the so-called Classical Language (CL). The term CL is principally used to describe the language used in the Quran and the earliest Arabic literature; however, over the centuries, CL has evolved into a simplified form known as Modern Standard Arabic (MSA) (اللغة الفصحى / *alfushaa allugha*). MSA is the official Arabic language that can be understood by all speakers of Arabic. As a result of colonial influences in the region, each Arab country has its own local dialect in addition to MSA. Colonisation affected the language in terms of word borrowings from English or French and different ways of speaking, but these vernacular forms of Arabic are for the most part mutually comprehensible across the region.

MSA (*alfushaa*) is the official language of 22 Arab countries across North Africa, the Middle East and the Gulf region. MSA (referred to as 'Arabic' in this research) is the form of Arabic used in intellectual life, and is taught not only in educational institutions in the Arab world but also in universities around the world. It is also the standard language most commonly used in books, news broadcasts, formal speeches, movies, etc. Most academic resources, such as natural language processing tools, and most available parallel corpora that include an Arabic pair are written in MSA (Abdelali 2004). As MSA is universal as well as being the variety of

[3] https://internetworldstats.com/stats7.htm

Arabic most frequently used in linguistic research, this study has adopted MSA as its subject of investigation.

Arabic is characterised by many particular features that distinguish it from other language families such as English. The following section describes some of the characteristics that have been selected for evaluation in this research.

1.3.1.2 Word order

One of the major differences between Arabic and English syntax lies in the word order. English maintains a strict word order: subject-verb-object (SVO). As such, it differs significantly from Arabic, which is regarded as far more flexible. This flexibility means that the sense of a sentence can be expressed in various ways by using the same surface words but in a different order: VSO, SVO, VOS or OSV (Elming 2008). However, the basic word order in Arabic is the VSO pattern, known as a 'verbal sentence', where the verb precedes the subject. The pattern of SVO, where the verb follows the subject, is arrived at by moving the subject into the initial position in the sentence – this is known as a 'nominal sentence' (Abdul-Raof 1998, cited in Sado Al-Jarf 2007). From a stylistic perspective, the linguistic difference between verbal sentences and nominal ones is one of emphasis: if the emphasis is on the doer, the sentence begins with a noun (i.e. subject) and the word order is SVO, but if the emphasis is on the deed (i.e. action), the sentence begins with a verb and the word order is VSO (Habash, 2007: 294).

To simplify the idea of difference conveyed by word order, take the two English segments[4] below. If the word order is ignored, the segments will provide an exact match – each segment has the same four words and a question mark, yet the segments have different meanings:

(1) Will you do it?

(2) Do you will it?

[4] http://www.city-data.com/forum/writing/1115620-two-sentences-have-same-words-but.html

The scenario above, however, can be applied to segments in Arabic and the meaning can be identical, as in the example of two versions of the Arabic sentence below – they express the same meaning in English:

(1) سيفرح الطفل بلعبته الجديدة / *sayafrah altifl bilaebatih aljadida*

(2) الطفل سيفرح بلعبته الجديدة / *altifl sayafrah bilaebatih aljadida*

(English translation: The child will be glad about his new game.)

In the Arabic examples, sentence (1) begins with the verb سيفرح / *sayafrah* / followed by the subject الطفل /*altifl/* , while sentence (2) begins with the subject followed by the verb; however, the meaning is identical whichever the order, and can be expressed in a single English translation despite the different sentence structures. This feature raises the following question in the context of TM retrieval: if one of these versions were given to a translator as a source text to translate but their TM database contained the other version, would the TM matching metrics accurately compute the high similarity? And if not, why not? The experiment in Chapter Four, section A, sets out to investigate this question, which has practical and commercial implications for translators and CAT tool developers. With regard to automatic translation, the question is whether the MT systems are able to construct a verb-initial sentence. If not, this may lead to less fluent translations between Arabic and English. The experiments in Chapters Six and Seven are designed to answer this question.

1.3.1.3 Morphology

Morphology studies the way morphemes combine with each other to create new word forms and express different meanings. A characteristic of Arabic is the richness of its morphology – in particular, its high incidence of inflectional morphology – leading to the creation of a huge number of surface forms, in contrast to the simpler morphology of English. Inflection, as an aspect of morphology, involves changing the grammatical form of the same word in order to express changes in tense, gender, number, etc. This results in an increase in the forms of words in Arabic corpora, while the number of occurrences of each form is correspondingly relatively low. According to Al-Kabi et al. (2013), owing to its rich morphology, an Arabic corpus has more surface forms than an English corpus of the same size. The inflection of verbs is called ‘conjugation’ and the inflection of nouns, ‘declension’. This research selects verb inflections to represent the morphology in its investigation of TM retrieval, since in Arabic the verb form is regarded as the crucial part of a sentence due to its preference for a verbal sentence structure.

The overwhelming majority of Arabic verbs have roots consisting of three characters in which the inflectional affix (i.e. a character) that shapes the template can only be positioned as a prefix or a suffix, while the affix string may encompass one or more characters. Habash (2010) states in his *Introduction to Arabic Natural Language Processing* that verb inflections have a limited number of patterns: ten basic templates for a three-character root and two templates for a four-character root. This means that the triliteral (three-character-root) verb can be transformed from one template into another by simply attaching a prefix (an initial attachment) or suffix (a final attachment). For example, the Arabic root “س م ع” has two genders: masculine (يسمع) /yasmae/, (he listens) and feminine (تسمع) / tasmae / (she listens); three grammatical numbers: singular [e.g. (أسمع) / 'asmae / (I listen)], dual (يسمعان) / yusmiean / (they listen) and plural [e.g. (يسمعون) /yasmaeun/ (they listen)]; two main tenses: the perfective pattern [e.g. (سمع) /sumie/ he listened), the imperfective pattern [e.g. (يسمع) /yasmae/, (he listens) or (سيسمع) /sayasmae/, (he will listen) (Ameur et al. 2020).

The verb conjugation involves the creation of new stems from the verb’s root (the base of the verb form) using specific verbal templates. Neme (2011) explains that the combination of a root with a pattern produces an inflected form in which the root signifies a morphemic abstraction for a verb, while the pattern is a template of characters (indices) surrounding the root consonants. The verb’s tense – and other aspects such as gender and number – are generally represented using the rules of inflectional verb morphemes. Tenses are used in either the perfect or imperfect form; the former indicates the past tense while the latter indicates the present or future tense. The language uses a unique inflection system: for example, verbs in the past tense are often designated by suffixes, whereas verbs in the present or future tense are often identified by a prefix. Numbers are classified as plural, dual or singular, with two gender categories, feminine and masculine. The number and gender features can be integrated with the verb’s tense and expressed in single-word forms (Habash 2007). For example, when the prefix يـ precedes the root فعل / *faʕala* (‘do’), the new form يفعل / *yafʕalu* (‘he does’) indicates the verb is in the imperfect tense and represents the third person singular masculine; when the suffix ـت follows the form فعل / *faʕala* (‘do’), the new form فعلت / *faʕalat* (‘she did’) indicates that the verb is in the perfect tense and signifies the third person singular feminine. Thus, four grammatical functions are expressed in a one-word form (Shamsan and Attayib 2015).

Attached morphemes of verbs may be affixes or clitics – affixes (i.e. prefixes, suffixes) attach to the stem, while clitics (proclitics and enclitics) attach to the stem after affixes. A clitic is a linguistic unit that is pronounced and written like an affix but is grammatically independent (Alqudsi et al. 2014). In other words, multiple affixes and clitics can appear in one word, with the result being that some words contain a meaning that can only be expressed in English by a whole sentence. The example in Figure 1.1 (below) shows the construction of the Arabic word فسيأكلونها / *fasyakulunaha* ('and they will eat it').

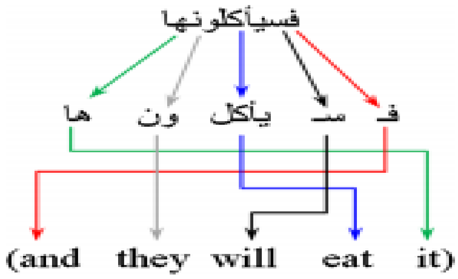


Figure 1.1: An example of word inflection in Arabic (Ezzeldin and Shaheen 2012:282)

As seen in Figure 1.1, a single Arabic token فسيأكلونها ('and they will eat it') is formed by و ('and'), ف ('will'), the base lexeme يأكل ('eat'), the plural subject pronoun ون ('they') and the singular object pronoun ها ('it'). The word-form example above has, in addition to the word root, two prefixes and two suffixes; its meaning can only be conveyed in English by a full phrase using five words.

The process of combining affixational and clitic morphemes can involve a diversity of morphological and phonological adjustment rules for a single word, giving rise to problems in MT or TM retrieval. In TM retrieval, the similarity metrics may find a combination of morphological inflections problematic. According to Planas and Furuse (1999: 331), the matching measurement would have difficulty in recognising that sentence (3) in the sequence below is more similar to sentence (1) than is sentence (2).

(1) The wild child is destroying his new toy.

(2) The wild chief is destroying his new tool.

(3) The wild children are destroying their new toy.

To exemplify this scenario in Arabic, the two source sentences below are composed of the same units although one of these is inflected differently (one sentence includes a different verb-inflectional character).

(1) جهزت الأم الطعام / *jhzt al'umu altaeam* / The mother prepared the food.

(2) تجهز الأم الطعام / *tujahiz al'umu altaeam* / The mother prepares the food.

In the above example, the verb جهزت in sentence (1) refers to the past tense, ending with the suffix ت, while the word تجهز in sentence (2) refers to the present tense, beginning with the prefix تـ. The deletion of the suffix ت and insertion of the prefix تـ produced a different tense. In terms of TM retrieval, this example raises the following question: if one of these sentences were given to a translator as a source text but their TM database contained the other version, would the algorithm penalise a different combination of the inflectional affixes heavily? The experiment in Chapter Five, section A, investigates this question.

With automatic translation, the analysis of inflection affixes gives rise to a complicated process of analysing and generating Arabic words; morphological analysis is one of the challenges facing corpus-based MT approaches as it has the effect of increasing the problem of data sparsity. 'Data sparsity' occurs when an MT system's training data does not cover all the input tokens sufficiently: that is, some input may not appear at all in the MT's training data or may appear but not in statistically useful numbers. Allison et al. (2006) describe data sparsity as the phenomenon of not observing enough word forms in a corpus to model a language accurately. The question is whether NMT systems translate grammar correctly between Arabic and English; if not, this could potentially lead to the production of less fluent translations. The experiments in Chapters Six and Seven set out to resolve the question of what happens when an MT system receives input which includes several inflected words that are not (or are less frequently) found in the training data.

1.3.1.4 Orthography

Arabic script has two types of symbols for writing words: letters and diacritic marks. Its spelling consists of 28 basic letters (it can be extended to 36 by using the Hamza variation) and nine diacritic marks to represent short vowels.

- **Hamza varieties**

Hamza (ء) is the/a glottal stop (') in Arabic; it takes multiple forms in written texts: it can be a single letter when it is written alone (ء)[5] (Habash and Rambow 2007), as in the word-final ماء */ma'an/* 'water' or it can be combined with a letter (character-marker) where it and its carrier become a diacritic marker. It is placed above or below the letter Alif (أ / Â , إ/ Â), as in أمي /umi/ 'Mum' and إقامة /iiqama/ 'accommodation', respectively; above a Waw (ؤ / ŵ), as in كؤوس /kawuws/ 'Cups'; and above an Alif-Maqsura (ئ / ŷ), as in شاطئ /shati/ 'beach'. Furthermore, an Alif-Hamza can be placed in addition to an initial-word position, in mid-word position as in يسأل /yas'al/ 'he asks' or word-final position as in ملجأ /malja/ 'Shelter'. The difficulty is that the insertion of Hamza on its carrier is controlled by a set of complex rules. For this reason, general Arabic texts often include Hamza variations with un-Hamzated (without Hamza) forms, especially the Alif-Hamza (أ /Â /), (إ / Â); thus, they are written as bare-Alef (as in ا) since the meaning is clear from the context (El Kholy and Habash 2010). For example:

(1) أمي تحبني كثيرا / *'umiy tahabani kthyraan* / Mum loves me so much.

(2) امي تحبني كثيرا / *amy tahabuni kthyraan* / Mum loves me so much.

In the above examples, the Hamza-Alif of the word أمي /umi/ 'Mum' in sentence (1) is written as Hamza above Alif (أ), while the word امي /umi/ in sentence (2) is written without the Hamza; the meaning is identical in both cases as the context renders the meaning clear even if the orthography is incorrect as the Hamza marker is omitted. This feature raises the following question in terms of TM retrieval: if one of these versions were given to a translator as a source text but their TM database contained the other version, would the TM matching metrics consider the omission of the Hamza marker as a minor or major difference? The experiment in Chapter Five, section B, investigates this question.

[5] Hamza varieties are highlighted in yellow.

With regard to translating automatically, the omission of the Hamza marker may lead to an increase in both ambiguity (the same form corresponding to multiple words) and sparsity (multiple forms of the same word) since this creates close synonym forms. According to Habash and Sadat (2006), Hamza allows suboptimal orthographic variants of the same word to coexist in the same text.

- **Diacritics**

Another orthographic feature of Arabic is that its script does not have dedicated letters denoting short vowels but uses diacritic marks to represent them – so-called 'altashkil' (تَشْكِيل) or 'ḥarakāt' (حَرَكَات). Diacritics are carried by case endings either above or below the letters; they relate to text vocalisation and are used to add information about the pronunciation and meaning of words that could help resolve the forms' potential lexical and semantic ambiguity (Ameur et al. 2015). In some cases, the lexical meaning of words (i.e. the homographs) can be determined simply by inserting a diacritic mark (Habash 2010). Figure 1.2 (below) shows that a single form, عقد / *eaqad* /, has multiple meanings according to its diacritic mark.

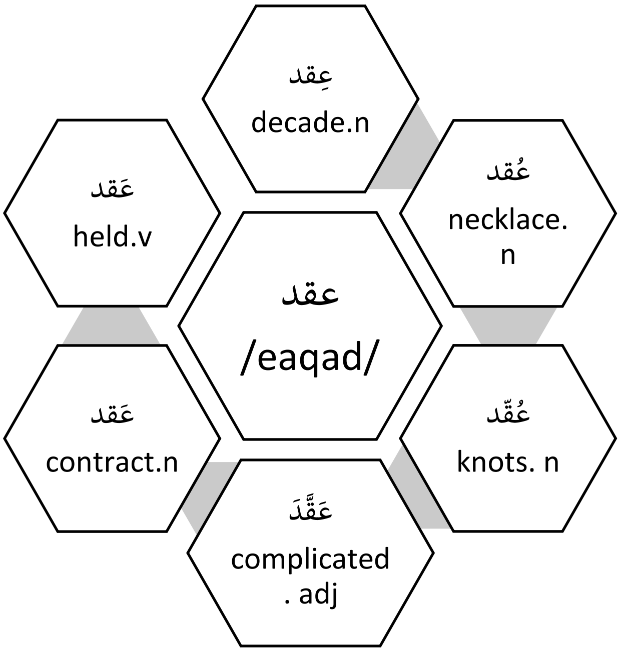


Figure 1.2: Multiple meanings of the word forms عقد / eaqad/

As seen in Figure 1.2, the intended meaning of the base-form عقد depends on its diacritic mark – the meaning changes as the diacritic marks change.

- The form عِقد ‘decade.n’ has a kasrah (كسرة) below.
- The form عُقد ‘necklace’ has a dammah (ضمة) above.
- The form عُقّد ‘knots.n’ has a dammah (ضمة) and shaddah (شدة) above.
- The form عَقَّد َ‘complicated. adj’ has a fatḥah (فتحة), shaddah (شدة) and fatḥah (فتحة) above.
- The form عَقد ‘contract.n’ has fatḥah (فتحة)above.
- The form عَقد ‘held.v’ has fatḥah (فتحة) above.

The issue that arises here is that the optionality of using diacritic marks in Arabic may lead to texts entirely without these markers; nevertheless, the intended meaning can be predicted from the context (Habash and Sadat 2006). Take the example below:

(1)لبست الفتاة العِقد / *labisat alfatat aleiqd* / The girl wore the necklace.

(2) لبست الفتاة العقد / *labisat alfatat aleuqad* / The girl wore the necklace.

In the above examples, the word العِقد / *aleiqd* / in sentence (1) is written with the diacritic mark *kasrah*, while in the sentence (2) the word العقد / *aleuqad* / is written without the mark; the meaning is identical whichever form is used since the context renders the meaning clear even if the orthography is incorrect due to the absence of the diacritic mark. This feature raises the following question in relation to TM retrieval: if one of these versions were given to a translator as a source text but their TM database contained the other version, would the TM matching metrics consider the absence of the diacritic mark a major or a minor difference? The experiment in Chapter Five, section A, is intended to answer this question.

Generally speaking, the linguistic features mentioned above represent significant challenges to the state of the art in both TM retrieval and MT. When Arabic is the source language, TM’s matching mechanism may face difficulties in handling its flexible word order, morphological inflections and omission of Hamza markers. With automatic translation, unless the MT systems construct a verb-initial sentence, produce morphologically correct forms and predict the context in a case of absent diacritic marks, this may lead to the production of less fluent translations from Arabic to English (the difficulties are likewise transferred to the target side when translating from English to Arabic).

1.3.2 Translation memory retrieval

1.3.2.1 Introduction to translation memory

The rise in the use of TM systems in the translation market in the early 1990s led to the establishment of the newly developed field of computer-aided translation (CAT) tools (Hutchins 1999). By the early 1990s, the component had become commercially available, and TM was soon widely accepted by translators. Hence, TM is the core component under scrutiny in this study of translation tools.

TM retrieval involves a process of recalling a set of translation records from a database which are algorithmically calculated to be of potential use in translating an input string (Baldwin 2010). The system was originally designed to support translators, allowing them to reuse the translations of repeated segments without having to re-translate the same text. Once the translation process starts, the TM system presents proposals of translations from its database matches which the translator can accept or modify as they wish, saving them a significant amount of time (Macklovitch and Russell 2000).

TM systems do not work in the same way as MT systems – they do not translate without human intervention. Instead, they perform a retrieval process that recalls relevant (similar) segments that have been previously translated by human translators. One of the potential benefits of the TM help boosting consistency of expression within and between documents (Moorkens 2012).

1.3.2.2 Translation memory database

An important aspect of obtaining useful matching results is the efficient storage of segments in the TM database. This database is a bank of parallel data comprising previous translations that are stored as translation units (TUs) – that is, segments of source language along with their translations. A segment means any meaningful unit: a word, phrase, sentence or even a paragraph. According to Bloodgood and Strauss (2015), how well the TM database of previously translated segments is matched to the texts to be translated is crucial to retrieving useful translations.

There are three main methods of building a TM database. First, during the translation process itself: when new segments are translated, they are automatically stored in the TM along with their translations. Secondly, through the import of a database from either a TM created with the same TM system or from parallel data (a ‘corpus’) available in the format of TMX (Translation Memory eXchange). A TMX file (an XML supported by all CAT tools) can be imported into or exported from any TM system. The corpus is defined here as a collection of naturally occurring examples which are stored on a computer to permit investigation using a special translation tool. A corpus typically contains annotations of meta-information (i.e. indexes attached to segments) that facilitate the efficiency of the matching metrics when retrieving translation segments. Thirdly, a TM can be created with the help of an alignment tool by placing a source text alongside its translation and aligning corresponding source-language (SL) and target-language (TL) segments into translation units (Somers 2003).

This study used the first two types of TM database to address the research objectives. In the experiments in Chapter Four, a corpus was imported to use as a TM database, while in the experiments in Chapter Five, the researcher created his own TM by alignment.

1.3.2.3 Translation threshold

TM systems are often user configurable with a translation threshold and only provide proposals with a similarity higher than or equal to this threshold score. The translation threshold is a mechanism to limit the number of unhelpful proposals that are provided to the translator. Therefore, unlike the suggestions in MT systems, a TM system does not inevitably supply proposals for each input. O'Brien (2007: 196) states that the translation of segments for which there are no or low matches is associated with the heaviest cognitive effort, while the translation of segments for which there are exact matches is associated with the least. 'Cognitive effort' here refers to the amount of brain activity and knowledge needed to accomplish the translation (Krings 2001:179). Translators generally have the ability to set a translation threshold to reduce the number of less useful fuzzy matches if they calculate that reviewing these would be a waste of time.

The translation threshold sets a minimum match level to filter out matches of lower similarity. Many translators prefer to set the threshold somewhere between 70% and 75%. Bloodgood and Strauss (2015) suggest that a 70% value is an ideal threshold if the translator is to avoid excessive cognitive effort. Further, some systems like Trados Studio use 70% as the default fuzzy match threshold, which means that when the degree of match between a source file segment and a TM source is less than 70%, they are dealt with as having less usability and are not displayed to the translator. By filtering the fuzzy matches, the TM similarity metrics only retrieve translation pairs that have computed scores higher than the match threshold; however, the danger is that this could mean that if a match is scored too low, the translator will not see potentially useful information.

This study used fuzzy matches and the 70% translation threshold to address the research objectives: the experiments in Chapter Four used this threshold in the pre-translation function in order to measure the usability of the TM proposals.

1.3.2.4 Similarity metrics

During the retrieval of proposals from the TM database, the measurement of the degree of similarity is based on a comparison between source language texts. The function of the similarity metrics, therefore, is to quantify the usefulness of the TM translations. Hence, Bloodgood and Strauss (2015) stress the need for effective similarity metrics.

One of the most significant functions of TM algorithms is the ability to match a source sentence against the database. When given a new sentence of text to translate, the matching algorithm looks for source language sentences in the TM which are identical (exact match). If an exact match is found, the TM metrics return the target-language version of that match (or matches). As mentioned above, if there is no exact match, the system metrics estimate the degree of similarity, and then display the close match (fuzzy matches), highlighting the differences by means of a fuzzy score. As similarity can be calculated in many ways, the question is how do the matching metrics in TM tools measure the text similarity?

The developers regard the matching algorithms used in TM systems as commercial secrets; nevertheless, it is widely believed that segment retrieval is measured by string similarity metrics based on some variation of edit distance, such as Levenshtein distance (Bloodgood and Strauss 2015; Simard and Fujita 2012). A distance function measures the dissimilarity between two strings of text: identical strings have a distance of (0), while the less similar two strings are, the larger the distance between them. Similarity metrics supply a similarity score, usually presented as a fraction between 0.0 and 1.0, or a percentage ranging from 0% to 100%. To convert a distance (d) into a similarity score, some normalisation (l) constant is used to constrain the distance measured to the range (0-1 or 0-100%). The similarity score is then defined as $1 - d/l$. To ensure that the smallest value of the similarity score is (0), the value of (l) has to be the highest possible value of (d). Using such a normalised metric allows a comparison of the extent of similarity between different string pairs, regardless of whether they are short or long. As various similarity metrics and their normalisation constants work in different ways, this could result in a difference between the similarity scores for a string pair when measured by two different similarity metrics. Thus, the different similarity metrics not only have an influence on the ranking of TM suggestions, but also on whether TM suggestions are displayed at all if constrained by a particular translation threshold (Wolff et al. 2016).

Segment-based edit distance between the source string and the target string is the smallest number of edit operations required to transform one of the strings into the other (Levenshtein 1966). The operations of edit metrics can be classified as either 4-operation edit similarity or 3-operation edit similarity.

1) 4-operation edit similarity

This metric is based on the string distance function, the edit operations as described as

> […] segment equality (segments s_i and t_j are identical), segment deletion (delete segment s_i), segment insertion (insert segment t_j after s_i in string S), and segment substitution (substitute segment s_i for segment t_j). (Baldwin 2009: 203)

Substitutions are in fact the combination of two operations, involving a simultaneous deletion and insertion operation. Dynamic programming (DP) algorithms are used to determine the minimum edit distance between a given string pair, following the 4-operation edit distance formulation of Wagner and Fisher (1974). According to Wolff et al. (2016: 23), although the details of matching metrics of CAT tools are a commercial secret, an informal investigation of a free application (such as OmegaT) indicates that the 4-operation edit similarity metric is used by some TM system implementations.

2) 3-operation edit similarity

This is similar to the 4-operation edit similarity metric, but the underlying distance function does not count substitution as one of its basic operations. A substitution is thus patterned as an insertion and a deletion (two operations), with the third being the identity operation. Baldwin (2009) has evaluated that the 3-operation edit similarity – which is identical to the ‘sequential correspondence’ method that determines the maximum sequential substring match between two strings (Baldwin and Tanaka 2000) – gives the best performance.

Other researchers appear to concur:

> The issue of the word order may be solved by using methods not as extremely tied to the order of characters/words, such as [that] suggested in Baldwin (2009). The suggestion of using three operation edit distance rather than four operation edit distance may be beneficial (Wolff et al 2014: 4404).

Another evaluation method was proposed by Bloodgood and Strauss (2015), whose study compared a set of string-based metrics for TM retrieval: edit distance metrics and matching methods specifically designed for fuzzy matching, such as percent match and n-gram precision, which act on unigrams and longer n-grams. Percent match calculates the percent of unigrams (tokens) in input segments that are found in the source TM, while n-gram precision is inspired by the n-gram precision underlying the BLEU score for MT evaluation. Weighted percent match (WPM) uses inverse document frequency (IDF). IDF is used to give more weight to important words or less weight to morphological variants. The authors found that a weighted n-gram precision measure performs better than the edit distance metrics, according to the judgment of human translators. They also found that a weighted n-gram precision measure retrieves a better-rated match than edit distance metrics, and correlates best with human judgments. The study concludes that it is useful to preserve the local context in fuzzy matches by designing a weighted version of n-gram precision in which translators can set the preferred length of matching spans themselves.

The present study investigates whether the TM metrics can recognise segments including a move operation (re-ordering) operation as highly similar. A ‘move operation’ means that two components in a segment can exchange positions without changing the segment’s meaning, unlike a substitution operation that is recorded as two edit operations, simultaneously deleting and inserting components (Wolff et al., 2016). Shapira and Storer (2002,), whose studies were among the first to discuss the problem of edit distance in relation to substring moves, considered the issue to be computationally complex. Muthukrishnan and Sahinalp (2000, 2002) discussed the problem of edit distance with block edit operations such as a move operation, they proposed a parsing algorithm “approximate nearest neighbour” search for calculating edit distance blocks; it is considered as a sequence comparison with block operations. The same factor was developed by Cormode and Muthukrishnan (2007), when considering the edit distance with move operation only, they embed the strings into the L_1 vector space, then use L_1 distance between the two strings vectors to get an approximate result to the block edit distance. Meanwhile, Baldwin and Tanaka (2000), in their study on the effect of word order on translation retrieval, state that texts that maintain their segment order will supply closer-matching translations than those that include the same words but in a different order. The move operation scenario can be seen in the flexible word order of Arabic. Arabic's flexibility means that the sense of a segment can be expressed in various ways by using the same surface words but in a different order (Elming 2008), which means a move operation is required.

Overall, it appears that the edit distance metric is not efficient enough in measuring similarity where strings are not exactly the same; it fails when required to recognise two segments which might have the same meaning but a different word order. According to Šoštarić (2018), the edit distance metrics have several disadvantages when applied to the measurement of strings: they do not allow for changes in word order and also face problems in dealing with inflectional phenomena. As a consequence, although the edit distance metric performs efficiently with an exact string sequence, it does not perform well with segments that are highly similar but whose sequence is not exactly the same.

1.3.2.5 Matching types

'Matching' is a process whereby the input is compared for similarity with the TM sources, and the corresponding TM translations are then retrieved. TM systems display different types of matches according to the similarity of the segments. If an identical source segment is found in the database, the TM algorithms offer translation(s) which are aligned with the source as a 100% or 'exact match'. An exact TM match is a character-by-character match between a source language segment and a TM source. If the source segments do not match any of the TM sources precisely, the system typically performs a 'fuzzy search' (a technique it uses to find matches in the database that may be less than the exact match). The fuzzy search retrieves segments that are similar but not identical, expressing the match as a percentage ranging from 0% to 99%. If the dissimilarity between the segments is minor (i.e. the TM segment differs only slightly from the input segment), the matching score is high; if, on the other hand, the dissimilarity is significant, the matching score is lower (for more information about fuzzy match bands, see Chapter Three, section 3.1.7.1).

Some TM systems permit the translator to view more than one fuzzy match simultaneously, arranged in descending order of match score. Translators then have the option to modify translation suggestions by editing one or other of the proposals into the desired accurate translation (Bowker 2002).

Hence, the aim of this research is to investigate the performance of current TM similarity metrics when retrieving Arabic source segments that include complex linguistic features, such as differences in word order, inflectional affixes and the omission of the Hamza marker, and to determine to what extent these features affect the fuzzy matching scores.

1.3.2.6 Integration

In the past, TM and MT were used separately (independently) in translation workflows. In recent years, however, TM systems – although they can still be used on their own – have begun to offer the functionality of MT integration. If these technologies are combined, the translator is offered, in addition to TM retrievals, MT suggestions that help them come as close as possible to producing a perfect translation. Using MT suggestions in a static way means that the MT providers offer a translation for the entire sentence. According to Simard and Isabelle (2009), the baseline approach to TM-MT integration is to use an MT suggestion to translate a query sentence when a sufficiently similar translation cannot be found in the TM database.

Some studies have explored the use of an MT system as a back-up in cases where no highly similar source sentence can be found in the TM database. The back-up approach extracts suggestions from the MT systems which are then fed into the TM system. Federico et al. (2012) state that although TM systems offer some advantages over MT systems in terms of retrieving translations for previously translated segments, the TM metrics can sometimes fail to retrieve matches above a certain translation threshold. In such cases, the MT system is often used as a back-up solution.

The viability of MT deployment for TM users can be considered in different ways. Some CAT tools, like Trados Studio and Memsource, develop MT systems for their own use in their language service businesses and for sale to translators. Other tools, such as DVX and memoQ, are dependent on connecting to a specialist provider's MT service. The typical deployment of MT is through an application programming interface (API) plug-in that makes MT systems' output available within the CAT translation environment (Hu et al. 2018). Thus, the CAT environment combines resources from MT, TM and terminology management tools to produce translations.

A dialogue box in the project settings allows the translator to configure the MT plugins – for example, which MT provider to use and how much. Typically, it looks up every segment in the translation memories of the project and inserts the best match (the exact match and matches of a certain translation threshold and above). If there are no appropriate TM matches, an MT suggestion is generated and inserted into the editor so that the translator can use the MT output as a first gist translation before reviewing it (Olohan 2021).

Some popular CAT tools that use MT plugins are described below:

- **DVX**

DVX is one of the CAT tools that deploys the direct integration of MT with TM proposals. Unless the translation editor displays exact and fuzzy matches from the TM database, segments from MT suggestions are displayed in the translation editor. MT engines, like Google Translate, Microsoft Translator and MyMemory, are supported by DVX and are available for use during translation (Déjà Vu X3).[6]

- **memoQ**

The memoQ application has been able to connect with MT engines through plug-ins since 2018. The MT providers that are supported by memoQ include Google Translate, Microsoft Translator and DeepL Translate (memoQ manual).[7]

- **Memsource**

Memsource offers MT integration by either accessing Memsource Translate or a third-party MT provider. Memsource Translate uses artificial intelligence (AI) to automatically select the optimal free MT engine, such as Google Translate, Microsoft Translator or Amazon Translate, to translate the text according to the language pair, domain and type of content. If Memsource Translate is not used, third-party MT engines can be accessed via an API (Memsource).[8]

Memsource produces a quarterly report, based on Memsource Translate data, on the performance of different MT engines across various language pairs and domains, providing a high-level analysis of their quality. First, it analyses a document to determine its domain; then, it looks for the optimal MT engine based on the past performance of each engine in a given domain, source language and target language. The most recent report concludes that no single

[6] Machine Translation – Atril Solutions (zendesk.com)
[7] Edit machine translation settings (memoq.com)
[8] https://help.memsource.com/hc/en-us/articles/360012620459-Machine-Translation-Overview

MT engine can always provide optimal quality as this not only varies according to the language pair but also depends on the domain (Memsource MT report 2020).[9]

- **Trados Studio**

Trados Studio has developed its own MT system – Language Weaver – for use in its language services. Language Weaver is an adaptable NMT platform that can be used for translating content. Segments not leveraged from the TM database can be automatically translated, but the translator has the option to accept or amend these translations if necessary. In addition, Trados Studio can connect to a number of MT providers, including Google Translate, Microsoft Translator and ModernMT (Trados Studio website).[10]

It is worth noting that pre-translating with the MT option activated can offer perfect translations or increase pre-translation results in terms of both quantity and quality. However, these services often require the translator to sign up for a paid account or purchase a licence for the software if they wish to install and use the engine.

1.3.2.7 TM and fuzzy match repair

Many contemporary CAT tools now offer a functionality known as ‘fuzzy match repair’. It is based on a structured cooperation between TM standard resources and MT that often makes use of essentially the same mechanics of semantic substitution. If a source sentence differs from a TM match in a single subsegment, the fuzzy match repair mechanism tries to ‘repair’ the portion of the segment that does not match the source segment: it gives the system the data to generate the correct translation of the source sentence, even though that translation is not present in the TM.

The functionality of fuzzy repair is available in some of the CAT tools, including DeepMiner at DVX, MatchPatch at memoQ, and UpLIFT at Trados Studio.

[9] https://go.memsource.com/machine-translation-report
[10] What is Machine Translation? (trados.com)

- **DeepMiner in DVX**

DeepMiner statistical extraction is a feature of DVX that extracts a greater amount of information from the standard resources (TM database, Termbase, Lexicon), along with the MT results, to fill in parts of a non-translated segment. It first carries out sophisticated cross-analyses of those databases on the fly in order to 'mine' the translations of terms and phrases embedded in them. If the system identifies the translated equivalent of the non-matching part of the segment, it uses conventional fuzzy matches and/or these mined terms and phrases to create fuzzy match repairs, offering improved translations. It then uses an MT provider to enter suggestions where no other matches are found.[11]

- **MatchPatching in memoQ**

MatchPatching is a memoQ feature that is used to 'repair' the translation from standard resources (TM database, Termbase) and MT. When setting up MatchPatching with TM, the system tries to 'patch' matches automatically, using not only translation memory fragments but also suitable terminology hits. memoQ looks up differences in fuzzy matches. Usually it patches fragments with one or two differences because MatchPatch needs high match scores. When a fragment is found in the text, it will appear on the list of suggestions. Patching with MT occurs when MatchPatching cannot patch a fuzzy match using the standard resource. memoQ shows a patched match with an exclamation mark before the match rate, and applies a penalty even though a patched match may be perfect.[12]

- **UpLIFT in Trados Studio**

Trados Studio uses the UpLIFT Fragment Recall technology to find matches at the subsegment level that are part of a previously saved translation unit. The fragments are matched accurately and retrieved automatically from the TM when no match has been found, making it simpler to maximise existing resources. With UpLIFT, fuzzy matches can be repaired by automatically performing a set of amendments such as deletion, insertion, movement and punctuation

[11] https://atrilsolutions.zendesk.com/hc/en-us/articles/205540701-Using-D%C3%A9j%C3%A0-Vu-X3-A-Tutorial
[12] Translation results list (memoq.com)

changes. Furthermore, the feature can present the origin of repair: a TM, Termbase or MT.[13] The UpLIFT feature was designed by Flanagan (2014) to improve the recall of subsegments from a TM, and was integrated into SDL Trados Studio 2017. It was tested as part of this research and the results can be seen in in Chapter Four (for more details, see section 4.1.2.5).

1.3.3 Machine translation development

1.3.3.1 Introduction to machine translation

Machine translation is a means of transferring texts automatically from a source language into a target language without human intervention. The history of MT dates back around 70 years to when the first full-time researcher in the field, Yehoshua Bar-Hillel, began his research at MIT (1951), and an influential demonstration of MT was performed by Georgetown University and IBM in 1954 (IBM). In the Georgetown-IBM experiment, 60 Russian statements were translated into English, marking a milestone in MT. However, the ALPAC (Automatic Language Processing Advisory Committee) report of 1966 (Hutchins 1998), which concluded that MT was slower, less accurate and more expensive than human translation research, at least in the United States, was almost completely abandoned except for some companies such as Systran that built a system for the United States Air Force in 1970. In Canada, France, USSR and Germany, however, research continued. The METEO System for example, developed at the Université de Montréal, Canada in 1977, was designed for the translation of the weather forecasts from English to French. Significant uptake of MT did not begin until the 1980s.

MT developments can be categorised according to their core methodology into two main paradigms: rule-based machine translation approaches that rely on linguistic rules, including direct (also referred to as dictionary) MT, transfer-based MT and interlingual MT; and the corpus-based (also referred to as data-driven) MT approach. The latter was developed as an alternative approach for MT systems in the attempt to overcome the difficulties faced by rule-based machine translation. The data-driven approach, which relies on large bilingual parallel corpora to obtain translation knowledge, is divided into sub-approaches: example-based MT;

[13] https://www.trados.com/blog/how-to-use-uplift-fragment-recall-and-fuzzy-repair-in-sdl-trados-studio.html#fragmentrecall

statistical MT; hybrid MT; and, since 2015, neural machine translation. The different approaches will be discussed below in the order of their historical appearance.

1.3.3.2 Rule-based machine translation

RBMT depends on hand-coded linguistic rules for the source and target languages. The system applies a large collection of grammatical rules and translation dictionaries – lexicons that contain morphological and even syntactic information, as well as semantics – in three different phases: analysis, transfer and generation. The rules are deployed by a combination of linguists and computer scientists. As mentioned above, the RBMT's architecture can be classified as three sub-approaches: the direct approach, transfer approach and interlingua approach.

Direct translation is the oldest MT approach and depends on the use of a large and comprehensive lexicon – this is the reason why it is also called the dictionary-based approach. It requires only a little syntactic and semantic analysis of the source; words in the source language are translated directly into the target language (i.e. word-for-word translation) (Hutchins and Somers 1992:72).

The main drawback of word-for-word translation – the fact that it may not be semantically comprehensible or convey the intended meaning of the translated text – led to the development of the transfer approach, in which the translation process is usually carried out in the different stages of analysis, transfer and generation. In stage one (analysis), a parser is used to provide a syntactic representation of the source language sentence. In the next stage (transfer), a set of linguistic rules specific to the source and target languages transforms the syntactic representation of the source into an equivalent representation in the target language. In the final stage (generation), a target language morphological analyser is used to generate the target text. However, the main weakness is that it is language-pair-dependent, which makes adding a new language pair very expensive. In other words, linguistic rules are needed for a new source-language analysis as well as for a new target-language generation (Nirenburg and Wilks 2000).

The interlingua approach was developed to resolve the issue of the complexity of the transfer approach. It uses an auxiliary language based on an intermediate representation of the meaning of the source text to form the basis of the target-text generation – that is, the source language text is transformed into a neutral inter-language, independent of any natural language, and the target language text is generated from the representation. The interlingua-based translation process requires only two monolingual elements: first, an analysis of the source text into a universal language-independent representation of its meaning; next, the generation of translated texts from the representation of the source's meaning using the lexical units and syntactic constructions of the target language. However, the difficulty of the interlingual approach lies in the building of language-neutral meaning representations (Dorr et al 2004)).

1.3.3.3 Example-based machine translation

Example-based machine translation (EBMT) systems are trained from a large collection of parallel corpora, in which a sentence is translated by analogy (i.e. text similarity). The EBMT model, introduced in the 1980s (Nagao 1984), consists of three basic steps: the input source is first deconstructed into short fragments to match example fragments in the available corpora; it then looks for translational equivalence in the target language; and, in the final step, the translated fragments are recombined into the target sentence (Hutchins 2005).

Despite the essential distinction between the systems of MT and TM, EBMT and the TM component have in common the re-use of previous source-and-target language translation pairs and the method of matching. The difference lies in the fact that EBMT automatically extracts the corresponding translations and combines them into the target output (the machine controls the translation), while TM suggests highly similar matches and the translator selects the best match – modifying it if necessary – for the production of the desired translation (the translator controls the translation) (Somers and Diaz 2004).

1.3.3.4 Statistical machine translation

Statistical Machine Translation (SMT) is a paradigm that generates translations based on the theory of probability, whose parameters are learned automatically from the analysis of a very large bilingual dataset. A probabilistic model entails collecting statistics about events and calculating the probability distribution (Brown et al. 1990, 1993). This means that the probability of something occurring depends on different variables likely to impact the event.

The difference between SMT and RBMT lies in the acquisition of translation knowledge: the RBMT approach always requires the manual development of linguistic rules, while the SMT system pursues a corpus-driven approach to acquiring translation knowledge.

The SMT paradigm is usually implemented in three components: a bilingual translation model; a reordering model; and a monolingual target-language model. First, the translation model uses the frequency of phrases appearing in a very large aligned bilingual training corpus to find the proper source/target translation combination; the more frequently the SL phrase is repeated in a corpus in parallel with a specific TL string, the more probable it is that the target translation is correct. The reordering model then provides probabilities for reordering the translated phrases relative to their original position in the target translation. Finally, the language model renders the translated text fluent in the target language (Koehn et al. 2003). Because they derive their information from corpora, language models and translation models are the most significant models in SMT. The language model is based on an n-gram architecture (n-grams are all the combinations of adjacent words or characters of length n that can be found in the source text). Translation models can take various alignment forms depending on the language units used: word-based, phrase-based or syntax-based models. Word-based models estimate sentence translation probability based on using words as the atomic unit, while phrase-based models calculate sentence translation probability based on using phrases as the atomic unit (Koehn 2009).

1.3.3.5 Hybrid machine translation

The motivation for developing the hybrid machine translation (HMT) approach stems from the failure of any single technique to produce a high level of accuracy. A hybrid architecture uses multiple MT approaches in a single system in order to integrate the advantages of different approaches. This system-combination approach takes one or more outputs from each system and then merges the results at a word, phrase or sentence level: for example, the combination of rules post-processed by statistics or guided by rules. The former is where texts are translated using a rule-based method and statistics are later used to adjust the output; the latter applies rules to pre-process the input in an attempt to better guide the statistical model (Chan 2014). However, according to Xuan et al. (2012), the hybrid MT approach faces similar difficulties to that of SMT, especially in terms of the requirement of a large bilingual database.

1.3.3.6 Neural machine translation

Neural MT (NMT) has become the mainstream architecture for MT since 2016. The neural network approach uses a single model – the encoder-decoder architecture – to train directly on the source and target text, no longer requiring the three-model systems used in the SMT approach (the encoder is a model used to encode a given text into a continuous vector; the decoder is used to transform the state vector into the target language). In principle, the ability of the encoder-decoder architecture to encode the source text into a vector representation also gives the model the ability to produce different decoding systems to translate into different languages.

The earliest significant attempts to use neural networks in MT were made by Kalchbrenner and Blunsom (2013). Their study, which laid the foundations for the use of encoder-decoder architectures in MT, used a convolutional encoder (i.e. linear sequential circuits) to encode a source input, then a decoder to generate the translation through a recurrent neural network (RNN). The model was further developed by Sutskever et al. (2014) and Cho et al. (2014). The model uses one Long Short-Term Memory (LSTM) to read the input sequence, one timestep at a time, to obtain a large fixed dimensional vector representation, and then another deep LSTM to decode the target (output) sequence from the vector. LSTMs are a very special kind of recurrent neural network which is capable of learning long-term dependencies. Remembering information for long periods of time is practically their basic behaviour. This end to-end NMT typically consists of two recurrent network processes in which a sentence is treated as a sequence of words or sometimes characters: an encoder maps the input sequence of variable length to a point in a continuous vector representation (i.e. a numerical summary of an input sequence), resulting in a fixed-length vector or so-called 'context vector' (i.e. the fixed-length vector means a fixed-size representation of input sequence where the output must be the same length). A decoder then generates a target sequence, again of variable length, starting from the context vector. The transformation of source sentences into a sequence of vector representations and translation generations are learned and performed using word embedding, a method of transferring words from a vocabulary into a vector representation space where each word is represented in hundreds of dimensions (e.g., 0,0,0,0,1,0,0,0). This type of word representation allows words of similar meaning to have a similar representation (Cho et al. 2014).

However, it was observed that although the pure encoder–decoder model performed well with short sentences, it was inefficient when handling long sentences. The difficulty stems from the internal fixed-length representation that must be used to decode each word in the output sequence. In other words, it is difficult to use a fixed-size representation to capture all the semantic details of long sentences, especially those that are longer than the sentences in the training corpus. Bahdanau et al. (2014) proposed a remedy for this issue: extending the basic encoder-decoder network model by incorporating an attention mechanism which learns to align and translate jointly (i.e. it aligns the words from source to target and translates them at the same time). The attention encoder-decoder model considers a sequence of vector representations of the source sentence generated by a bi-directional RNN encoder as input, and then learns to align and translate simultaneously by reading the vector representations during translation with a decoder (i.e. a bidirectional RNN reads the source sequence in both forward and backward directions). This explains interactive translation prediction – using interactive typing (for example, Lilt translates before the whole sentence is read) (Santy et al. 2019). An attention mechanism is designed to predict the alignment of a target word in relation to the context vectors (source words). With this approach, the attention encoder-decoder model does not encode a whole input sentence into a single fixed-length vector; it encodes the input into a sequence of vectors and then chooses a subset of these vectors when the decoder produces the target sentence, which helps it to cope effectively with long input sequences. In other words, each word is encoded through a vector, then the words' vectors (i.e. the vectors of multiple words) are gradually combined to give a vector representation of the whole sentence, where each vector represents the meaning of all the words read so far (the encoder). Once the entire sentence is read, the decoder begins by producing the translated words one at a time, each time focusing on a different part of the input sentence to gather the semantic details required to produce the next output word (synced review blog).[14]

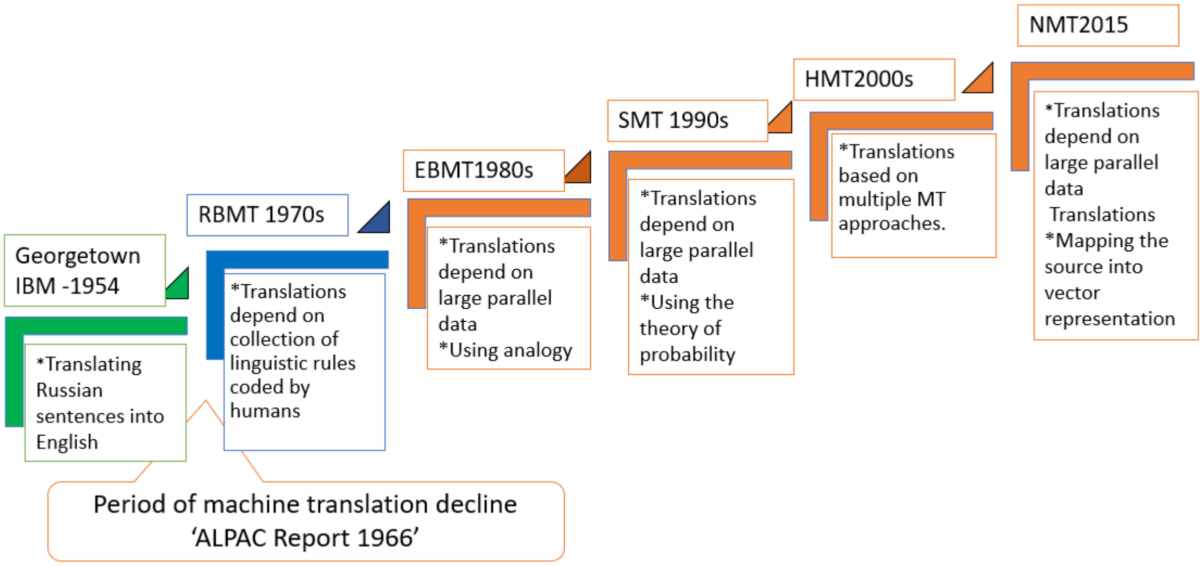
The techniques of vector representation and the attention-based mechanism worked well in improving MT output; however, their performance appeared to be conditional on the provision of large amounts of parallel data, and this is not available for low-resource languages. This led to the suggestion of a complementary solution: to supplement the parallel data by using monolingual data. Sennrich et al. (2015) attempted to enhance the decoder network model by incorporating monolingual target sentences into the training data so as to boost translation fluency. Their study introduced back translation as a method of leveraging the target monolingual data so that an automatic translation system can be initially trained on translating from the target to the source language on the available parallel data. Using target-side

[14] A Brief Overview of Attention Mechanism | by Synced | SyncedReview | Medium

monolingual data as additional parallel training data can significantly enhance the decoder model.

Another technique for overcoming a lack of parallel data in some languages is a multitask learning system in which MT learning occurs without parallel data. Johnson et al. (2017) introduced a multilingual approach to train a single model for translating between multiple language pairs. 'Zero-shot translation' (translating between two language pairs for which no parallel data was applied at the time of training) uses a shared vocabulary and a special token in the input sentence to specify the target language. This approach, first demonstrated in 2015, reported promising results close to the state of the art.

Figure 1.3 (below) summarises the timeline of MT evolution.



Georgetown IBM -1954
*Translating Russian sentences into English
RBMT 1970s
*Translations depend on collection of linguistic rules coded by humans
EBMT1980s
*Translations depend on large parallel data
*Using analogy
SMT 1990s
*Translations depend on large parallel data
*Using the theory of probability
HMT2000s
*Translations based on multiple MT approaches.
NMT2015
*Translations depend on large parallel data
Translations
*Mapping the source into vector representation
Period of machine translation decline
'ALPAC Report 1966'

Figure 1.3: Timeline of MT evolution, (Maučec and Donaj 2019).

Due to its success, the encoder-decoder RNN architecture has become the heart of NMT architecture. The neural architecture contains different models, such as single fixed-length presentation and the encoder-decoder with attention. Furthermore, different resources and sizes of training data can potentially be applied across the NMT systems. Hence, this study has focused its evaluation of MT on an investigation into the performance of a set of NMT systems.

MT systems have been improved in recent years by switching to the NMT model. Moreover, the translation process has become more focused on human-computer interaction due to two key developments: interactivity and adaptivity.

An interactive MT system attempts to predict – autocomplete – the translation the human user is about to type. Whenever the prediction is incorrect and the user changes it, a new prediction is offered until the MT suggestion matches the user's expectations. Thus, rather than post-editing the MT output, the translator sees the translations as suggestions they can either choose to use or reject. Meanwhile, an adaptive MT system learns from corrections on the fly and is continuously trained – that is, it learns from the translator's edits of its proposals (Daems et al. 2019). The Lilt tool, which was tested as part of this study, offers the translator interactive MT and adaptive MT.

1.3.4 Using MT for fuzzy match repair

A study by Ortega et al. (2016) has implemented an editing TM matches approach (or fuzzy match repair) that focuses on correcting TM matches using SMT systems. The fuzzy repair feature is a technique that automatically edits high-scoring fuzzy matches. In the automatic fuzzy match repair technique, the translator is offered a choice from a set of repaired translation proposals that comes from a translation unit whose source segment is similar to the segment to be translated. A further study also used this approach and reaffirmed that a fuzzy match repair technique can be beneficial for the quality of the TM output (Bulté et al. 2018).

Fuzzy match repair is gaining greater influence among modern tools as it is regarded as a reliable method of repairing the TM suggestions by using the system's own resources or MT proposals, providing the translator with a more useful translation suggestion and reducing post-editing effort. Unfortunately, details about exactly how the fuzzy repair methods work are not available.

1.3.5 Translation tools evaluation criteria

An evaluation process is essential for measuring the progress in translation tools, and also for providing a way of distinguishing between competing systems.

1.3.4.1 Translation memory evaluation

TM measurement allows the user to retrieve segments from the database which partially match the input. According to Whyman and Somers (1999: 1270), as the functionality of TM systems can be likened to information retrieval software, its performance measures (recall and precision) can be used in the evaluation of TM performance. The definition of these terms depends on what is being measured: in the context of TM, 'recall' measures how many of the available matches are retrieved by TM systems, while 'precision' measures how useful or accurate the actual matches and translation suggestions are, however they are measured (Flanagan 2015).

In this study, 'recall' refers to fuzzy matches of input segments with segments retrieved from the TM database and 'precision' refers to the similarity of segments retrieved from all the valid segments in the database. One of the research hypotheses is that there is a possibility of deficiencies in the TM algorithm systems when faced with specific linguistic issues, causing poor recall but higher precision. As a result, the subsequent low fuzzy matching may have a negative effect on the suggestions offered by the TM, meaning that the translator would not see potentially useful proposals.

1.3.4.2 Machine translation evaluation

The quality of MT systems' output can be evaluated either manually or automatically.

- **Manual evaluation**

Manual evaluation entails judging the raw MT output using properly qualified evaluators – experts in translation or linguistics – to rate its adequacy and fluency according to an agreed scale (White et al. 1994). An adequate translation is a translation that preserves the meaning of the input and does not add, lose or distort any information. A fluent translation, on the other hand, is a text in the target language that is both grammatically correct and natural. Adequacy and fluency are generally considered the two most desirable features for a correct translation; the human evaluator will measure these attributes separately in each of the MT translations, but they are occasionally averaged to give the MT output a single numerical value (Callison-Burch et al. 2007).

- **Automatic evaluation**

Automatic metrics are measures that are employed to calculate the correlation of a candidate translation (the output of the MT system) with one or more gold standard translations that have been produced by a human translator and are considered benchmarks for assessing MT quality. The use of metrics to evaluate the improvement of an individual MT system during its development, as well as to compare different MT systems, depends on the availability of such gold standard translations, which are used as reference translations. This means that a metric should give a high score to machine-translated texts that translators score highly, and a poor score to those that human evaluators award low scores (White 2003). The most commonly used MT evaluation metrics are BLEU, METEO, TER, and recently LEPOR.

1.4 Research question

The issues raised in the above introduction to the background of the research topic lead to the main research question: which translation tool (TM or MT) can best handle Arabic linguistic features when translating between Arabic and English? It is necessary to mention that the tools' usability has been assessed according to its function.

1.5 Thesis structure

The thesis is organised into eight chapters. The current chapter introduces the research: it gives a statement of the problem the study sets out to address, summarises the background to translation technologies and to the Arabic language, and presents the research questions.

Chapter Two reviews the previous findings in the literature on Arabic<>English translation that are of relevance to this research.

Chapter Three details the methodology chosen to measure the performance of the translation tools: its first section presents the method used to evaluate TM retrieval, while the second section describes the method employed to evaluate MT systems.

Chapters Four and Five describe the experiments used to evaluate the TM systems. Chapter Four studies the retrieval of segments which include different syntactic structures: section A investigates the retrieval of segments that include a move operation, while section B compares the retrieval of segments which include a move operation with the retrieval of segments containing a one-edit operation. Chapter Five studies the retrieval of segments which include a morphological feature: section A investigates the retrieval of segments which include an inflectional affix, while section B describes the retrieval of segments which omit the Hamza marker.

Chapter Six is a transitional chapter describing the testing of a neural machine system using the same data as used in TM retrieval.

Chapter Seven presents an analysis of the quality of the output of multiple neural MT systems using adequacy and fluency judgements in addition to automatic metrics.

Chapter Eight concludes the thesis with a discussion of the research findings. It highlights the study's contribution to the field, identifies its limitations, and suggests ways in which it could be extended in future.

CHAPTER TWO

RELATED RESEARCH

2.0 Literature review

In order to contextualise the scope of the current investigation, the following chapter reviews the relevant previous evaluations of TM and MT performance. It is divided into two main sections: section 2.1 focuses on studies evaluating the TM component, while section 2.2 discusses research that uses adequacy and fluency approaches to investigate the performance of NMT.

2.1 Translation memory

2.1.1 Translation memory retrieval

In the process of retrieving suggested translations from the database, the TM system compares source texts by measuring the string similarity using some variation of the edit distance metric. The degree of similarity is determined by fuzzy matching algorithms – when the TM fails to find an exact match between the segments to be translated and the TM sources, the fuzzy matching algorithm helps find any matches that are above the match threshold. The fuzzy matching is determined by comparing strings of characters, meaning that the TM matching metrics use a string comparison of the input segment and the TM source segments (the segments that the translator is dealing with) (Reinke 2013).

The advantage of using TM systems is that they give translators the ability to reuse previously translated segments through the retrieval of close as well as exact matches. However, the weakness of the matching metrics, which are based on the similarity of character strings (formally known as the Levenshtein distance method – for more details, see section 1.4.2.4), reduces the retrieval of close matches. In response to this weakness, many researchers have pointed to the need for similarity measurements that go beyond character-string comparisons. Macklovitch and Russell (2000), for example, maintain that TM systems are limited by the rudimentary techniques employed for approximate matching. They cite Planas and Furuse (1999: 331), who claim that unless the TM matching metric is able to do morphological analysis, it will face difficulties in recognising that sentence (3) in the sequence below is more similar to sentence (1) than is sentence (2).

(1) The wild child is destroying his new toy.

(2) The wild chief is destroying his new tool.

(3) The wild children are destroying their new toy.

These difficulties arise because the TM matching metrics implement edit distance. In this case, the metrics will measure that (1) and (2) have greater similarity as they differ in only four characters, whereas (1) and (3) differ in nine characters. Macklovitch and Russell (2000) explain that using linguistic information, such as named-entity recognition and morphological processing, could improve the TM systems' matching processes. Wang (2014:62,63) illustrates the performance of the Levenshtein distance method in non-Western writing systems such as Chinese, where the meaning is expressed through a combination of characters, word order and contextual information. For example, '微软视窗' and '微软窗口' are Chinese translations of 'Microsoft Windows'. They both convey exactly the same meaning and only differ in one character; however, their fuzzy matching score, calculated using the Levenshtein distance method, is only 50%. As a result, these features hinder the efficiency of string similarity metrics.

Somers (2003:39) highlights a different drawback of the TM matching metric: the segments (4) and (6) in the example below would be retrieved in a higher match than (5).

(4) Select 'Symbol' in the Insert menu

(5) Select 'Symbol' in the Insert menu to enter character from the symbol set

(6) Select 'Paste' in the Edit menu

This retrieval is due to the fact that (4) and (6) differ in only two words, while (5) has eight extra words. As the matching metrics measure only differences, not similarities, the TM system ignores the fact that segment (5) matches (4) better since it includes the entire segment of (4). Somers concludes that the matching metrics need more advanced techniques that incorporate linguistic knowledge, such as inflection paradigms, synonyms and grammatical alterations, if TM performance is to be improved.

Mitkov (2005) points to another drawback: the metrics of TM systems perform matching at the sentence level but not at a sub-sentential level. The author gives the following examples:

(7) Select 'Shut down' from the menu and click on 'Shut down'

(8) Select 'Shut down' from the menu

(9) Click on ‘Shut down’

If the translator is to translate either the segment or clause (8) or (9), a sufficiently high match would not be proposed due to the inability of standard TM metrics to identify sub-segment matching.

In terms of syntactic structure, the metrics of TM systems may not be able to match sentences that are semantically equivalent but expressed in a different word order. For example, the pair of sentences ‘Microsoft developed Windows XP’ and ‘Windows XP was developed by Microsoft’ cannot be a high match because their similarity score is 43%, which is far below the common translation threshold (70%), as Mitkov and Corpas (2008) (cited in Wang 2014: 61) point out. Likewise, Šoštarić (2018: 28) holds that edit distance metrics have several disadvantages when applied to the measurement of strings because they do not allow for changes in the word order, while Gow (2003:27) states that Trados Studio provides the segments ‘*Prendre des mesures de dotation et de classification*’ and ‘*Connaissance des techniques de rédaction et de révision*’ a match score of 56%. This is because half of the word forms are the same and they occupy exactly the same position, even though the words in common are only function words.

Furthermore, Chatzitheodorou (2015: 26) states that the English-Italian sentences pairs (below) share the same meaning, although they do not share the same lexical items:

(10) Press ‘Cancel’ to make the cancellation of your personal information.
(11ST-EN) Press ‘Cancel’ to cancel your personal information.
(11T- IT) *Premere ‘Cancel’ per cancellare i propri dati personali*.

He explains that the word-based string edit distance between sentences (10) and (11ST) is 70%, due to their syntax. As this is a relatively low score, translators may not be offered this as a match.

Gupta et al. (2016b) also state that because the retrieval process is limited to edit distance-based measures that operate according to surface-form matching, the TM metrics are unable to identify the semantic similarity between the following two segments: ‘I would like to congratulate the rapporteur’ and ‘I wish to congratulate the rapporteur’. Even though the two segments express the same meaning, the TM matching metrics based on Levenshtein edit distance award them a similarity percentage of only 71%.

In terms of experimental studies of TM systems, this review of the literature reveals that, to date, there has not been any that directly addresses the way TM retrieval deals with free word order or inflection words; however, some studies have investigated TM performance in relation to Arabic-English translation. Quaranta (2011) may have been the first to address the performance of TM in Arabic<>English translation in a study that investigated the potential difficulties that users of SDL Trados (2007) face in the translation of Arabic texts. She found that linguistic differences, such as words with prefixes, suffixes and infixes, led to decreasing fuzzy matching scores, and concluded that a morphological analysis tool is necessary to overcome the problems deriving from Arabic's complicated morphology. Meanwhile, Thawabteh (2013: 83) has addressed the applicability of TM retrieval of segments differing only in the Hamza marker (character marker) by evaluating the difficulties the MA students faced when translating a text from Arabic into English. The researcher gives the following example:

(1)

وانجبت القدس العديد من الكتاب و الشعراء

/ wainjabat alquds aledyd min alkitab walshueara /

Several writers and poets were born in Jerusalem!'

(2)

وأنجبت القدس العديد من الكتاب و الشعراء

/wainjabat alquds aledyd min alkitab walshueara /

Several writers and poets were born in Jerusalem!

In the first sentence, in the word انجبت , Alif is written without Hamza (bare Alif), while in the second, the word أنجبت includes the Alif-Hamza (Hamza above Alif). The examples above, although semantically identical, provided an 86% similarity. In terms of diacritic marks, Thawabteh gives the following examples:

(3a) ماأجمل القدس! /*maajml alquds* / How beautiful Jerusalem is!

(3b) ماأجملَ القدس! /*maajml alquds* / How beautiful Jerusalem is!

(3c) ماأجملُ القدس؟ /*maajml alquds* / How beautiful Jerusalem is?

The examples above are highly similar in terms of orthography, but syntactically different, thus produce different semantic meanings. Although the word ماأجمل in version (3a) does not include any diacritic mark while version (3b) includes the diacritic mark *fatha* (َ), the meaning is the same, however version (3c) includes the diacritic mark *ḍamma* (ُ), and the meaning is the different. In example 3b, the phrase of ماأجملَ with the fatha is used to create an exclamation, so it expresses the meaning 'the most beautiful'. In contrast, in example 3c, the phrase ماأجملُ with the *ḍamma* is used to create a question, it seeks more details. Though semantically different, 3a and 3b receive a similarity score of 75% and 3b and 3c 84%. The study concludes that including or omitting the Hamza character has no effect on translation retrieval since there are no semantic differences. Nevertheless, the inserting or absence of diacritic marks is important and may have a deleterious effect on meaning. A further study by Alanazi (2019) highlights that one of the potential difficulties CAT tools face when dealing with Arabic-English translation is the spelling output for texts that include Hamza variations and diacritic marks.

Turning to the retrieval of segments requiring minor editing in different language pairs, Wolff et. al (2014) have proved that a similarity metric based on edit distance is likely to miss several useful suggestions. They analysed two linguistically unrelated language pairs (English to French and English to Hungarian) in two different translation memories with very different properties. Their study investigated which useful suggestions would not be selected through the source text similarity, and found that the largest category of missed opportunities was composed of segments that were orthographically different but semantically similar, such as sentences which included synonyms, paraphrases, active/passive variations and abbreviations. In terms of the active/passive variation, the missing suggestion was due to changes in the word order.

More recently, Baquero and Mitkov (2017) performed an experimental investigation into the detection of similarities in sentences with minor revisions, using a small TM for English-Spanish translation. Their aim was to transform, either individually or together, a set of lexical and syntactic rules for input segments from both languages. In terms of syntactic rules, they changed sentences in the active voice into the passive and vice versa, and using a further rule, changed the order of words, phrases and clauses within the sentences (i.e. the syntax of the source segments was converted into a new structure but remained semantically identical). In terms of changes in lexical rules, they replaced single- or multiple-word units with synonyms. The multiple-change rules they applied were as follows:

- Rule 1 replaced a one- or two-word unit with its pronoun and changed the order of a word, phrase or clause.
- Rule 2, in addition to replacing a one-word unit, changed the active voice into the passive voice or vice versa.
- Rule 3 replaced a noun with its pronoun, changed active into passive, and changed the word order.

The two sets of tests were then translated using four TM systems: namely, Trados Studio (2017); Wordfast Pro; OmegatT; and memoQ (the dates and version numbers of the last three systems are not mentioned). The results showed that the TM systems all failed to return the segments as scoring above the default translation threshold; in particular, segments including syntactic transformations and memoQ performed the poorest. In other words, the segments performed least well across all tools tested and MemoQ performed worst across all rule types. Although Baquero and Mitkov’s study experimented with transforming the rules – resulting in multiple consequences at once – rather than editing a single inflection affix, as in the current research, it addresses the same issue of retrieving segments that would require only minor editing.

The conclusion that can be drawn from these studies is that TM systems will not offer translators useful proposals in cases where the segment to be translated is written in a syntactically or semantically different way to otherwise highly similar segments in the TM database, because the TM algorithm cannot recognise the similarity. This omission will occur when the matching score drops below the translation threshold, meaning that the translator may have to translate the sentence from scratch, despite the fact that a useful translation is to be found in the TM database. These results suggest that the TM matching metrics need to be enhanced with natural language processing capabilities.

2.1.2 The development of translation memory matching metrics

The weaknesses of the TM matching metrics (character-string comparisons) led to attempts to go beyond simple surface-form comparison by introducing linguistic information and paraphrasing. Various studies have focused on semantics or syntactic techniques for improving the matching in TM systems. If viewed in relation to the development of their linguistic information processing abilities, we can identify three generations of TM systems.

Most first-generation TM systems appear incapable of language processing, except for a few systems such as Déjà Vu. They perform basic morphological (shallow) processing that does not require additional linguistic information. The vast majority of commercial TM systems currently available in the translation industry belong to this generation (Benis 2003: 24 cited in Mesa-Lao 2020: 104).

The second generation of TM systems use some language processing capabilities in order to apply morpho-syntactic analysis to the segments. This includes using 'chunk' technology to break down the source and target texts. So far, a few commercial TM systems are known to belong to this generation. Similis,[15] for example (Planas 2005), attempts to apply structural and syntactic information to the database. The system splits sentences into syntactic units (chunks), uses a monolingual lexicon and algorithm to attach a grammatical annotation to the chunks, then indexes these as translation units. Thus, when the system searches for a match, it looks not only at sentence level but also at the chunks, increasing the possibility of finding a match.

A different TM system, so-called Translation Intelligence,[16] learns from the content in the TM database to further break down a sentence into smaller segments; each segment is annotated with grammatical information to create translation patterns. As a result, when the system searches for a match, it uses a deep-structure pattern recognition technique, increasing the possibility of finding a match (Grönroos and Becks 2005).

[15] Developed by Lingua et Machina.

[16] Developed by Master's Innovation.

A further TM system developed and tested by Hodász and Pohl (2005), the Meta Morpho[17] TM system, stores multi-level structures, including linguistic information, and retrieves sub-sentential segments. The method involves three levels of similarity – the surface form of the word, the lemma of the word, and the class of the word – in order to calculate the similarity between two segments. The work is based on the method proposed by Planas and Furuse (1999) which extends the edit distance metrics to incorporate lemma and parts of speech, together with the surface form, when calculating the similarity between two segments. Hodász and Pohl also added the detection of noun-phrases that are tagged either by a translator or automatically by an aligner developed especially to improve TM matching for morphologically rich languages like Hungarian. The preliminary results show that the morphological analysis and parsing to determine similarity between two source language segments improves the matchings in the TM. However, the second generation of TMs applies the principle of TM to chunks rather than sentences.

A new generation of TM systems has been proposed that would be able to process segments not only at string-matching and partial syntactic analysis (subsentential) levels, but also at a semantic level (processing paraphrases of texts). Significant research has focused on enhancing the metrics of TM with semantic processing techniques, such as finding semantically equivalent sentences through syntactic and semantic analysis, using a lexical database.

Pekar and Mitkov (2007) have proposed the creation of a third generation of TM tools by introducing the concept of 'semantic matching' – that is, the syntactic and semantic analysis of TM segments. Their study shows how to improve TM matching by using the syntactic structure of sentences. In this method, syntax-driven semantic analysis is used to process sentences over tree graphs, followed by lexicosyntactic normalisation. (A 'tree graph' represents the syntactic analysis of a sentence generated by a parser.) The similarity between the syntactic-semantic tree graphs is then calculated. However, despite giving a valuable insight into the retrieval of better segments, this approach is generally regarded as not feasible for practical implementation due to the amount of processing time required.

[17] A linguistically enriched translation memory for Hungarian-English translation.

Another method that combines the matching metric of Levenshtein distance with syntactic information has been proposed by Vanallemeersch and Vandeghinste (2015). This uses a shared partial subtree metric that compares two parse trees by identifying the overlapping subtree structures they share. The authors claim that their method transfers the complex parse trees used by Pekar and Mitkov into more easily manageable string form, while retaining all the information contained in the nodes. They tested this method on an English-Dutch dataset from Europarl. Their study found that although the combination of fuzzy matching metrics with linguistic knowledge provides significant added value, the results are preliminary and, as the authors note, the tree-matching method is ‘prohibitively slow’.

Other work has focused on approaches which incorporate paraphrasing into TM matching and retrieval processes. An important strand of research has tried to address the weakness of the Levenshtein distance metrics by employing similarity metrics that, when combined with paraphrasing, can identify semantically similar segments even when they differ at the token level.

Utiyama et al. (2011), for example, have proposed a method of searching TMs using paraphrases. This model retrieves sentences from the TM that have the same meaning as the input even if the actual words of the two sentences do not match exactly. Its approach finds that paraphrasing is useful for TMs in terms of both the precision and recall of the retrieval process. However, the downside is that it limits TM matching to exact matches only.

Another approach (Gupta et al. 2015, 2016a; Gupta and Orasan 2014) offers a semantically enhanced edit-distance method by introducing a paraphrase database into the edit-distance metric during the matching process. The extra paraphrase TM database contains semantic information such as lexical, phrasal and syntactic paraphrases. Paraphrases in the PPDB dataset are extracted using a statistical method. Both automatic and human evaluation have shown that paraphrasing improves TM matching and retrieval.

Chatzitheodorou (2015) has taken a similar approach with a method that uses NooJ (a linguistic development tool that is able to create equivalent paraphrases of the source texts; it contains a large lexicon, along with a large grammar set) to create paraphrases of support verb constructions (SVCs) of source translation units. This is in order to improve the matching as much as possible when searching in the TM, so that the sentences share the same meaning but not necessarily the same lexical forms. An evaluation of this method by translators showed that it helps the translation process by speeding it up; however, the number of fuzzy matches drops due to out-of-vocabulary words in NooJ.

Meanwhile, Timonera and Mitkov (2015) have also shown that systems that are enhanced with a clause-splitting and paraphrase function increase their match retrieval. In order to split both the input and the TM segments into clauses, a rule-based module is used at a pre-processing stage (i.e. it is employed before a segment is searched for) in the TM database. The results showed that more sub-segments can be retrieved if clause splitting is applied.

Similarly, Šoštarić (2018) has investigated whether sophisticated metrics and the inclusion of linguistic features could retrieve better TM suggestions, according to both automatic and human evaluation. The translation dataset used in this study was English to Swedish, and was extracted from the European Commission's Directorate-General for Translation (DGT). In the experiment, the data was pre-processed and then parsed. The study concluded that if the matching metric is enhanced with the paraphrase resources, it can achieve a significant improvement.

In a very recent study, Ranasinghe et al. (2020) claim that most of the methods that try to capture semantic similarity in TM were trialled on small databases and are not appropriate for the large TMs normally employed by translators. These researchers, therefore, have introduced an approach that relies on encoding sentences into embedded vectors in order to improve the matching and retrieval process; this means that text similarity is calculated using deep learning (vector representation) rather than texts. The experiment employed the Universal Sentence Encoder for English released by Google (Cer et al. 2018). The results showed that universal sentence encoder architectures handle semantic textual similarity better than the edit distance metrics, even when the word order is changed. It appears to be a promising method, especially for the retrieval of segments expressed in Arabic's free word order.

Another approach towards overcoming the problems of surface-form-based metrics proposes the use of MT automatic evaluation metrics in the context of TMs. Simard and Fujita (2012) experimented with several MT evaluation metrics – BLEU, NIST, METEOR and TER – to calculate similarity as well as to test retrieval quality. They found that those metrics that capture a particular linguistic aspect correlate badly with evaluation metrics that are not susceptible to such features: for example, BLEU gives the highest score to matches retrieved using n-grams (matches based on n-gram precision) as a similarity measure, meaning that the metric gives the best scores to the matches it retrieves itself.

Vanallemeersch and Vandeghinste, (2015) have measured the combination of Levenshtein distance metric and TER. In this study, the evaluation set first calculated the similarity of each input segment with the TM source, and then calculated the evaluation score for the TM target, based on TER, as compared with the reference translation from the evaluation set.

Šoštarić (2018), meanwhile, has examined whether the matching metrics, in addition to edit distance and TER, METEOR, BEER and SPS, as well as the inclusion of linguistic features, could retrieve better TM suggestions, according to both automatic and human evaluation. The study targeted the improvement of the performance of the metrics in matches below the 70% value: the matches which would not be seen by the translator as suggestions in a default translation threshold. The study found that the combination of metrics outperform the edit distance baseline according to the human evaluation, while they come close to or outperform the edit distance baseline according to automatic evaluation in the matching range above the 70% translation threshold.

2.2 Machine translation and the Arabic language

The MT evaluation part of this study focuses on neural network translation. It divides previous studies on the subject into the evaluation of pre-neural network Arabic MT and research into neural network Arabic<>English translation.

2.2.1 Arabic-English translation research pre-neural network

The late 1970s saw the emergence of the first English-to-Arabic translation system, developed by Weidner Communications Inc. This depended on the direct approach; however, due to its use of word-for-word translation, the system could not deal with the complexity of Arabic morphology and syntax (Farghaly 2010) and was superseded in 2002 by the first Arabic-to-English translation system to use the transfer approach: the SYSTRAN MT system (Farghaly and Senellart 2003).

Salem et al. (2008) later produced a RBMT system, the so-called UniArab system (universal MT system for Arabic), which used RRG (role and reference grammar) to accommodate the features of the Arabic source language. This model provided linguistic information for the key grammatical components of a sentence, such as verbs, nouns, pronouns, etc. Nevertheless, although it produced good translations of short sentences, it failed to provide an adequate treatment of the semantics of lexical units due to its limited lexicon, and it did not deal properly with lexical homographs.

Hatem and Omar (2010) proposed an improvement to the transfer approach by attempting to discover rules that would resolve the word-ordering issue when translating from Arabic into English (the rule-based approach). However, the major limitation of all these methods is the difficulty in writing rules that cover all of Arabic's linguistic features.

In terms of the EBMT approach, Phillips et al. (2007) have experimented with the use of morphological information to improve the quality of EBMT when translating from Arabic to English, regardless of the size of the corpora. Their study tested different methods of generalising morphology in order to increase the quality and coverage of the training data. They found, however, that although using the morphological generalisation contributed to an increase in the number of potential matches (recall), it also resulted in the generation of additional irrelevant fragments, thus affecting precision.

The accuracy of SMT was also affected by Arabic's rich morphology and different syntactic structures. As a result, research was focused on using morphological reduction to stem or lemmatise words. Many studies developed word-segmentation routines for SMT to provide morphological analysis and improve the coverage of the lexicon in translations between Arabic and English. Lee (2004) may have been the first to call for segmenting each word in Arabic into the sequence of 'prefix(es)-stem-suffix(es)' to increase the quality of Arabic-to-English translation. However, although morphological segmentation delivers some improvements as the corpus becomes larger, the improvement diminishes at a corresponding rate.

Zollmann et al. (2006), Habash and Sadat (2006), and El Kholy and Habash (2012) have reached similar conclusions. In English-to-Arabic translation, words with complex inflection have to be generated on the target side; the negative effect of data sparsity is a problem for the output. Addressing this problem, Badr et al. (2008) used the morphological decomposition of Arabic as a target language during training, and described the different techniques for tokenisation and detokenisation. They found that morphological tokenisation (i.e. segmentation of the input sequence of orthographic symbols into elementary symbols or tokens) and detokenisation (i.e. a process of morpheme-to-word conversion) helped to produce a better output of Arabic, especially with smaller-sized corpora. The same trend was observed by El Kholoy and Hanbash (2010, 2012) in relation to small training datasets. Both studies highlight different aspects of the fact that the best segmentation is dependent on the direction of the translation. Although Arabic word segmentation is shown to significantly improve output quality in SMT, increasing the amount of training data leads to a reduction in the gain in output quality.

The statistical model, therefore, finds the differences in syntactic structure between the two languages problematic. A key challenge is how to bring Arabic source morpho-syntax to bear on the lexical and word-order choice of the English target string. Bisazza and Federico (2010) tested a chunk-based reordering method based on identifying and moving a clause initial verb on the Arabic side of a word-aligned corpus. The technique handled the most important cases of reordering verbs in Arabic-to-English translations, focusing only on the issue of VSO sentences. Meanwhile, the results of a study by Badr et al. (2009), which concentrated on reordering English source sentences based on rules applied to the source-side parse during training and decoding, showed an improvement in the reordering of short sentences but not of long ones.

In a more recent study, Alqudsi et al. (2019) proposed a hybrid strategy that integrates the rule-based approach with a statistical algorithm in order to address the problems of word ordering and ambiguity. Their results reaffirm that the improvement of Arabic<>English translation is conditional on the size and relevance of the available training data.

Sakhr,[18] the first commercial bidirectional MT service to focus on Arabic<>English translation, applies a hybrid architecture that optimises rules and a statistical model to produce translations; it uses linguistic analysis to understand morphological, lexical, semantic and syntactic contexts when translating Arabic into English. Almahasees (2020) has tested Sakhr against NMT systems (Google and Microsoft) and places it third in the adequacy and fluency ranking (for more details, see the following section).

[18] http://www.sakhr.com/index.php/en/solutions/machine-translation

2.2.2 Neural network translation and Arabic<>English translation

Neural machine translation has become the mainstream approach (see Chapter One, section 1.4.3.6), principally because it enhances the quality of the translations in many language pairs. According to Bojar et al. (2016), the news translation task of the 2016 Workshop of MT (WMT) was to use both the automatic evaluation tool BLEU and human evaluators to analyse the performance of a number of submitted online translation systems for 12 language pairs (i.e. English paired with each of the following languages: Czech, German, Finnish, Russian, Romanian and Turkish). The results revealed NMT systems as ranking above PBSMT (Phrase Based Statistical Machine Translation) and online systems for six out of the 12 language pairs. Toral and Sanchez-Cartagena (2017) also tested nine language pairs (English and Czech, German, Romanian and Russian – and vice-versa – plus English and Finnish) with engines trained for the news translation task at WMT16. The output of the NMT systems achieved better BLEU scores than the PBMT output for all the language pairs apart from Russian<>English and Romanian<>English.

In so far as Arabic<>English translation is concerned, the literature on the NMT paradigm can be placed into three categories: research concerned with developing the NMT approach; studies comparing the performance of the NMT paradigm and PBSMT; and evaluations of the efficacy of NMT systems.

2.2.2.1 Arabic and neural machine translation

Research studies concerning Arabic<>English NMT can be categorised as follows (Ameur et al. 2020):

- **Pre- and post-processing**: the research focuses on improving the quality of NMT systems by performing basic morphological pre-processing of the Arabic source language.
- **Morphology, vocabulary and factored NMT**: the research investigates the effect of incorporating linguistic knowledge sources, via additional tools, into baseline NMT systems.
- **Multilingual and low-resource translation**: the research focuses on the training data of MT systems by using multilingual NMT with both rich-morphology and low-resource settings.

To begin with **pre- and post-processing**, Sajjad et al. (2017) have investigated the effectiveness of three data-driven segmentation schemes to split the words morphologically – namely, sub-word segmentation based on Byte-Pair Encoding (BPE) (Sennrich et al., 2016), characters used as a unit of learning (character-level encoding), and word embedding learnt using character-CNN (Convolution Neural Network). With sub-word segmentation, BPE (a data compression algorithm) is used to split words into smaller units (a sequence of characters), the sequence is not linguistically motivated. With character-level encoding, the fully character-level embedding treats the source sentence as a sequence of letters, segmenting words (and the spaces between them) as characters. With character-CNN, the model takes character-level input and learns word embedding, then the embedding is given to the encoder as input. What distinguishes the fully character-level encoding and the character-CNN is the fact that the former gets word-level embedding, as in the case of unsegmented words, while the latter gets word embedding that is richer morphologically. The study examined whether the pre- or post-processing components are avoided by learning segmentation (including part-of-speech tagging) directly from the training data. The segmenting results compared against Arabic morphological segment tools MADAMIRA[19] (a morphological analyser that generates a list of possible word-level analyses) and Farasa[20] (a segmenter that uses a variety of features and lexicons for segmentation). Their study reported that the BPE segmentation produced the best BLEU scores, and even outperformed MADAMIRA in the Arabic-to-English translation direction. On the POS (part-of-speech) tagging, it was found that the results of the two models of character-embedding were closer to those of the morphological analysers than the BPE.

Oudah et al. (2019), meanwhile, compared the effect of different tokenisation schemes on neural and statistical Arabic<>English MT models in multiple data sizes and domains. The morphology-based schemes they employed were Simple Tokenisation (Raw), which splits off punctuation and numbers; Penn Arabic Treebank (ATB) tokenisation, which splits all clitics except definite articles; and Decliticisation (D3), which relies on the linguistic rules of the source. They also used the segmentation tools, MADAMIRA and BPE. The empirical results

[19] https://camel.abudhabi.nyu.edu/madamira/

[20] http://qatsdemo.cloudapp.net/farasa/

showed that the morphology-based segmentation scheme (ATB) was useful to both the SMT and NMT models, and that a slight improvement to SMT could be achieved if it was combined with BPE. The study also found that the combination of the ATB tokenisation tool with the BPE segmentation tools provided the best results for the SMT models, although not for the NMT models, which suffer from long sentences. The study also concluded that the effectiveness of the tokenisation scheme is based on the type of model to be trained and the type of data available.

Turning to the research studies that have tried to improve the NMT models by incorporating various sources of linguistic knowledge (**morphology, vocabulary and factored NMT)**, the majority of works dealing with Arabic word embedding evaluation use special techniques. Shapiro and Duh (2018), for example, have tried to improve Arabic-English translation quality by combining training data with linguistic information. Their study proposed a method that extended the traditional word embedding model into a morphological one by allowing it to include lemmas. MADAMIRA was used to provide the lemma to predict context words. The model was tested on Arabic-to-English translation tasks using a small corpus of data taken from TED talks. The authors found that the use of morphological word embedding outperformed standard word embedding on an Arabic word-similarity task and as initialisation of the source word embedding in NMT systems for low-resource settings.

Meanwhile, Ataman and Federico (2018) have proposed a Linguistically-Motivated Vocabulary Reduction (LMVR) method to improve the quality of NMT when dealing with morphologically rich languages. The authors tested two unsupervised vocabulary reduction methods in NMT: BPE and the LMVR method, which generates a new vocabulary by segmenting words into sub-lexical units based on their likelihood of being morphemes and on their morphological categories. The two methods were tested on ten translation directions involving five morphologically rich languages, including Arabic. The results showed that input representations were comparable to those obtained from the subword units generated with the LMVR method.

Recently, Ding et al. (2019) investigated the optimal vocabulary size for NMT models that use subword units. They performed a wide range of experiments using different numbers of BPE merge operations on multiple NMT architectures, including encoder-decoder model and multiple language pairs including Arabic. The results of their study showed that the number of the BPE merge operations had a significant effect on the performance of NMT systems. The overall conclusion was that the best outcome when translating into lower-resource languages was obtained with smaller vocabularies.

Ataman et al. (2019) proposed an NMT decoding method that models word-formation via a hierarchical latent variable that simulates the process of morphological inflection. The model generates words one character at a time by composing two latent representations: the first is used to represent the lemmas and the second one, the inflectional features. The researchers evaluated their proposal against subword and character-level decoding methods for translating from English into three morphologically rich languages: Arabic, Czech and Turkish. In terms of the Arabic-English translation experiments, the results showed that using a hierarchical decoding model was more advantageous than using the subword and character-level models.

Liu et al. (2019) have also proposed a method that allows the sharing of source and target word-embedding features to enhance lexical word representations and the interactions between the source and target words. The bilingual features of source and target words provide a closer relationship between source and target word embeddings, and also reduce the number of model parameters used for word representations. Each word embedding (either source or target) is composed of two features: shared features, which incorporate bilingual lexicons to improve the NMT models, and private features that are used to capture the monolingual words. The proposed method, which was tested on five language pairs including Arabic-English, showed performance that was significantly over the Transformer baselines. It seems that although many applications show the usefulness of word embedding, the downside of these embedding models is that they require additional annotated data or specific linguistic tools.

Among those researchers interested in the use of **multilingual NMT under both rich- and low-resource settings,** Almansor and Al-Ani (2018) have addressed Arabic's lack of sufficient parallel datasets by evaluating a character-based hybrid NMT model that combines both recurrent and convolutional neural networks. The reason behind combining the recurrent neural network (RNN), which works with sequential data at both word and sentence levels, and the convolutional neural network (CNN), which embeds vectors for the sequence of the input, is to scale long sequences of data. The model was trained using small parallel datasets from the IWSLT Arabic-to-English and English-to-Vietnamese evaluation sets. The combined model reported noticeable improvement, whichever the language pair.

Nishimura et al. (2019) also investigated the utility of multi-source NMT which incorporates multiple source inputs (from different languages). The model was trained on multilingual corpora, which contained multiple source languages and the target language, so that information from different source languages could be used to generate the target language. They also tested methods that learn and translate from incomplete multilingual parallel corpora in which some source or target translations may be missing. They used a UN multilingual corpus from which they selected French, Spanish and Arabic as the source languages and English as the target language. The multi-source NMT method showed translation performance over the one-to-one NMT baseline.

Tan et al. (2019), meanwhile, have developed a framework in which languages are grouped into multiple clusters and then trained on one multilingual model for each cluster. The framework tested two methods for language clustering: (1) language clustering based on language family and (2) language grouping based on language embeddings for similarity measurement and clustering via an embedding vector obtained by training all the languages in a universal NMT model. Then, the two methods were evaluated on the translation of an IWSLT dataset of 23 languages (including Arabic) into English, and from English into the 23 languages. The first clustering method placed Arabic and Hebrew in the same cluster (i.e. the same language family) and the second placed Arabic, Hebrew and Persian in the same cluster. The language clustering for multilingual NMT showed that the language embedding method outperformed the language family method in almost all scenarios.

2.2.2.2 Comparing neural machine translation and PBSMT performance

Almahairi et al.'s (2016) work was one of the first pieces of research to investigate Arabic-English translation using neural MT. The authors compared an attention-based NMT with a PBMT model, using a variety of configurations in the pre-processing of the Arabic text. Their results revealed that the NMT's translation performance was comparable to that of the PBMT, and the proper pre-processing of the Arabic text had a similar impact on both models. They further observed that NMT was more robust in dealing with an out-of-domain test set in comparison to PBMT. Junczys-Dowmunt et al. (2016) also performed a comparative experiment, testing PBSMT and NMT systems across 15 language pairs and 30 translation directions, including Arabic-to-English and English-to-Arabic, using a UN parallel corpus. Sentences longer than 100 words were discarded. The results showed that the NMT approach performed either on a par or better than the PBSMT with all the translation pairs, while for translation pairs that had Arabic as a source or target language, the gains in BLEU were respectable.

2.2.2.3 Evaluation of NMT systems in Arabic-English translation

Most of the previous evaluations of NMT using the TAUS parameters of adequacy and fluency have compared the neural architecture with traditional MT approaches and established that the neural model generates translations of greater fluency than pre-neural systems (Castilho et al. 2017). In contrast, the current study uses adequacy and fluency scales to focus directly on the quality of NMT. Very few studies have evaluated the NMT approach in Arabic-English translation using adequacy and fluency approaches – the sub-topic of this research.

Almahasees (2017) has evaluated the performance of Google Translate and Microsoft Bing Translator when translating from Arabic into English using the BLEU evaluation method and a literary text. He found that both systems' output produced an inaccurate translation due to syntactic errors, although there were also some lexical errors, and both systems offered relatively similar translations, showing that an automatic evaluation metric is insufficient for providing a full analysis of MT output. His study recommends the use of a combination of human and automatic evaluation. Almahasees (2018) then extended his work by applying a linguistic error approach to the analysis of the translation of small, simple journalistic texts from Arabic into English. Both systems produced higher accuracy in terms of orthography and grammar at the word level but major linguistic errors in terms of collocation.

It appears that, to date, there have been few evaluations using the approaches of adequacy and fluency that focus on the quality of NMT systems' translations between Arabic and English. One study by Almahasees (2020) has tested Google Translate and Bing Microsoft Translator against an Arabic MT system (Sakhr), using texts from different domains. The experiment, which he ran twice (in 2016 and 2017), investigated the efficiency of Google and Bing before and after the switch to a neural network, and compared the results against the hybrid approach used by Sakhr. The evaluation methods were based on error typology and the TAUS translation quality criteria of adequacy and fluency, and the evaluations were performed by four native Arabic-speakers with a good standard of English. The study found that the systems improved after the neural-network switch, with Google outperforming the other two systems in terms of adequacy and fluency, regardless of translation direction.

A further study by Abdelaal and Alazzawie (2020) has examined the quality of Google NMT output when translating informative texts from Arabic to English by rating adequacy and fluency procedures and annotating the errors. The researchers themselves analysed the error typology of the output, and used a questionnaire asking evaluators to rank the two procedures on a scale of 1 to 5. The levels of adequacy were rated by four bilingual speakers, while four monolingual native English speakers rated the levels of fluency. The study found that the semantic adequacy of the system's output was higher than its fluency, which contained errors.

The current study differs from the above experiments in its use of multiple NMT systems, a segment corpus, automatic metrics and two translation directions. Although its analysis is based on the black-box method, it has benefited from reviewing the results of previous evaluations of NMT using corpus-based methods. The review has shown that when NMT was compared with traditional MT approaches, it consistently outperformed them (Almahairi et al. 2016; Alrajeh 2018; Oudah et al. 2019).

2.2.3 Development of MT evaluation

Due to the rapid development of MT systems, MT evaluation has played a vital role in allowing us to assess how well MT systems perform. The MT systems' output evaluation can be achieved either manually or automatically (see section 1.4.4.2). As manual evaluation – although it provides better results – is time-consuming and thus too expensive to employ on a frequent basis, this study also uses automatic evaluation metrics to measure the performance of MT in a more low-cost and less time-consuming way.

2.2.3.1 Manual evaluation development

The earliest human evaluation criteria for MT focused on the intelligibility and fidelity of the output. According to the ALPAC report (Carroll 1966), a translation assessed as intelligible should be readily understandable, while high fidelity means that the translation distorts the intended meaning of the original sentence as little as possible. Human evaluation methods were later developed to also include adequacy, fluency and comprehension (White et al. 1994; White 1995). A further development of manual evaluation methods was led by Linguistics Data Consortium (LDC 2005). Its manual judgment methodology developed adequacy and fluency approaches judged on five-point scales that are defined as follows: adequacy: 'How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?' (i.e. how much of the meaning expressed in the reference translation is also conveyed in the MT output); and fluency: the extent to which the translation is 'one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker' (i.e. does the translation deliver both grammatical correctness and idiomatic word choices) (Görög, 2014).

When developing the traditional evaluation of MT quality, the Translation Automation User Society (TAUS) published three different approaches: adequacy and fluency approaches that were scored according to a four-point scale, and error typology.[21] In terms of the error typology, the translation quality assessment framework was derived from the TAUS Dynamic Quality Framework (DQF) and Multidimensional Quality Metrics (MQM). These methods worked independently before they were integrated to create a standard error typology in 2014. Lommel et al. (2015) presents the harmonised DQF-MQM error typology version.

The seven main error categories of the DQF-MQM version are: accuracy, fluency, terminology, style, design, local conventions and verity, as well as a varying number of subcategories within each of these. The standard error typology is detailed in TAUS (2016),[22] with examples that reflect a functionalist approach and a domesticating strategy – for example, the use of the Fahrenheit scale in a text translated into French is considered an error. The harmonised DQF-MQM error typology template, which is available for download through the TAUS website,[23] offers translators a standard and dynamic model that provides the evaluator with a common vocabulary with which to identify and categorise translation errors. In addition to classifying the error type, the severity level of each error is defined. The harmonised system distinguishes between five levels of severity, in which levels 1,2 and 3 are errors that are critical, major and minor, respectively, while 4 (neutral) is used to mark other information that is not an error, and 5 (kudos) indicates exceptional achievement. This system also introduces the idea of weight, meaning that some error types may be given a heavier penalty than others. According to Lommel (2018), this standard error typology, which combines the two separate methods of

[21] https://www.taus.net/think-tank/news/press-release/taus-releases-translation-quality-evaluation-best-practices

[22] https://www.taus.net/qt21-project#harmonized-error-typology

[23] https://www.taus.net/qt21-project#error-typology-template

quality assessment into a unified systematic framework, has had considerable uptake in industry, research and academia.

In terms of adequacy and/or fluency evaluation, TAUS states that quality evaluation using adequacy and/or fluency approaches is less costly and time-consuming to implement than an error typology approach. The evaluator judges how much of the meaning is represented in the translations (adequacy) using a four-point scale. Similarly, they also judge on a four-point scale the extent to which the translation is well-formed grammatically, contains correct spellings, and is experienced by a native speaker as using natural/intuitive language.[24]

Although, TAUS employs the definitions of adequacy and fluency given by LDC 2005, it uses a four-point scoring scale for each segment, whereas LDC uses a five-point scale – however, according to Koehn and Monz (2006), it is very hard to perform a manual evaluation by scoring a translation on a graded scale according to a five-point system. The TAUS adequacy and fluency framework was used in this research to evaluate the MT systems' output (see Chapter Seven).

A number of studies have used adequacy and fluency methods to manually evaluate MT. Koehn and Monz (2006), for example, have evaluated MT performance with six language pairs, using BLEU (automatic evaluation) and manual evaluation for adequacy and fluency using a five-point scale. They found that the manual evaluation of translation was overly time-consuming and less consistent. However, Turian et al. (2006), who also used adequacy and fluency approaches in addition to automatic metrics to evaluate English-Arabic and English-Chinese language pairs, concluded that, even though human judgment of MT output is often inconsistent and not very reliable, automatic evaluation metrics are even less reliable and therefore cannot replace manual evaluation. The fluency and adequacy methods have also been widely used to assign scores in International Workshop on Spoken Language Translation (IWSLT) evaluation campaigns (Paul et al. 2010; Federico et al. 2012).

[24] https://www.taus.net/index.php?option=com_rsfiles&layout=preview&tmpl=component&path=Articles%2Ftaus-adequacy-fluencyguidelines-may2013.pdf

Overall, manual quality evaluation is criticised for being relatively expensive, subjective, inconsistent and time-consuming. Furthermore, some researchers have found that it is relatively hard to define the scales used in manual evaluation (Lavie 2010), particularly as it can be difficult to reach an agreed definition (Callison-Burch et al. 2011).

However, the main motivation for using manual evaluation (adequacy and fluency criteria) to estimate the translation quality in Arabic-English language pairs in the present research has been the lack of any other studies using these methods in this context. In relation to this, Ali et al. (2020) have introduced an automatic tool that does not require access to the reference translation or the source in order to assess the quality of translated content in Arabic-to-English translations. A black-box method is used to extract a feature set from the Arabic-English language pair, while a five-point scale is used to measure the adequacy and fluency. The model then utilises different machine learning algorithms to predict the quality scores of unseen translated texts at runtime. The study concludes that the tool succeeds in predicting both fluency and adequacy. A further reason behind the selection of adequacy and fluency criteria is that there are still very few studies evaluating NMT using adequacy and fluency approaches (Castilho et al., 2017),

2.2.3.2 Automatic evaluation development

When automatic metrics evaluate the quality of machine translation output, they have to assign quality scores that correlate with the human judgment of quality. However, they do not directly evaluate adequacy or fluency; rather, they indirectly evaluate these criteria by comparing the generated sentence with human-created reference translations (Stent et al. 2005).

Automatic metrics use a number of basic methods to compare MT output against the gold standard translation(s).

The first method computes the overlap of words between the two translations using n-gram matching. This method became the basis of the most widely used MT metrics such as BLEU, NIST and METEOR. BLEU was one of the earliest metrics to use n-gram matching, and has also become one of the most popular automated metrics for evaluating MT quality (Papineni at al. 2002). It depends on lexical similarity n-gram measures (individual words and continuous word sequences of different lengths) to compute the scores. The range of BLEU scoring lies between 0 and 1, where values close to 1 indicate that the similarity of the two translations is high, while values close to 0 indicate that it is relatively low. Although BLEU produces good correlations with human judgment at the document (or system) level, it has been widely criticised for its inadequate or inaccurate evaluation at the segment level, which has a low correlation with human judgment. As a result, the NIST (National Institute of Standards and Technology) introduced a metric to improve the BLEU baseline correlation with human judgment. The NIST metric calculates the informativeness of each n-gram by adding weights, while BLEU calculates n-gram precision using an equal weight for each one (Doddington 2002). Meanwhile, METEOR (Metric for Evaluation of Translation with Explicit Ordering) is designed to improve the correlation with human judgments of MT quality at the segment level (Banerjee and Lavie, 2005). This metric uses a sophisticated and incremental word alignment method to consider stem, synonym and paraphrase matches between words and phrases, along with the standard exact-word matching. Unlike BLEU and NIST scores that depend on precision only, METEOR's score uses recall in addition to precision. Banerjee and Lavie (2005) have proved that METEOR's use of a combination of precision and recall produces a better correlation with human judgment at the sentence or segment level than metrics based on precision alone. Although METEOR introduced recall in the calculation of its score in order to overcome one of BLEU's drawbacks, not all of its features support Arabic. The specific difficulty when translating to and from Arabic is the lack of resources available for linguistic analysis of, for example, word stems, with Porter stemmer, or synonyms, with the standard Wordnet (El Marouani et al. 2017).

The second method used by automatic metrics is the computation of the distance between two translations using the Levenshtein edit distance. Word error rate (WER) is one of the early automatic evaluation metrics to use Levenshtein distance to match two sequences (Su et al. 1992). A modification of the edit distance measurement was introduced later in the translation edit/error rate (TER) metric (Snover et al. 2006), in which the word order (the shift operation) was considered as a single editing step rather than two edit operations. TER is described as working with measuring mismatches – counting transformations – rather than with n-gram methods that measure accuracy (Babych 2014).

An alternative method of measuring the distance between MT output and the gold standard translation(s) is introduced in LEPOR (Han et al. 2012). LEPOR (Length Penalty, Precision, n-gram Position difference Penalty and Recall), and its variants hLEPOR and nLEPOR, represent a language-independent model. The metric optionally uses the part-of-speech information of the candidate translation and gold standard translations. It increases the penalty for a translation that is shorter or longer than the reference translations, institutes an n-gram word order penalty, and combines the penalties with precision and recall measures. This is seen as a fix for the traditional metrics' weaknesses in relation to the problem of either using too many linguistic features or no linguistic information, and has been shown to yield higher correlations with human judgement than METEOR, BLEU or TER.

The above metrics are mentioned only briefly as not all of them were used in this study. However, the hLEPOR variant (Han et al. 2013), which uses a set of enhanced factors for evaluating the correlation between the MT output and the reference translations, was selected for testing, and the results were compared with the precision-based-BLEU (Papineni at al. 2002). Hence, further details about BLEU and hLEPOR are provided below.

BLEU

BLEU uses a measure of n-gram precision based on lexical similarity to compute a score, where the n-gram is the degree of overlap between a candidate translation and its reference (gold-standard) translation(s) of a sequence of unigram (single), bigram (two), three-gram or four-gram words, etc. Precision is a widely used criterion in the MT evaluation tasks.

The n-gram precision means that BLEU calculates the ratio between matched n-grams and the total length of the candidate translation:

$$(P = \frac{the\ number\ of\ n{-}grams\ matched}{candidate\ length})$$

Equation 2.1: Precision calculation

This means there are two main steps involved in calculating the-n-gram precision. The first involves counting the number of matched n-grams regardless of the positions in which they occur – reordering is not penalised. The next step involves producing an overall score, which is computed by combining modified n-gram precision scores using a geometric mean and weighted by a brevity penalty. The brevity penalty is a factor that is used if a candidate length is shorter than the reference length in order to prevent short candidates (relative to their references) from receiving a high score. The key function of a brevity penalty is to penalise sentences that are shorter than the reference. BLEU focuses on n-gram precision only; the component of recall is disregarded.

Recall is the ratio between matched n-grams and the total number of reference translations:

$$(R = \frac{the\ number\ of\ n{-}grams\ matched}{reference\ length}).$$

Equation 2.2.Recall calculation

BLEU scores on a scale between 0 and 1, where the higher the score value, the better the quality of translation – in other words, the higher the number of matches between a candidate translation and reference translations, the better the translation. Further, BLEU can use multiple reference translations: the number of valid references for a certain source is not limited to one, and this affects the correlation score positively – the more references per candidate there are, the higher the BLEU score.

Some researchers claim that metrics based on lexical similarity – such as BLEU – focus on the exact matches of the surface words in the output, and thus perform better in capturing the fluency of the translation than its adequacy (Lo et al. 2012:243). In terms of evaluation of Arabic-English, there are constrained and limited metrics that can be used in Arabic MT evaluation, one of these metrics is BLEU (El Kholy and Habash, 2012). The study states that the standard MT metrics such as BLEU or TER, although have been widely used for evaluating Arabic MT, are too simplistic and inadequate for morphologically rich target languages such as Arabic. Furthermore, it is seen as language biased, performing well on some special language pairs but weakly on other languages pairs. Further, Hadla et al., (2015) examined the translation quality of Google Translate and Babylon MT engines from Arabic to English using BLEU and METEOR 1.5 metrics. The results show that BLEU measurement was closer to human judgment than METEOR.

The hLEPOR model

The hLEPOR (the harmonic mean of enhanced Length Penalty, Precision, n-gram Position difference Penalty and Recall) model is a language independent machine translation MT metric with reinforced factors (Han et al. 2013). The metric, increases the penalty for a translation shorter or longer than reference translations, institutes an n-gram word order penalty, and combines the penalties with precision and recall measures. This is seen as a treatment of the traditional metrics' weaknesses regarding the problem of using too linguistic features or no linguistic information (Han et al., 2012). The model assigns different weights to three factors: sentence-length penalty, n-gram-based word order penalty, and the harmonic mean of precision and recall. Further, this design combines the performance metrics for words and POS (parts of speech) in order to calculate the final score.

1) Length penalty (LP)

To indicate, for example, redundant information (a longer candidate translation), hLEPOR has developed an enhanced length penalty (LP) by increasing the penalty for a translation that differs in length to the reference translations; it applies the penalty to both shorter and longer output translations. LP is calculated by:

$$LP = \begin{cases} e^{1-\frac{r}{c}} & if \;\; c < r \\ 1 & if \;\; c = r \\ e^{1-\frac{c}{r}} & if \;\; c > r \end{cases}$$

Equation 2.3.Length penalty calculation in hLEPOR

where c indicates the output's sentence length and r indicates the sentence length of the reference translation. If the length of the output is the same as that of the reference translation, there would be no penalty, but if the output length is larger or smaller than the reference translation, hLEPOR would calculate an LP, although the penalty would be light.

2) N-gram position difference penalty

The hLEPOR metric also introduces an n-gram position difference (NPD) penalty. The value of NPD is designed to compare the word order of the sentences in the candidate translation and the reference translation. Han et al. (2017) have further developed a matching metric that was initially introduced by Wong et al. (2008). The strategy is based on an n-gram alignment that considers the neighbouring words of the MT output matches, and is provided with a formulaic measuring function. The normalised n-gram position difference penalty (NPosPenal) is calculated by

$$NPosPenal = e^{-NPD}$$

Equation 2.4.Normalisation of n-gram position difference penalty in hLEPOR

where NPD indicates the n-gram position difference penalty. To calculate the NPD value, there are, for the most part, two steps: n-gram word alignment and score measuring.

The first stage is the context-dependent n-gram word alignment which takes into account the surrounding context of the potential word pairs and assigns higher priority to pairs that produce a better match between the MT output and the reference translation. In the second stage, a numbering system is applied to the positions of words in the MT output, and the numerical position of a word is then divided by the length of the translation for normalisation. A similar method is applied to numbering the positions of words in the reference translation.

3) Harmonic mean of precision and recall

The hLEPOR metric applies tuneable parameters for precision and recall by using a weighted harmonic mean of precision and recall (the harmonic mean is a type of numerical average that is calculated by dividing the number of observations by the reciprocal number in the series).

The harmonic mean (α, β) is calculated by:

$$Harmonic(\alpha R, \beta P) = (\alpha + \beta)/(\frac{\alpha}{R} + \frac{\beta}{P})$$

Equation 2.5.Harmonic mean of recall and precision in hLEPOR

where R is recall, P is precision, and α and β are adjustable weights – α and β are two parameters Han et al designed to adjust the weight of recall and precision.

Thus, the formula of hLEPOR is:

$$hLEPOR = Harmonic\left(w_{LP}LP, w_{NPosPenal}NPosPenal,\ w_{HPR}HPR\right)$$

Equation 2.6.The formula of hLEPOR metric

where wLP indicates the weight of the length penalty (LP), and wNPosPenal indicates the weight of the n-gram position difference penalty (NPosPenal), while wHPR indicates the weight of the harmonic mean of precision and recall.

There is the option to combine linguistic information (parts of speech) with the evaluation of the MT system's word performance. Firstly, a score is calculated for the surface words (hLEPOR word) – i.e. the proximity of the candidate translation to the reference translation. Then, a score is calculated on the extracted POS sequences (hLEPOR POS) – i.e. the degree of proximity between the corresponding POS tags of the candidate translation and the reference translation. The final score of hLEPOR is produced by the combination of the two subscores.

The hLEPOR metric has been tested on translations from English into other European languages: Spanish, German, French and Czech and from these languages into English using ACL-WMT11 corpus. The hLEPOR scoring results were compared with the scores of TER, BLEU and METEOR, and its metric gave a promising performance with the languages tested by the metrics without using any external tool or data sources; LEPOR yields better correlation results with human judgment at system-level (Han et al. 2013). In the ACL-WMT 2013 Metrics Task, hLEPOR also had the highest Pearson correlation score with human evaluation on a new language pairs – English-to-Russian and the reverse direction (Han et al. 2014). Marzouk and Hansen-Schirra (2019) used the hLEPOR model and TERbase metric to evaluate the Google NMT output in German-to-English translation. Their results showed that there was no significant difference in the scores of both metrics.

The evaluation of MT systems' output in the research is to use both manual and automatic evaluation; the lack of using the adequacy and fluency methods has been the main motivation behind using them in this research, whereas a comparison between BLEU – one of the most popular automated metrics, and hLEPOR – as an advanced metric.

2.3 TM integration with state-of-the-art NMT

Although MT has seen several significant advances, there are still some doubts about the quality of MT translations; translations produced by human translators (such as TM translation units) or through post-editing remain the gold standard. A number of studies have demonstrated that SMT and TM can be integrated to improve translator productivity: Bulté et al. (2018) have proposed a fuzzy match repair technique using a system that integrates SMT and TM, while Ma et al. (2011) have put forward an approach that merges similar segments from the TM into the source sentence and then uses the translations of matching segments between the input and the TM fuzzy match to constrain the SMT output. Way (2020), however, states that although MT can be successfully integrated with existing TM systems, researchers and developers have yet to get MT to produce TM data for NMT.

Other studies have investigated different MT frameworks in an attempt to improve the quality of MT output by using similar translations retrieved from a TM database. Such translations would offer the advantage of being the product of a human translator.

Ortega et al. (2019; 2020) have developed the idea of 'fuzzy match repair' as a method to improve the quality of NMT output. The fuzzy match-repair technique was applied to correct TM matches using SMT systems (Ortega et al., 2016). Recent implementations of this approach propose a two-step process to generate improved translations. This approach relies on first presenting the proposal of the fuzzy match repair produced by the translation unit, then improving it through an automatic post-editing technique that helps boost the quality of the MT before showing it to the translator. When presented with a segment for translation, the fuzzy match repair technique creates a set of suggestions. The selected suggestion is then given as input to the automatic post-editing system that corrects any errors it finds. The results show that the combination of these two techniques can significantly improve translation quality and outperform the pure NMT model.

In the last few years, research has focused on ways of augmenting NMT training data with fuzzy TM matching by leveraging information retrieved from a TM database. Augmented translation is a form of human translation carried out within an integrated technology environment that offers translators access to subsegment adaptive lookup (from MT, TM and Termbase) to support their work (Muravev 2019).

Bulté and Tezcan (2019) have introduced a neural fuzzy repair method based on concatenating similar translations to source sentences, whereby translations of fuzzy matches are displayed with the source sentence and the MT network learns to use this additional information. This method is based on finding similar source sentences to those in a parallel corpus with which to augment limited data, and then re-using the translation of the original source sentences as translations for the similar source sentences found – in other words, incorporating target segments retrieved from the TM with the source segment so that the NMT architecture can use them to produce translations closer to the TM matches. Two tests using this data augmentation approach have been carried out on two language combinations, English to Dutch and English to Hungarian, using the TM of the European Commission's Directorate-General for Translation (DGT). The input sentence was augmented with the translation retrieved from the TM that showed the highest matching score. The selection of fuzzy matches was made according to a simple similarity measurement between each source sentence and all other source sentences from the TM. The fuzzy source sentences with a similarity score above a given threshold were then stored with their corresponding target sentences. The results show that the fuzzy repair method extends the parallel data and increases the quality of the MT output considerably.

Xu et al. (2020) have extended the data augmentation methods for training NMT to make use of similar translations by experimenting with additional source-side features in order to distinguish between related and unrelated words, and by employing distributed sentence representations. The main integration technique is to retrieve fuzzy, n-gram and sentence-embedding matches to boost NMT performance by using similar translations. The idea is to feed the neural model with information on both source and target sides of the fuzzy matches, using n-gram fuzzy matching to collect similar sentences from translation memories, and to also enlarge the similarity so as to include semantically related translations retrieved using distributed sentence representations. A test carried out on an English-French language pair, using multiple data sets and domains, showed that the different types of similar translations provide consistent improvements in the neural model's accuracy.

In their recent research, Tezcan et al. (2021) have proposed developing a 'neural fuzzy repair' method by using sub-word-level segmentation in fuzzy match combinations to maximise the coverage of source words. This method employs vector-based sentence similarity metrics for retrieving TM matches in combination with alignment-based features on overall translation quality. It aims to maximise the added value of retrieved matches within the neural fuzzy repair paradigm. A test was run on eight language combinations – English<>Hungarian, English<>Dutch, English<>French and English<>Polish – using the DGT-TM of the European Commission's translation service. The study reaffirms that integrating fuzzy matches into NMT through data augmentation leads to a considerable increase in estimated translation quality.

In the same context, a very recent TAUS webinar discussed current efforts to build a dedicated NMT engine: two case studies presented by Lilt and Systran looked at how to increase the quantity of a client's specific data and use it to build a dedicated NMT engine. The case studies, which used sample data from TAUS, show promising results.[25]

To sum up, it appears that it would be helpful to use fuzzy match repair in combination with source–target concatenation to improve the robustness of NMT models. Such TM-NMT integration would not only enhance the advantages of translations produced by a human translator, but could also result in increasing the quality of MT training data.

Emerging from the current research detailed in the literature review, these are the new sub-questions that I want now to investigate. Each of these questions, which is based on the translation tool function, is followed by the hypotheses the research sets out to test.

[25] ttps://info.taus.net/optimize-your-training-data-webinar

The research questions (RQs) concerning TM retrieval are as follows:

RQ1: How useful are the matching metrics of TM systems in retrieving segments that are semantically identical but different in structure?

The hypothesis is that when a TM source segment has a sub-segment move, an otherwise semantically identical input segment may not be in the same word order as the TM source. Thus, when an Arabic segment that requires a move (re-ordering) operation is retrieved from the TM, the matching metrics may prevent two highly similar segments that differ only in their word order from being ranked as close matches, thereby depriving the user of valuable information.

RQ2: To what extent does the TM matching algorithm measure a combination of the inflectional affixes as a word or character intervention when retrieving a segment (which may be combined with a diacritic mark)?

It is expected that if TM systems have some linguistic knowledge, the penalty would be very light. This would be useful to translators since a high-scoring match would be presented near the top of the list of proposals.

RQ3: Is the absence of a Hamza marker weighted as a minor or major difference by the TM matching metrics?

If TM systems have some orthographic knowledge or provide a spell-check of the input, it is predicted that the system would be able to detect orthographic errors and deal with them as minor differences.

The question relating to MT systems is:

RQ4: What degree of adequacy and fluency can different NMT systems achieve when translating between Arabic and English?

Due to a lack of training data resources and its rich morphology, MS systems may produce low-fluency output in translating the Arabic-English language pair. However, the systems may produce relatively adequate translation due to the advantage of using document-level texts- the context sentences are similar.

RQ5: Is there a difference between the scores of BLEU and hLEPOR metrics when translating between Arabic and English?

BLEU is based solely on n-gram precision scores, while the score of hLEPOR is based on three factors: an enhanced length penalty, an N-gram position difference penalty and the harmonic mean of precision and recall. Thus, hLEPOR may calculate better scores.

RQ6: Based on the scores obtained from the evaluation methods, which one of the NMT systems provides better translation when translating between Arabic and English?

Although the MT systems selected in the study use the NMT approach, the systems may respond differently in handling Arabic linguistic features where Arabic is either the source or the target language.

2.4 Summary

This chapter has presented an assessment of the previous research into the challenges faced by the matching metrics of TM in coping with the retrieval of segments that include highly similar meaning, and also studied some methods suggested to develop the TM matching metrics. It reviewed the proposed key stages for improving Arabic<>English translation before the arrival of the NMT approach, and then examined the studies that used adequacy and fluency criteria to evaluate the performance of neural MT systems when translating between Arabic and English.

The review has revealed the lack of experimental studies evaluating the TM retrieval of Arabic segments with a different word order, morphological inflections and the omission of Hamza marker. Furthermore, it has found there is a deficit of studies using a combination of adequacy and fluency parameters and automatic evaluation metrics to assess the quality of the output of NMT systems in terms of Arabic<>English translation.

The current study attempts to address these gaps in the literature by, on the one hand, conducting an experimental investigation into the performance of TM retrieval, and on the other, assessing the output quality of NMT systems using both human and automatic evaluation. The aim of the research is to determine which translation mechanism – retrieval translation or automatic translation – is more effective in translating between Arabic and English. The next chapter will describe the methodology employed by the study, explaining in detail its design, evaluation methods, experimental setups, test data and selection of translation tools.

CHAPTER THREE

RESEARCH METHODOLOGY

3.0 Introduction

Chapter Two identified a lack of experimental evaluations of the retrieval of TM systems and NMT systems when translating between Arabic and English. To redress this lack, this study has proposed to undertake a comparative investigation into multiple TM and MT systems, using an evaluation method based on a black-box approach. This chapter begins by detailing the methodology chosen to investigate the TM systems (including a pilot study), before describing the methodology used to analyse the performance of the NMT systems.

3.1 Methodology of the TM study

The level of objectivity was the crucial factor in the choice of methodology for this evaluation of computer-based translation tools. This section outlines the methodology used to test the TM systems: it begins by describing a pilot study that was run before the main study took place (3.1.1), followed by a description of the research paradigm framing the study (3.1.2), a discussion of the specific evaluation method (3.1.3), an outline of the definitions of variables (operationalisation) (3.1.4), a description of the data used in the study (3.1.5), an illustration of the procedures used for selecting the TM systems to be tested (3.1.6), and finally, a description of the way these TM systems work (3.1.7).

3.1.1 Pilot study

3.1.1.1 Introduction

A pilot study was conducted as preparation for the main experiments described in Chapters Four and Five. It was designed to verify the validity of the variables and parameters that would be implemented in the main investigation, and to detect any possible flaws in its design. The pilot, which was conducted in December 2017, tested a mainstream TM system (Trados Studio 2017), with a view to finding whether the matching metrics would provide useful fuzzy matches, and to verify the approach of the main study. Below is a description of the pilot’s setting, a discussion of the retrieved matches, and the identification of some of the issues requiring modification.

3.1.1.2 Experimental setup

3.1.1.2.1 Data used in the pilot study

The pilot study investigated whether a TM system is able to retrieve segments that include a modified linguistic feature as useful matches. The experiment was designed in two phases: Exercise One was designed to retrieve segments that were semantically identical but had a different word order; Exercise Two was designed to retrieve segments that were exactly the same except for the inclusion of a different inflectional affix. The content of the two exercises was extracted from the MeedanMemory corpus.

3.1.1.2.2 Test set preparation

The first step in the evaluation was to produce a systematic test suite. A test suite is a collection of test cases related to the same test work that are intended to be used to test a software program to show that it has some specified set of behaviours. This was built manually by choosing random segments from MeedanMemory,[26] the first open-source Arabic-English translation resource. MeedanMemory,[27] which was released in 2009, can be obtained from the GitHub platform and is formatted with a TMX extension that facilitates import into CAT tools. It contains around one million words in aligned Arabic-into-English segment pairs in different domains.

In order to extract the exercise segments from the MeedanMemory corpus, it was opened using a TM system, since its interface removes meta-information and tags of translation unit pairs. The translation units were then copied and saved in a Word file. The test segments were chosen as follows:

- *Exercise One*: A set of 15-word sentences in Arabic (syntactically, either verbal or nominal) was extracted from MeedanMemory, each containing a different type of subject unit (i.e. a single or multiple-word unit) and ranging from five-to-seven words in length.

  Having extracted the sentences, the verb and subject constituents were reversed.

[26] https://mailman.uib.no/public/corpora/2009-November/009513.html

[27] https://github.com/meedan/news-memory

- *Exercise Two*: A set of 15-word sentences in Arabic was extracted from MeedanMemory, ten containing a combination of different inflectional affixes to the verb (i.e. the base-form of the verbs with a prefix or suffix attached) and five including the Hamza marker. The morpheme was either one or two characters, and the length of the sentences ranged between five and seven words. In addition, the optional diacritic marks were inserted on some segments.

  Having extracted the sentences, the inflectional affix attached to the verb was replaced with a different inflection morpheme of the same size and in the same location but indicating a different grammatical meaning. The intervention affected the verb form only; the aspects of the subject remained the same. Regarding the sentences that included Hamza, the marker of Hamza was removed.

The two exercises were uploaded into Trados Studio[28] as files for translation, while the whole MeedanMemory was imported as a TM file. Trados Studio was chosen because it is the most widely known of CAT tools and considered a leader in the field.

3.1.1.3 Retrieved matches

3.1.1.3.1 Matches retrieved in Exercise One

Trados Studio's translation editor displayed fuzzy matches that scored above the default threshold (70%), while a number of sentences were not returned as TM proposals. The TM system displayed fuzzy matches for ten sentences, ranging from 73% to 81%, while five were given no match (the scores were literally zero) although the TM database contained the pre-change sentences. It was further observed that although some sentences shared the same length, they were given different matches.

3.1.1.3.2 Matches retrieved in Exercise Two

Trados Studio offered matches ranging from 73% to 89% for the 15 segments. The matches were distributed between the low, middle and high level of fuzzy matches, although the only difference between the input and the TM source was a morphological morpheme or the

[28] The version used in the pilot study was Trados Studio 2017.

omission of Hamza. Also, some sentences with the intervention in the same position were given different matches.

3.1.1.4 Issues arising from the preliminary experiment

The emergence of these issues in the preliminary experiment showed that the main study's methodological framework needed improvements. Hence, in terms of the retrieval of segments containing a reordering fragment, it was decided that:

- the input sentences should be enlarged to include segments longer than seven words and shorter than five words, thus allowing the researcher to identify whether the length of the sentence surface forms has an effect on fragment retrieval;
- the string move should be systematic, comprising one, two, three words, etc., as this would establish whether the move was treated as a number of discrete words or as an undifferentiated block;
- the translation threshold should be reduced in order to display the lower fuzzy matches as this could help identify how the matches were measured.

In terms of the retrieval of segments including a different inflection affix and the omission of Hamza marker, the improvements identified were:

- the input sentences should include sentences shorter than five words;
- the inflection affix morpheme should involve only one variable (a single-character or two-character morpheme) as this would help the return of consistent scores;
- the test data of inflectional affixes should be separated from the test data of the Hamza variation.

The researcher examined the content of MeedanMemory to establish whether it could be used to increase the test data and found that it could meet the requirements for enlarging the segments of different lengths that include a string unit move; however, it did not have the representative data to increase the segments that include different inflectional affixes and Hamza variation. As a result, it was decided to either create a special TM or look for a larger TM in order to find representative data for the main experiment.

3.1.1.5 Summary

The results of the exploratory work carried out in the two preliminary exercises showed a variability in the retrieval of fuzzy-matching scores, paving the way to potentially significant findings in the main study once the enlarged test data and the modification of some of the variables were taken into consideration. The testing of the instruments' effectiveness and variety of matches provided the research with a more solid grounding. The findings of the pilot study led to the decision that the main investigation should implement a quantitative analysis, increase the population and size of the test segments, and expand the investigation to cover more CAT tools, resulting in a comparative evaluation of TM systems.

3.1.2 Research method paradigm

This section describes the evaluation technique used to collect and analyse the data, as displayed in the experiments in Chapters Four and Five. The aim was to evaluate the TMs' retrieval practice when faced with a specific linguistic difference between Arabic and English, and the theoretical framework includes experiments constructed to answer the different aspects of this main research question (see Chapter One, section 1.3). The following sections present an overview of the research method (3.1.2.1), a description of how it was applied (3.1.2.2), and an explanation of the translation direction the study employed (3.1.2.3).

3.1.2.1 Research method

The TM study chose to use quantitative research methods to test the causal relationship and generalise the results. Williams and Chesterman (2014: 63) state that the types of information involved in quantitative research can lead to generalisations based on objective observations and to uncovering patterns in the research. In order to implement research based on quantitative results and to validate the hypotheses in the current study, the matching metrics of TM systems were expected to quantify the differences between the input and the TM sources in numerical data. The use of quantitative research allows for the statistical comparison of results and the future repetition of the experiment.

It was recognised that it would be impractical to use qualitative data in this study due to the difficulty of finding a sufficient number of users to cover all the CAT tools it set out to analyse. The research strategy, therefore, was based on an experimental investigation, seeking new data (fuzzy matches) of observable phenomena (segments that are highly similar but include linguistic features in Arabic). Williams and Chesterman (2014: 58) explain that empirical research 'seeks new data, new information derived from the observation of data and from experimental work; it seeks evidence which supports or disconfirms hypotheses, or generates new ones'. Denscombe (2014) confirms that experimental studies are usually conducted using quantitative data, as the data collected can be measured in such a way as to transform it into usable statistics.

3.1.2.2 Research design

The theoretical framework of this study comprised an experimental design to establish the relations of cause and effect when variables are manipulated. The causal model is a powerful one as it contains the other factors and is capable of forming different kinds of research hypotheses (Chesterman 2000, cited in Saldanha and O'Brien 2013). As a result, the independent variables (linguistic changes) were manipulated and their effect on the dependent variables (fuzzy matches) measured to test the hypotheses. The steps in this causal relationship design are shown in Table 3.1, below.

Research questions	**Independent variable (cause)**	**Dependent variable (effect)**
TM matching measurement and re-ordering operation	Multiple-word string move	Fuzzy match
TM matching measurement and inflection affix combination	Replacing the combination of inflection-affix	Fuzzy match
TM matching measurement and omitting Hamza character	Omitting Hamza marker	Fuzzy match

Table 3.1: The independent variable and dependant variable used in the TM study

The independent variables were applied at different levels (i.e. different sentence lengths) to see how the retrieved matches differed. In terms of the cause component, three different test sets were extracted from TM databases. Each evaluated a linguistic feature: in the first test, fragments from sentences were moved to a different place in the sentence; in the second, a single inflectional affix was exchanged with a different one; and in the third test, the Hamza marker was omitted from the test sentences. It was anticipated that once the test data was translated, the TM systems would provide fuzzy matches due to the intervention of independent variables.

The methodology was chosen following the pilot study, which showed that sentence length could play a role in the matching measurement. The pilot proved the efficacy of the quantitative research method in evaluating TM systems, and provided an opportunity to address the issues arising in the preliminary study design.

The choice of a quantitative methodology also counteracted the threat to internal and external validity innate in experimental research. Internal validity refers to the degree of confidence that the cause-and-effect relationships test is reliable, while external validity refers to the extent to which results from the study can be generalised. The internal validity of this study was ensured by representing replicated multiple samples of each event, while its external validity was ensured by studying natural linguistic events in Arabic to reflect anticipated circumstances.

The examples below show how the people would use the word-order phenomena in Arabic:

Example 1

The different orderings of the three Arabic words(ate), (the boy), and (the apple) convey slightly different meanings (Alqudsi et al., 2014 cited in Ameur et al. 2020)

I. ‘أكل الولدُ التفاحةَ’ / 'akal alwld altfah / in a Verb-Subject-Object order has the meaning of "the boy ate the Apple"
II. ‘الولدُ أكل التفاحةَ’ /alwld 'akl altufaah / in a Subject-Verb-Object order has the meaning of ‘‘(it is) the boy (who) ate the Apple’’
III. ‘التفاحةَ أكل الولدُ’ /altfaht 'ukil alwld / in an Object-Verb-Subject order has the meaning of ‘‘(it is) the Apple (that) the boy ate’’.
IV. ‘أكل التفاحةَ الولدُ’ /'akal altfaht alwld/ in a Verb-Object-Subject order has the meaning of ‘‘the boy ate the apple’’

Even though these versions are structurally different, they express highly similar meaning. However, all the above possibilities are accepted in the Arabic language, the first one (I) is the most common followed by the second (II).

Example 2

The two sentences below belong to the same context, in which the verbal and nominal structures were used with different word order. In sentence 1, the subject precedes its verb

(**السيسي يستقبل**), while in sentence 2 the verb is followed by the subject (**استقبل الرئيس المصري عبدالفتاح السيسي**). This shows that both versions are natural in use in Arabic.[29]

Example 2a

السيسي يستقبل المستشار بالديوان الملكي تركي آل الشيخ

Transliteration:

/ **alsiysi yastaqbil** almustashar bialdiywan almalakii turki al alshaykh/

English translation

Sisi receives advisor at the royal court, Turki Al-Sheikh

Example 2b

استقبل الرئيس المصري عبدالفتاح **السيسي** اليوم الاثنين، المستشار بالديوان الملكي رئيس مجلس إدارة الهيئة العامة للترفيه تركي بن عبد المحسن آل الشيخ.

Transliteration:

/ **aistaqbal** alrayiys almisriu eabdalfataah **alsiysi** alyawm alaithnayn, almustashar bialdiywan almalakii rayiys majlis 'iidarat alhayyat aleamat liltarfih turki bin eabd almuhsin al alshaykh./

[29] Retrieved 17.06.2021https://www.alarabiya.net/saudi-today/2021/05/24/%D8%A7%D9%84%D8%B3%D9%8A%D8%B3%D9%8A-%D9%8A%D8%B3%D8%AA%D9%82%D8%A8%D9%84-%D8%A7%D9%84%D9%85%D8%B3%D8%AA%D8%B4%D8%A7%D8%B1-%D8%A8%D8%A7%D9%84%D8%AF%D9%8A%D9%88%D8%A7%D9%86-%D8%A7%D9%84%D9%85%D9%84%D9%83%D9%8A-%D8%AA%D8%B1%D9%83%D9%8A-%D8%A2%D9%84-%D8%A7%D9%84%D8%B4%D9%8A%D8%AE

English translation

Today, Monday, Egyptian President Abdel Fattah **El-Sisi received** the advisor at the Royal Court and Chairman of the Board of Directors of the General Authority for Entertainment Turki bin Abdul Mohsen Al-Sheikh.

Despite the different position of the subject and verb in example 2a & b, they both indicate the same meaning; the subject-verb structure is used in the news heading, while the verb-subject was in the news details.

Such examples above show the validity of the study: despite the change to the data, alternative word order is possible in Arabic without significant change in meaning. In contrast, this kind of flexibility is not found in the English language, the different word order changes the meaning which is not the case in Arabic – in the data tested in the research. This is why the researcher was confident to make changes and edit the test data, remain natural in Arabic.

In terms of visualising independent and dependent variables, the study used graphs to visualise the results, in which the independent variables are represented on the 'x' or horizontal axis and the dependent variable on the 'y' or vertical axis.

3.1.2.3 Translation direction

The translation direction was Arabic to English: the study focused on the TMs' retrieval performance in relation to Arabic, using it as the source language, because of its rich morphology. English is one of the main languages these applications are designed to handle, so if it were used as the source language, the TMs would work in the same way as for any target language – as long as they can properly display Arabic translations. However, they could be expected to experience difficulties when working from source texts that are morphologically rich since the retrieval process is based on a comparison of source texts.

3.1.3 Evaluation method

The method of evaluating the TM systems (which is further illustrated in the experiments) was based on the information retrieval evaluation approach advanced by Whyman and Somers (1999), in which the TM is treated as a black-box component – a black-box evaluation looks at the input and output of the system, not its mechanics: the focus is on its functionality. In other words, a black box is a testing method in which the functionalities of software applications are tested focusing mainly on the inputs and outputs without knowing their internal code implementation and it is entirely based on software requirements. The study also employed precision and recall measures. It used constructed tests from randomised samples extracted from the TM database as input and the TM systems' matching scores as output.

The type of evaluation tool this study chose to employ was a test suite. According to Lehmann et al. (1996), this comprises a carefully constructed set of examples focusing on a specific linguistic phenomenon in controlled factors; it can also be used as an investigative tool to determine how a system deals with grammatical features. To accommodate the need to refine the quantitative data, the design of the test suite included independent and dependent variables, in order to establish the causality between them.

The test segments in terms of the retrieval of segments containing a reordering fragment were enlarged to a three-to-ten-word length, instead of the five-to-seven-word length used in the pilot study. Likewise, the test segments regarding the retrieval of segments including a different inflectional affix and the omission of Hamza marker were enlarged to a three-to-seven-word length, instead of the five-to-seven-word length used in the pilot study. A short (one-clause) segment range was specifically selected, as the TM database could be expected to contain a greater number of segments; one of the great weaknesses in the TM system is that as segments get longer, the likelihood of a highly similar match becomes correspondingly lower.

The architecture of the translation project was based, on the one hand, on using a corpus as a TM file that included the original segments, and on the other, on translating the source file consisting of the test segments. This meant that the test experiments and the TM data came from the same parallel corpus. To make the comparison as fair as possible, the same input was used as a test suite for each of the selected CAT applications.

3.1.4 Operationalisation

In order to collect the quantitative data of a study, the researcher needs operational definitions of the abstract concepts, turning them into measurable observations. In other words, concepts that cannot be directly measured must be operationalised to make it possible to measure variables in a consistent way. This section describes how the key concepts in this investigation were operationalised.

3.1.4.1 Useful fuzzy match

The concept of a 'useful fuzzy match' in this study indicates a 70% or above matching score. The selection of this percentage was based on Bloodgood and Strauss's (2015) study (see section 1.4.2.3). Further, 70% is the default value in the selected computer assisted translation tools. The measurement was used when retrieving a segment that included a move operation: the retrieved segments that provided a match of 70% or above were considered useful. This formula was used in the experiment in Chapter Four, section A and B.

3.1.4.2 High fuzzy match

A 'high fuzzy match' indicates the ratio of retrieved matching scores ranging between 85% and 95% (see section 3.7.1). This measurement was used when retrieving inflectional verb-variation sentences, where the input and the TM source were identical except for the inflectional affix. The retrieved segments that provided a high fuzzy match were considered to be those shown at the top of the list of TM proposals (memoQ blog;[30] Trados Studio blog[31]) This formula was used in the experiment in Chapter Five, section A.

3.1.4.3 Nearly exact match

A 'nearly exact match' indicates the matching scores ranging between 95% and 99% (see section 3.7.1). This was measured when retrieving a segment with a character marker omission, meaning the input and the TM source were identical except for a difference in the character marker. Thus, the retrieved segments that provided a nearly exact match handled the omission of a Hamza marker as a minor difference (memoQ blog; Trados Studio blog). The concept was employed in the experiment detailed in Chapter Five, section B.

The adoption of 70% as a useful fuzzy-match threshold was based on the ideal translation threshold suggested by Bloodgood and Strauss (2015). However, the concept of fuzzy matches is not standardised, and differs between CAT tools. Translators who work with TM generally

[30] https://docs.memoq.com/current/en/Things/things-match-rates-from-translation-m.html

[31] Fuzzy match grids in SDL Trados Studio | Signs & Symptoms of Translation (signsandsymptomsoftranslation.com)

accept a discounted translation rate for sentences with a higher level of fuzzy matches, as do many agencies (Trados Studio blog).

It is worth noting that the above three variables are all relative to segment length. This allows for a direct comparison between segments of different lengths when analysing the data for each variable.

3.1.5 Database used in the study

The parallel data used for this study was Arabic-English. It is worth mentioning that the data used in the main study is classified as:

3.1.5.1 MeedanMemory

As MeedanMemory, which was used in the pilot study, met the requirement for increasing the test data for the move operation (see section 3.1.1), it was used as a TM file in the experiments in sections A and B in Chapter Four.

3.1.5.2 TM database created by researcher

However, one of the pilot study's conclusions was that MeedanMemory could not meet the requirement of increasing the test data for the retrieval of the morphological features (a different inflection affix and the omission of Hamza marker); there was no representative data for each feature. Furthermore, it was difficult to find a corpus that included specific inflectional verb-variation segments or rich segments with the Hamza variant. For this reason, it was necessary to create a specialised TM file to achieve more effective and robust results. The TM file was created by collecting some generic Arabic segments from the internet; the length of segments ranged from 3 to 7 words.

When creating source segments that included inflection affixes, the verb-stem was generated from a three-character root, combined with a single character as a prefix or suffix, and formatted in four templates (i.e. verb stems) to represent the inflectional verb variations. At least three samples were used in each event: for example, 'يشرب الطفل الحليب الطازج صباحا.' / *yashrab altifl alhalib altaazij subahana* / 'The child drinks fresh milk in the morning'. In such example, if the prefix (يـ) is removed (deletion operation), the tense of the sentence changes into past. In the experiment, we removed such prefixes, so that the input string was different from the TM source by a single character. In relation to segments including variants of Hamza, word one of each segment included Hamza variants such as ‘أ, إ’ in an initial-word, mid-word ‘أ ,ؤ , إ , ئ’ and final-word position ‘أ ,ؤ , إ , ئ’ – Words cannot commence with the Hamza variants of ‘ؤ , ئ’ . Again, at least three samples were used in each event. The selection of the first word of the segments involved a deliberate decision to avoid any further consequences; an intervention in a different position in a word string may lead to different matches: for example, أجرة البيت كانت مكلفة جدا. / *'ujrat albayt kanat mukalifat jidana* / ‘The rent of house was very expensive’. In such an example, the Hamza above Alef is removed in the test segment ‘اجرة ’; the meaning of the sentence does not change. The size of the TM file was 105 Arabic-to-English aligned segments (60 segments representing inflectional verb affixes; 45 segments representing Hamza variation), while the length of segments varied between three and seven words. The researcher’s TM (hereafter referred to as RTM) was used as a TM file in the experiments in sections A and B in Chapter Five. The files were very small but this research and its results were seen as preliminary. The TM file can be found in Appendix One.

Once the Microsoft (MC) Word file to be translated had been created, it was uploaded as a translation project with Arabic as the source language and English as the target, and was then translated into English. As a result, the TM file was created including translation units of variable sizes. The TM file was transferred between the TM systems in the study, where the Arabic source segments were used as the test segments.

3.1.6 Computer-aided translation tools selection

There is a large number of potentially useful CAT (computer-aided translation) tools on the market, in which the TM feature basically performs the same core task of retrieving repeated segments, yet only a couple of these tools are commonly used by professional translators and students trained in their use. Others may be less well known but this does not mean they are not worth considering.

This study chose five of the popular CAT tools available on the market:

- **Trados Studio**[32] 2019 Pro is the market leader and the most widely used by translators (Moorkens and O'Brien 2017; Alanazi 2019).

- **memoQ 9.5**[33] is the second most popular tool among Arabic translators, after Trados Studio, according to Alanazi's survey.

- **Déjà Vu X3**[34] is a tool developed by Atril Language; one of the CAT tools that was recommended to be integrated in translator training in Arabic<>English translation (Al-Jarf 2017)

- **Memsource Cloud**[35] (2020 version) is a cloud-based tool developed by a company of the same name. The reason behind its development is the fact that a web editor may have matching measurements that differ from those of desktop editors (Memsource 2020).

- **OmegaT**[36] is an open-source TM software application. The fact that it is a free TM tool distinguishes it from the other tools mentioned in this list. Its operating system requires Java. According to Gupta et al. (2016.b), the application implements word-based Levenshtein edit distance with some additional preprocessing stage to improve the retrieval.

The first three and the last are workstation-installable tools, Memsource is a cloud-based application, and OmegaT is an open-source TM software application. Four out five of these tools are purely commercial and require a licence from their parent company before use; however, as this study was undertaken at Swansea University, the researcher was able use these tools under the terms of the university's institutional licence. The OmegaT tool, being open-source, was a free application.

[32] https://www.sdltrados.com/

[33] https://www.memoq.com/memoq-versions/memoq-9-5

[34] https://atril.com/

[35] https://www.memsource.com/

[36] https://omegat.org/

In order to learn how to use these tools, the researcher attended sessions of an MA module on translation tools at Swansea University, in addition to watching webinars.

3.1.7 How TM systems work

TM allows a translator to re-use any of the previously translated segments that it retrieves from its database and offers as identical or as similar segments.

The retrieval process is a mechanism for recalling translation units stored in the TM database. This allows the TM algorithm to retrieve the translation candidates in the target language(s) using percentage figures (i.e. TM systems check whether there are any matches between segments). If the system only shows exact (100%) matches, translators may miss out on useful segments; however, TM users can utilise 'fuzzy matches' to get more leverage from the database. The system uses matching metrics to find TM sources that partially match the input. Once found, it displays these segments according to their degree of similarity, ranging from the highest to the lowest (displayed as percentages), and highlights the differences between the input and the TM sources in order to alert the user to the dissimilar fragments (Reinke, 2013). This study analyses the usability of the matches that are not retrieved in this way.

3.1.7.1 Fuzzy match bands

Fuzzy bands are levels of matches that scored below the exact match and are displayed when the pre-translation analysis is performed. Different ranges of fuzzy bands are applied to determine the match percentage. The analysis figures show the number of segments and words with their match percentage in the translation files in each band.

The percentage of fuzzy match levels is very important in terms of the fuzzy match grids (rates); they can help translators or translation agencies to estimate the cost of the translation, and can also be used when estimating the amount of editing needed. Typically, the types of rates are classified as four main categories (memoQ blog[37]; Trados Studio blog[38]):

[37] Match rates from translation memories and LiveDocs corpora (memoq.com)

[38] Fuzzy match grids in SDL Trados Studio | Signs & Symptoms of Translation (signsandsymptomsoftranslation.com)

- Nearly exact match (95% -99%). TM systems not only show the nearly exact match for segments where the text is identical but also where other very small elements such as numbers, tags and punctuation differ.
- High fuzzy (85% -95%)
- Medium fuzzy (75% -84%)
- Low fuzzy (50% -74%)

In the last three levels, the source text is similar to the TM source in the match, but it is only a partial match so it requires editing. The degree of dissimilarity is represented by the TM penalty. Matches below 50% fall into a lower match band. If they are below the lowest match threshold, they are treated as no match at all.

3.1.7.2 Penalties

A TM penalty is a negative value that indicates the loss of reliability in a translation match: the match reflects the correspondence between a source segment in a file and a source segment in the TM, while the penalty reflects the differences between the two segments. The TM system displays the match percentage and the dissimilarity between the segments, but not the penalty; however, the penalty can be calculated by using 100% minus the given fuzzy match percentage – for example, a 75% match gets a 25% penalty. Thus, when a fuzzy match segment is inserted into the target cell, it usually needs to be edited in order to produce the desired translation (Azzano 2011). Further, a TM penalty can be manually specified, as in Trados Studio,[39] to apply a match when there are slight differences. For example, the translator can specify a penalty that applies if the formatting differs between the TM source and the input.

3.1.7.3 Pre-translation

Pre-translation is an automatic task in which matches from the TM are applied to the segments of the source file. When a TM system pre-translates a file to be translated, it looks up every source segment in the TM and then inserts the 100% matches. However, the quality of matches can be controlled by using a fuzzy match threshold – that is, the low match segments can be

[39] https://docs.sdl.com/783545/577388/sdl-trados-studio/tm-penalties

filtered out by using a minimum match threshold, ensuring that matches at the threshold and above are the only ones inserted into the translations. In the CAT tools selected, except Trados Studio, the TM systems accept whatever the value is, Trados does not accept values smaller than 30%.

3.1.7.4 Analysis statistics

Once the pre-translation has been processed, an analysis of the file can take place. Statistics are used to analyse a text automatically from different perspectives. One of the analysis functions comprises counting the segments and estimating the amount of repetition (i.e. matches) in order to estimate the amount of work involved in the translation job. In terms of price quotations, the statistics command can help find out how much to charge (e.g. a high number of fuzzy matches tends to lead to a bigger discount compared to translating everything from scratch).

3.2 Methodology of MT study

3.2.0 Introduction

This section describes the methodology – a black-box approach – the study used to analyse the translation quality of neural MT systems. Its objective was to discover the extent to which the NMT systems perform not only adequately but also fluently when translating between Arabic and English, and then to apply automatic evaluation to confirm which system performs better. The section begins by discussing the study's evaluation method (3.2.1) and mixed methods research (3.2.2), and goes on to describe the data used in the study (3.2.3) and the selection of a representative set of NMT systems (3.2.4).

3.2.1 Evaluation method

The black-box system of testing is the method commonly used to evaluate MT engines where the study is dealing with input and output (Whyman and Somers 1999). In this study, the output quality of NMT systems was evaluated manually by graduate and postgraduate students majoring in Arabic<>English and automatically using the BLEU and hLEPOR metrics.

- **Manual evaluation**

The researcher decided not to use the error-based human analysis method in the MT evaluation part of this research (Chapters 6 and 7) as it needs expert evaluators to identify and score all the errors that the machine makes. However, a part of the harmonised DQF-MQM error typology version is used in Chapter 6 (the transitional chapter where the test data consists of short, simple sentences) to identify the type of errors made by the MT system in the word order and morphology of simple sentences translated between Arabic and English.

The simplified scheme of the adequacy and fluency measures is used in the evaluation of MT systems' output (Chapter 7) as it is relatively easy for informants to understand and corresponds more closely to the kind of judgments that non-experts make.

The use of adequacy and fluency measures to assess the quality of MT systems' output has become reasonably popular among the MT community. Their ranking is typically used in investigations that comparatively evaluate output from different MT systems originating from the same source text. However, much research into Arabic<>English MT translation has been conducted using automatic evaluation; the use of manual evaluation, specifically ranking adequacy and fluency, is still limited (Ali et al., 2018). The relative lack of such research in this area prompted the decision to use the adequacy and fluency approaches in this study.

Following the TAUS quality criteria adopted by this study, the chosen evaluators were asked to participate in a survey in which they had to rate the adequacy and fluency of the translations of a selected source text, using a four-point Likert scale to identify their preferred translations. The question comes directly from the TAUS[40] quality evaluation guidelines:

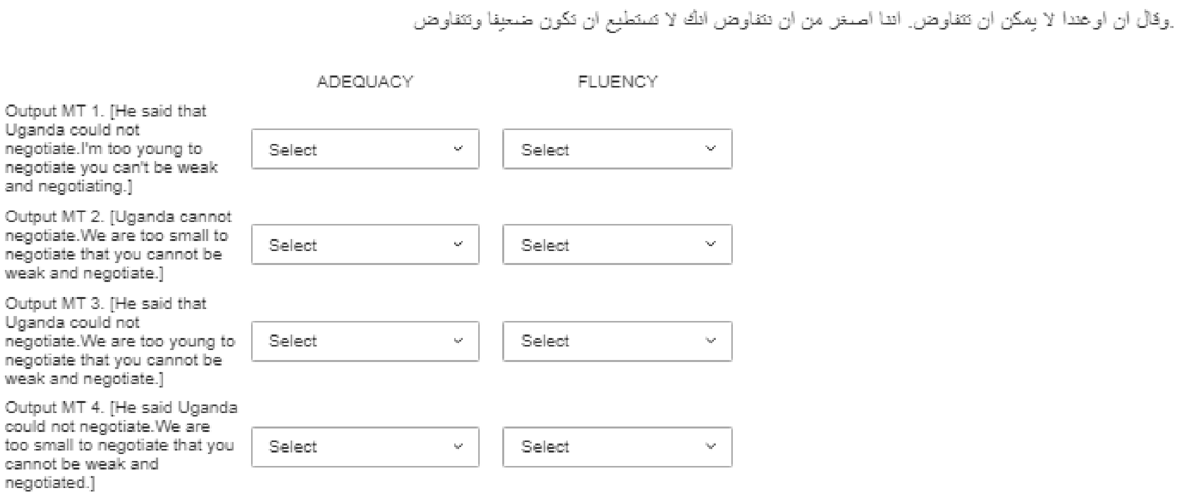
The adequacy of the translations was rated according to the response of participants to the question: 'How much of the meaning expressed in the source fragment appears in the translation fragment?'. A Likert scale of a 1-4 rating was used, where 1 corresponded to 'none of it', 2 to 'little of it', 3 to 'most of it' and 4 to 'all of it'.

The fluency was rated according to their response to the question: 'Is the target text well-formed grammatically so that it contains correct spellings, adheres to the common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker?'

[40] TAUS - The Language Data Network

The main question is this ‘Is the target text well-formed grammatically?’, the rest of the question is a clarification for the potential errors i.e. it is offered as simply guidelines for errors that non-expert evaluators may consider when they try to make judgment about the fluency. (TAUS). Likewise, a 1-4 rating scale was used, where 1 corresponded to ‘incomprehensible’, 2 to ‘dis-fluent’, 3 to ‘good’ and 4 to ‘flawless’.

As shown in Figure 3.1 (below), a source-text unit was provided, and four MT outputs were listed anonymously.



وقال ان اوغندا لا يمكن ان تتفاوض. اننا اصغر من ان نتفاوض انك لا تستطيع ان تكون ضعيفا وتتفاوض.

	ADEQUACY	FLUENCY
Output MT 1. [He said that Uganda could not negotiate.I'm too young to negotiate you can't be weak and negotiating.]	Select ⌄	Select ⌄
Output MT 2. [Uganda cannot negotiate.We are too small to negotiate that you cannot be weak and negotiate.]	Select ⌄	Select ⌄
Output MT 3. [He said that Uganda could not negotiate.We are too young to negotiate that you cannot be weak and negotiate.]	Select ⌄	Select ⌄
Output MT 4. [He said Uganda could not negotiate.We are too small to negotiate that you cannot be weak and negotiated.]	Select ⌄	Select ⌄

Figure 3.1: An example of ranking the four MT systems’ output in terms of adequacy and fluency (Arabic to English)

The expectation was that the output of the MT systems, all of which use an NMT approach, would be positive for adequacy, with possible low levels of ratings for fluency (Abdelaal and Alazzawie 2020). However, TAUS recognises that human judgments are inevitably subjective and therefore recommends recruiting as many evaluators as possible.

In order to estimate the inter-rater reliability judgment, the scores are calculated by the Fleiss’ Kappa statistical measure. Fleiss’ Kapp applies to the agreement on the evaluation among multiple raters. It is defined as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Equation 3.7: Fleiss' kappa formula

The factor 1- $\bar{P}_e$ gives the degree of the that is attainable above chance, and, $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance. It takes values between 0-1 where 1 indicates complete agreement, and there is no agreement among the raters, then <0. Fleiss., 1971). The interpreting of the K values is [< 0= poor agreement, 0.01 – 0.20= slight agreement, 0.21– 0.40= fair agreement, 0.41–0.60= moderate agreement, 0.61– 0.80= substantial agreement, 0.81 – 1.00= Almost perfect agreement]. Inter-rater reliability was assessed for each test segment in chapter seven (N=8) using Fleiss' kappa for adequacy and fluency scores.

- **Automatic approach**

Automatic metrics are rarely used on their own to assess adequacy and fluency, because they have not been proven to be robust. Adequacy and Fluency are often presented together with automatic scoring of string-based metrics (TAUS).[41]

For the purpose of the study, the automatic evaluation metrics BLEU and hLEPOR were applied. The BLEU metric (Papineni at al. 2002) was chosen since it is one of most popular metrics for MT automatic evaluation due to the ease of its computation. Further, the method of measurement is based on the n-gram similarity of a candidate to the reference translation. In addition, those who use BLEU can benefit from its language independence. The reason behind the choice of hLEPOR (Han et al, 2013) is that it is generally regarded as an advanced metric, and according to the best of this researcher's knowledge, it has not yet been involved in Arabic-English translation.

[41] Adequacy/fluency evaluation - knowledgebase (taus.net)

3.2.2 Mixed research method

In the mixed methods approach the researcher presents both the qualitative and quantitative data in the compiled collection. This type of research draws on the potential strengths of both qualitative and quantitative methods, allowing the researcher to explore the different perspectives and relationships that exist between the variables in the research questions. It requires a purposeful mixing of methods in the data collection, analysis and interpretation of the findings. Purposeful data integration at an appropriate stage in the research process enables the researcher to provide more complete data than would a separate research method (Creswell and Plano Clark 2007).

Triangulation, the most common approach in mixed methods research, is a model in which qualitative and quantitative data are collected simultaneously. In the two interactive phases, the quantitative phase plays a basic role, while the qualitative phase builds directly on the results of the first phase. In this way, the qualitative data is used to assist in explaining and interpreting the initial, quantitative results (Creswell and Plano Clark 2007). For instance, a researcher evaluating MT systems' output can identify the causes behind the resultant trends that show increases or decreases in translation quality and explain how these have come into being in the translation process. This design is used in this study to directly compare and contrast quantitative (numeric) results with qualitative findings.

In the first, quantitative phase of the study, the participants – who were asked to complete a survey – were purposively selected for their level of proficiency in Arabic-English language pairs and/or their native language (English) using quantitative measures (a four-point scale). Once analysed, the quantitative findings were compared with the themes that emerged from the qualitative analysis. These interpretive processes provided a context for trends within the quantitative collection: the qualitative findings helped explain the quantitative findings (Creswell and Plano Clark 2007).

The survey consisted of an online questionnaire – used to collect the participants' responses – in which they recorded their experiences of evaluating the output of MT systems using the adequacy and fluency approach. The questions, which the respondents were expected to answer as truthfully as possible, were designed to evaluate the adequacy and fluency of four translations generated by four different MT systems.

The study tried to use a range of sentences of a natural length and with multiple clauses, since the study used short sentences from a TM in the preliminary experiment in Chapter Six. The procedure of selecting the translation pairs presented in the questionnaire was based on the length of the source unit, which ranged from 19 to 39 words, as research has shown that the performance of NMT architecture degrades as the sentence length increases(Bentivogli et al. 2016; Ruiz et al., 2019). Translated sentences where the four MT systems' output happened to be identical or highly similar were omitted because evaluators did not have the option to indicate this in the rating scale.

One of the principal aims of the evaluation of MT systems is to put a valuable resource which includes representative data to the test; however, there is a lack of availability of Arabic-pair datasets. The dataset of the study was extracted from the LDC (Linguistic Data Consortium)[42] corpus, since this is a genuine data resource for Arabic- English translation pairs that has been used in previous studies, such as that of Almahairi et al. (2016); Alrajeh (2018); Oudah et al.(2019). The researcher was allowed access to the LDC2004T18 database thanks to the award of a scholarship from the Linguistic Data Consortium[43] in January 2019.

The Arabic-English corpus LDC2004T18[44] comprises one source file in Arabic and four English reference files. It contains approximately 68,685 sentence pairs, two million Arabic words and 2.5 million English words. Most of the sentences are taken from news articles. The study used sentence pairs which had less than 30 tokens either side. The selection was based on factors affecting the quality of the data and the anticipated translation problems pertaining to each language pair.

For the purposes of the study, ten texts in Arabic and ten in English were randomly extracted from the corpus. Each text consisted of several sentences – test one consisted of 117 sentences in Arabic, while test two included 109 sentences in English. With the advantage of using the document-level translation identified in the NMT research (Miculicich et al. 2018; Tu et al. 2018) and the fact that the neural network architecture can benefit from context, the study

[42] The Linguistic Data Consortium is an international not-for-profit organisation supporting language-related education, research and technology development.

[43] https://www.ldc.upenn.edu/

[44] https://catalog.ldc.upenn.edu/LDC2004T18

chose to keep the sequence of the sentences by selecting full texts from the corpus rather than extracting each sentence separately, without its context.

The questionnaire designed for this study consisted of two main sections. First, a section including a set of demographic questions was used to gather anonymised background information about the participants; it did not link to the participants personally. The questions asked for the following information:

- whether the participant uses MT systems to translate (in either direction) between Arabic and English;
- if so, whether the participant usually edits the MT systems' output;
- whether the participant has previously been involved in evaluating the output quality of MT in terms of adequacy and fluency.

Secondly, based on the question selection criteria mentioned in section 3.2.2.2, the researcher selected four Arabic sources, which were paired with their four MT systems' output in English, to use in the questionnaire. Likewise, four English sources were paired with their four MT systems' output in Arabic. The subset of the four translation pairs were randomly selected from the two test segments. The four MT systems' output was mixed and presented in random order to the participants using the free version of Zoho[45] survey tool.

The online questionnaire provided an information sheet, which supplied the respondents with the researcher's name, the university where the study was taking place, and an introduction to the study itself. This was followed by instructions on how to apply the adequacy and fluency scale, with a link to more information regarding the TAUS quality evaluation used in the study. The information sheet also emphasised that the survey was voluntary and that the data would be stored confidentially. Ethics approval for the study was obtained from the COAH Research Ethics Committee at Swansea University.

[45] Zoho Survey - Instantly sign up for free

The intended participants in this study were high-level translation students. The definition of 'high-level', for the purposes of this survey, indicated a person who was either a graduate or an advanced translation student (MA or PhD level). Unfortunately, as native English-speakers with Arabic as a second language were not available at the time of the study, the survey was distributed via email to translators whose first language was Arabic and second language was English.

The survey was distributed on 2 April 2020 and closed on 2 May 2020. The questionnaire was calculated to take up to 40 minutes to complete (see Appendix Two for the full questionnaire). By the end of the survey period, ten respondents had answered the questionnaire completely, although there were also 23 partial responses.

3.2.3 Data processing and analysis

The data used in the study was downloaded from the Zoho tool into an Excel file, and afterwards arranged and transferred into graphs for analysis.

To analyse the quality of the MT systems' output, each source sentence, along with the four versions of its translation, was evaluated in terms of its adequacy and fluency. This was followed by an holistic analysis of the results: first the average adequacy and fluency scores was calculated, followed by an overall ranking based on the mean score, and finally, in addition to the overall mean of fluency and adequacy scores, the BLEU and hLEPOR scores for each NMT system were calculated in order to obtain an overall picture of the evaluations and determine whether the human and automatic scoring corresponded.

More specifically, according to the translators' usability perspective, the 1-4 rating scale was classified as two levels: a 3-4 quality value and a 1-2 quality value. In the first, the scores of 3 and 4 refer to a high level of output quality, which potentially gives a clear indication of the usability of a translation – this could indicate to translators that post-editing the output would probably be worthwhile. Secondly, the scores of 1 and 2 refer to a poor level of translation, indicating that in practice it would not be useful and would require a major post-editing effort. In other words, a high level of adequacy suggests a translation that conveys 'most or all of the meaning', while a high level of fluency suggests the production of a 'good or flawless translation'. This level is represented by coloured bars in the charts presenting the findings. On the other hand, a poor level of adequacy suggests that the translation conveys 'little or none of the meaning', while a poor level of fluency suggests the production of 'disfluent or incomprehensible' output. This level is represented by light and dark grey bars in the charts (Castilho et al. 2017).

3.2.4 Neural machine translation systems selection

Machine translation systems have shown remarkable progress since the launch of the NMT approach. The NMT systems used in the experiment were Google Translate, Bing Microsoft Translator and Yandex Translate, in addition to the Lilt tool. The idea behind the diversity of the MT systems was that it would allow the study to infer which would respond better to handling Arabic linguistic features. The output of the systems were annotated to identify the types of errors committed.

- **Google Translate**,[46] a free NMT service, is the most popular MT system. In October 2020, it provided translation for 109 languages.
- **Bing Microsoft Translator**,[47] a free NMT service, is one of the most widely used MT systems. In October 2020, it supported translation for 73 languages.
- **Yandex Translate**[48] is a free NMT service, which has the potential to be a competitive system in the future. In October 2020, translation was available in 97 languages.
- **Lilt**[49] is a commercial translation tool. The reason behind involving Lilt, which has only recently begun to support Arabic (at the end of 2019), was to compare its translation quality against that of the free systems.

Lilt was also included in the selection of NMT tools because it uses a certain type of interactive and adaptive mechanism (see Chapter One, section 1.4.3.6) – interactivity refers to the ability of an MT system to autocomplete what the user is going to type, while adaptivity is a quality whereby it learns from corrections and is thus continuously updating itself (Daems and Macken 2019). This indicates that Lilt is an NMT system that could potentially be used in

[46] https://translate.google.co.uk/
[47] https://www.bing.com/translator
[48] https://translate.yandex.com/
[49] https://lilt.com/

future research. However, the main focus of this study was on the quality of the output of the free online systems. (see Chapter One, section 1.4.3.6).

3.3 Conclusions

This chapter has explained the methodology used in this study, describing the research design and the processes involved. The first part of the chapter presented the method used to evaluate TM retrieval when translating from Arabic to English. This section shed some light on the black-box approach – including the pilot study, the set-up of the test data, and the basic knowledge of TM systems – that was used to investigate the TM retrievals.

The second part of the chapter described the method used to evaluate the performance of machine translation systems when translating between Arabic and English. This section highlighted the selection strategies behind the manual and automatic methods used to perform the MT evaluation, and the methods used to analyse the performance of the NMT systems.

The findings of the experiments with TM systems will be discussed in Chapters 4 and 5, while the findings of the experiments with NMT systems will be discussed in Chapters 6 and 7. Finally, Chapter 8 will deliver a comparative evaluation of the findings for the two systems.

CHAPTER FOUR

TRANSLATION MEMORY RETRIEVAL AND SUB-SEGMENT MOVE: SECTIONS A AND B

4.0 Introduction

Chapter Four presents an experimental evaluation of TM retrieval for a move operation in Arabic – as mentioned previously, Arabic is a language with a flexible word-order. The chapter is organised into two main sections: section A investigates how a sub-segment (fragment) move affects TM retrieval, while section B examines the retrieval of segments with a three-operation edit distance. The results reveal whether the TM systems deal with the move operation as an edit operation. The chapter concludes with a summary of the findings.

SECTION A:

4.1 A comparative evaluation of the performance of TM in the retrieval of Arabic-English segments containing a sub-segment move

Section A investigates the way in which TM systems retrieve Arabic-to-English translation units when the Arabic source contains a sub-segment move, testing whether the TM metrics can recognise such segments as highly similar. Its hypothesis is that the current TM algorithms will prevent two highly similar segments that differ only in their word order from being ranked as close matches, thereby depriving the user of valuable information.

The section introduces the experimental setup (4.1.1), summarises (4.1.2) and discusses (4.1.3) the findings, and then evaluates the results (4.1.4). Finally, it presents the conclusions drawn from the study (4.1.5).

4.1.1 Experimental setup

In order to answer this research question and test the hypothesis mentioned above, the experiment was carried out using the MeedanMemory corpus (for more detail, see Chapter Three, section 3.2.5.1), and the corpus was then used as the TM. The test suite, which was saved in a Word document (*.docx), contained 95 segments, ranging from three to ten words in length (see Appendix Three for the test segments A). The segment length was chosen in the assumption that it would yield enough data to enable the researcher to draw a valid conclusion; the minimum three-word length represents the shortest possible full sentence (e.g. verb-subject-object – VSO). As it is easy to find segments sharing the same number of words, the length of the segments was selected according to the number of words they contained.

For the purposes of the experiment, it should be noted that the VSO- and SVO-ordered segments varied in terms of string length. When the unit was a noun, the subject string was the same in Arabic as in English and could be in single or multi-word form (see Chapter One, section 1.4.1.2). Accordingly, the subject constituent of the test segments contained different types of sub-segment units, comprising one, two, three or four words (1W, 2W, 3W, 4W). The rationale for testing sub-segment variations was to establish whether the move operation would be treated as a number of discrete words (multi-word units) or as an undifferentiated block (one chunk).

In addition, the agreement between verb and subject was taken into consideration: the source segments selected from the corpus were either verbal segments or nominal segments in singular form. This was considered the optimum way of avoiding the risk of the exchange of verb and subject components which would distort the meaning when building the test.

To summarise, the variables applied in the test segments were as follows:

- The routine of segments ranged from three to ten words in length.
- The subject string of each segment routine comprised four different events: a one-, two, three- or four-word unit.
- In each event, at least three samples were used to verify the results. This meant that the same match was repeated in the three different samples.

4.1.1.1 Sub-segment move

Having extracted the segments according to the testing procedures mentioned above, the sub-segment move (the reversal of the subject and verb components) was applied, as appropriate. The reverse operation occurred when the subject was moved into the verb position in each of the original segments, while the position of the object remained fixed. In this step, the component in the first position (the subject or verb component) was moved into the component in the second position to avoid any algorithm problems: for example, in terms of a one-word unit move, the original sentence تنشيء الكويت لجنة للطاقة النووية / tanshi' alkuayt lajnatan lilttaqat alnawawia / ‘Kuwait establishes a committee for nuclear energy’ has been changed into ~~تنشيء~~ الكويت تنشيء لجنة للطاقة النووية / alkuayt tanshi' lajnatan lilttaqat alnawawia / ‘Kuwait establishes a committee for nuclear energy’. Regarding a four-word unit move, the sentence

تخفق الوكالة الدولية للطاقة الذرية مجددا في اختيار خلفا للبرادعي / takhfuq alwakalat alduwaliat lilttaqat aldhariyat mujadadaan fi aikhtiar khalafaan lilbaradieii / 'The International Atomic Energy Agency fails again to choose a successor to ElBaradei', for example, has been changed into ~~تخفق~~ الوكالة الدولية للطاقة الذرية تخفق مجددا في اختيار خلفا للبرادعي / alwakalat alduwaliat lilttaqat aldhariyat takhfuq mujadadaan fi aikhtiar khalafaan lilbaradieii / 'The International Atomic Energy Agency fails again to choose a successor to ElBaradei'.[50]

The length of segments was also considered to ensure that the intervention did not exceed 50% of the segment length. This meant a three-word segment (i.e. routine) would only accept a one-word move (i.e. event), while a ten-word segment would accept all four types of sub-segment intervention. Table 4.1 (below) displays where each move (M) event could be applied in the length routine, and where it was not applicable (N/A).

Segment routines	**One-word M**	**Two-word M**	**Three-word M**	**Four-word M**
3-word segments	Yes	N/A	N/A	N/A
4-word segments	Yes	Yes	N/A	N/A
5-word segments	Yes	Yes	N/A	N/A
6-word segments	Yes	Yes	Yes	N/A
7-word segments	Yes	Yes	Yes	N/A
8-word segments	Yes	Yes	Yes	Yes
9-word segments	Yes	Yes	Yes	Yes
10-word segments	Yes	Yes	Yes	Yes

Table 4.1: The four different events moved

[50] Track Changes was used for the intervention.

The table above represents the four different types of intervention on the subject string – one-word move (1WM), two-word move (2WM), three-word move (3WM), or four-word move (4WM), while the verb was always expressed in a single-word string.

If it was necessary to translate the extracted segments from the TM without any changes, the matches would obviously be 100%. Thus, the VSO order became SVO, bearing in mind that the meaning of the two segments was identical in both versions.

4.1.1.2 Translating

Having processed the test segments, we then imported the test into the five CAT tools (see Chapter Three, section 3.1.6) as a translation file, while the translation project in each tool was based on using MeedanMemory as a TM file that included the original segments. The minimum translation threshold was set at 30% (See more in Chapter Three 3.1.7.3), and then the pre-translation function was run. This made it possible to discern the usability of the matches. As the test segments and the TM database were semantically identical, although the word order was different, it was considered desirable for the TM matching metrics to produce fuzzy matches at the match threshold or higher that could then be reused.

To see the fuzzy matches that were assigned, due to the sub-segment move, the findings associated with each segment routine and a string move event were assessed using a new match value.

4.1.2 Findings

The results provided by the five CAT tools show that their TM algorithms measured different matches. The full retrieved matches provided by the five selected systems are presented below, with the assumption that scores at a 70% (i.e. a 30% penalty) or higher are in the usability bracket, while scores that are lower are not suitable for re-use, and the translator would have to translate from scratch. Column charts are used to represent the retrieved scores that fall above and below the match threshold, in which the multiple-word string moves (independent variables) are represented on the 'x' or horizontal axis and the fuzzy matches (dependent variable) on the 'y' or vertical axis. Further, the TM systems are presented in alphabetical order.

4.1.2.1 Déjà Vu X3 output

The Déjà Vu X3 (hereafter referred to as DVX) findings indicate that the retrieved matches were found to occupy a consistent band according to the number of words in the test segment, whether these contained a one- or a four-word move. The matching scores decreased gradually as the length of the segments decreased. For instance, the retrieval of ten-word segments provided an 80% match; nine-word segments, a 77% match; eight-word segments, a 75% match, and so on, whichever the unit move. Further, the move of single- or multi-word units was dealt with as a one-chunk move. In the ten-word segment routine, for example, the match retrieved for moving a one-word, two-word, three-word or four-word unit was 80% (i.e. the same penalty was applied however large the unit moved). The matches of three-to-ten-word routines containing a unit move ranged from 34% to 80%. Figure 4.1.1 (below) shows the retrieved matches that each segment length (SL) produced for a unit move.

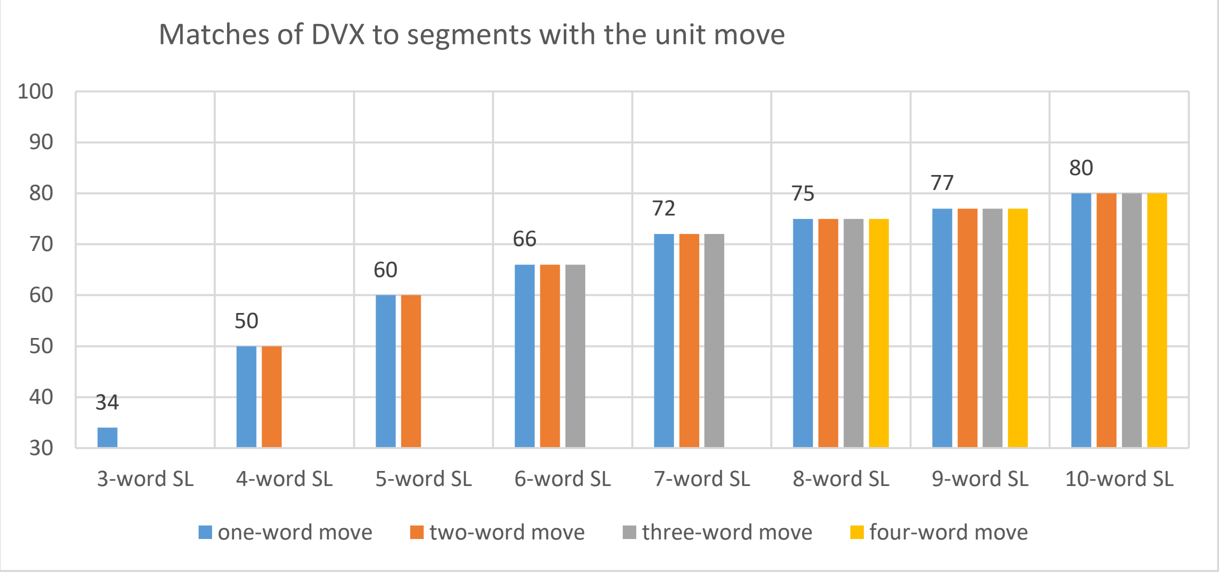


Figure 4.1.1: DVX matches for retrieving 3-to-10-word segments with the unit move

It can be said that the performance of the DVX TM was weak in terms of the retrieval of the move operation. The routines of three- to six-word segments provided matches of less than

70% (i.e. a potential translation threshold) whichever the unit move, meaning that the translators would not be made aware of these proposals.

4.1.2.2 memoQ 9.0 output

The matches produced for memoQ fell into two different categories: the scoring of five-to-ten-word segments was consistent, while the scoring of three- and four-word segments was inconsistent.

In terms of the routines of five-to-ten-word segments, the matches decreased as the number of words decreased. For example, the retrieval of ten-word segments produced a 79% match; nine-word segments, a 77% match; eight-word segments, a 75% match, and so on, whichever the unit move (DVX produced 80%, 77%, 75%, respectively). Further, the unit moves were treated as an undifferentiated block. For instance, with the ten-word segment, the match retrieved for moving either a one-word, two-word, three-word or four-word unit was 79%, (with DVX, the match retrieved was 80%). Regarding the three- and four-word segments, the matches were inconsistent regardless of how many words the segment contained. For instance, a 62%, 67% and 71% match was produced, even though the segments comprised three words including a one-word move. The matches of three-to-ten-word routines containing a unit move ranged from 62% to 79%. Figure 4.1.2 (below) shows the matching scores that each segment length (SL) produced due to the unit move.

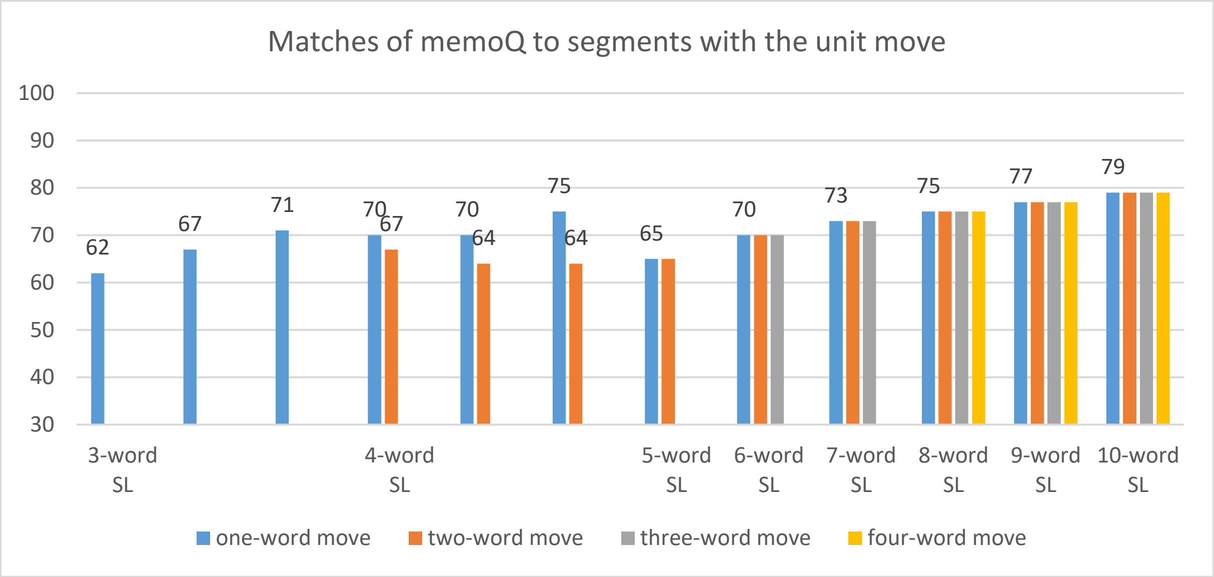


Figure 4.1.2: memoQ matching scores for retrieving 3-to-10-word segments with the unit move

As Figure 4.1.2 clearly shows, the match of less than 70% was given to the shorter segments, meaning that these segments would not be offered to translators as TM proposals.

4.1.2.3 Memsource Cloud output

The matching scores of Memsource, which were apparently scattered, were derived in a different way and were inconsistent. Hence, the experiment used the filter feature in the system's setting to sort the source's shortest segment first, based on the number of characters. When observing the fuzzy matches, the scores appeared to decrease as the total number of characters in the segment fell regardless of how many words a segment contained and the size of unit moved. Similarly, when the source was sorted according to the principle of the longest first, the matches appeared to increase as the total number of characters in the segment increased. For example, an 87% match, the highest score in the matches retrieved, was recorded for two nine-word segments: a segment of 70 characters containing a one-word move, and a segment of 64 characters containing a four-word move. In terms of the lowest matches, the shortest segments, which were three-word segments containing a one-word move, provided the lowest scores but in a different range: a segment of 14 characters was given a 42% match; a segment of 16 was provided with a 48% match; and a segment of 19 was given a 46% match, even though these segments shared the three-word length and the one-word move.

Due to the difficulties in displaying the large amount of scattered data with different segment length routines and different unit moves, the researcher has decided that the retrieval of segments that depended on the number of characters should be classified as follows: the recall for 15-30 characters in length was given a less-than 70% match, while the recall of 31-70 characters in length was given a 70% match or above. This suggests that around 28 out of 95 segments – representing 19% – were retrieved as low fuzzy matches.

4.1.2.4 OmegaT output

The matching scores obtained by OmegaT also dropped steadily as the segment length became shorter. Moreover, the algorithm classified the matching scores as three basic patterns according to the size of unit moved: a) pattern one was a one-word move; b) pattern two was a two-word move; c) pattern three was a three- and four-word move. For instance, in terms of a one-word move, the retrieval of ten-word segments produced a 90% match; nine-word segments, an 89% match; eight-word segments, an 87% match; and so on (DVX produced 80%, 77%, 75%, respectively; memoQ produced 79%, 77%, 75%, respectively). However, the matches reduced as the size of the unit move increased. For example, in the ten-word segments, the match retrieved for moving one word was 90% (as mentioned above), moving two words was 85%, and an 80% match was given when moving both three- and four-word units. The matches of three-to-ten-word routines including a unit move ranged from 62% to 90%. Figure 4.1.3 (below) illustrates the matching scores that each segment length (SL) provided due to the unit move.

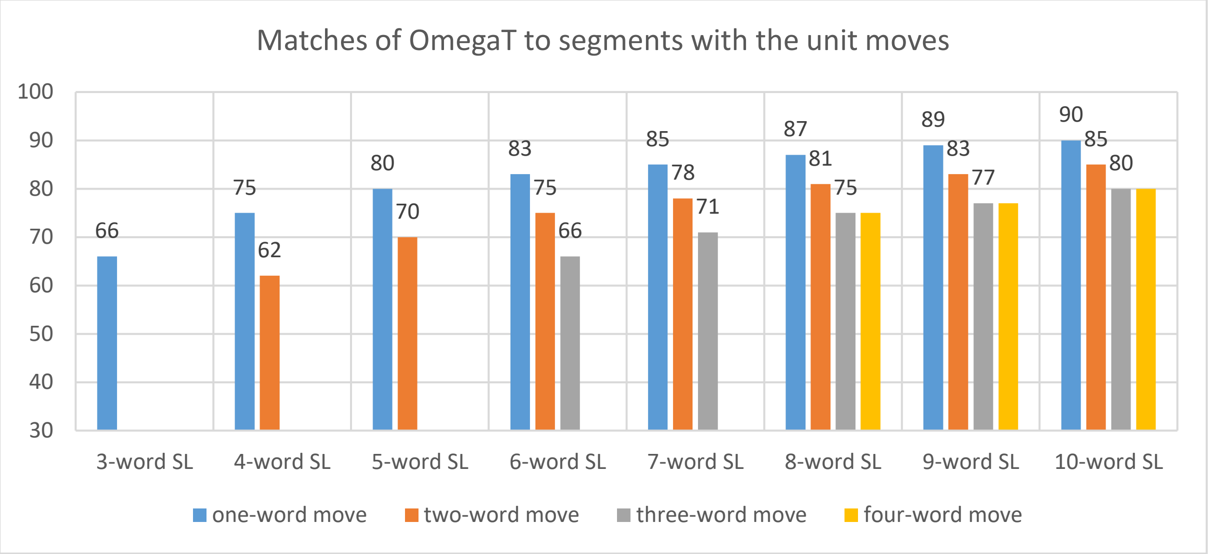


Figure 4.1.3: OmegaT matches for retrieving 3-to-10-word segments with the unit move

It was noted that the fuzzy scores produced for retrieving those segments with a one-word move were high enough to exceed the match threshold, apart from the three-word segment. Furthermore, the system treated the one-word move (i.e. a single chunk) better than it did a multiple-word unit. This shows that, in some cases, the OmegaT matching mechanism succeeded in tackling the pattern of a one-word move whatever the length of the segment. Section 4.1.3 provides a possible explanation of the mechanism the TM system uses.

4.1.2.5 Trados Studio 2019 output

The Trados Studio test was run twice, first with UpLIFT off and then with it on. The UpLIFT[51] technology includes matching based on fragments, thereby helping with the fuzzy match research.

- *UpLIFT off*

The matching values produced by Trados Studio also fell as the segment length became shorter. The matching rates, however, were consistently related to the segment's word length. The scores were classified as four essential patterns according to the size of the unit moved: a)

[51] UpLIFT technology in SDL Trados Studio

pattern one was a one-word move; b) pattern two was a two-word move; c) pattern three was a three-word move; and d) pattern four was a four-word move. With the one-word move, for example, the retrieval of ten-word segments produced an 87% match; nine-word segments, an 85% match; eight-word segments, an 83% match; and so on (DVX produced 80%, 77%, 75%, respectively; memoQ produced 79%, 77%, 75%, respectively; and OmegaT produced 90%, 89%, 87%, respectively). However, the match percentages were lower in pattern two (i.e. a two-word move) and pattern three (i.e. a three-word move). For example, the retrieved matches of the ten-word segments for moving one word were 87% (as mentioned above), whereas moving a two-word string and a three-word string produced 80% and 73% matches, respectively. Regarding the pattern of a four-word move, interestingly, the system assigned retrieved segments with such a unit move high percentages – for example, a ten-word routine with a four-word move was given a 93% match (DVX gave it 80%; memoQ, 79%; and OmegaT, 80%). The matches of three- to ten-word routines containing a unit move ranged from 49% to 93%. Figure 4.1.4 (below) illustrates the fuzzy matching scores that each segment length (SL) produced due to the unit move.

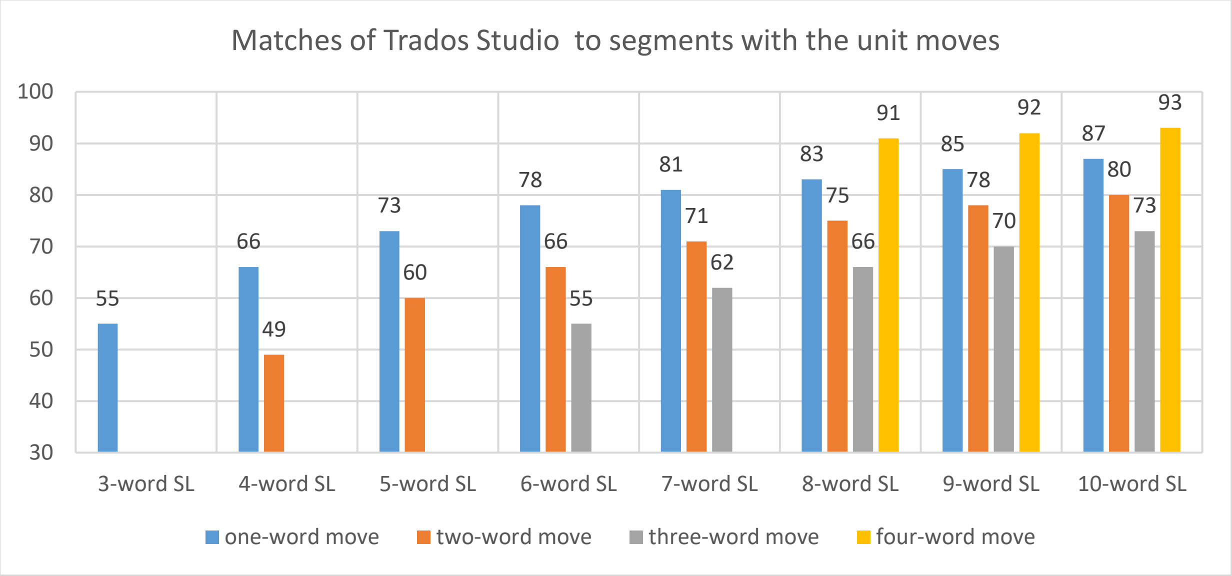


Figure 4.1.4: Trados Studio matches for retrieving 3-to-10-word segments with the unit move with UpLIFT off

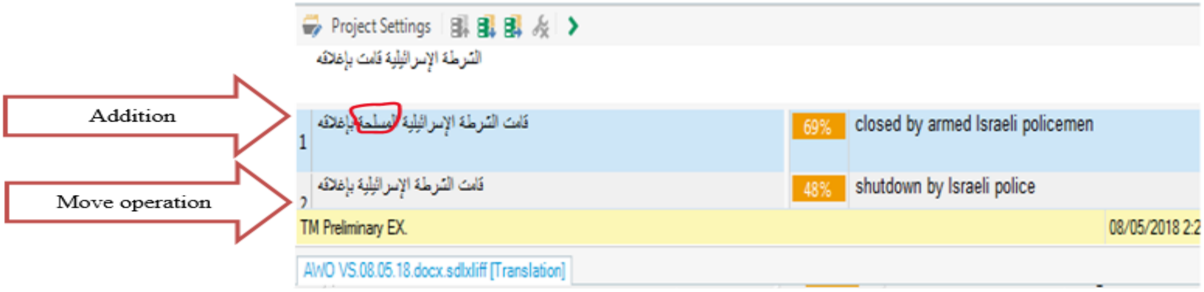
Figure 4.1.4 clearly shows that Trados Studio's low scoring not only covered the short segments but also encompassed seven- and eight-word segments, which were ranked higher than 70% in the other TM systems. However, it can be further noted that a potentially important result of the Trados Studio matching mechanism is that the scores of a four-word unit move, which would potentially occur in a long segment, significantly increased. The mechanism used is discussed in section 4.1.3.

- *Comparison of Trados Studio versions (2014, 2015, 2017, 2019)*

The study's experimental evaluation was primarily intended to test the retrieval capabilities of different translation tools; however, it also gave the researcher a chance to evaluate different versions of the same system. Trados Studio supplied a short-term licence on request, allowing access to the old versions of Trados Studio (2014 and 2015). The aim was to test whether UpLIFT, which was first introduced in the 2017 version, improved the retrieval of segments with a move operation. In order to test this, the evaluation in section A was run in May 2018 to compare the performance of Studio 2017 with that of the 2014 and 2015 versions. The three different versions produced the same score.

- *UpLIFT on*

The test was also run with UpLIFT enabled. The investigation of UpLIFT's retrieval capabilities, despite producing many scores identical to those provided when the UpLIFT was disabled, also produced some TM proposals with higher matches than those which included a move operation. The proposals seem to have undergone a three-operation edit (i.e., addition, deletion, substitution). Hence, this study used the results obtained when UpLIFT was disabled. However, these results provided motivation to compare the segment retrieval that included a move operation with those including a three-operation edit. Figure 4.1.5 (below) displays a screenshot of two TM proposals. Proposal one, which included a one-word addition, gave a 69% match, while proposal two, which comprised the same surface form but included a two-word move, provided a match of 48%.



الشرطة الإسرائيلية قامت بإغلاقه

1	قامت الشرطة الإسرائيلية المسلحة بإغلاقه	69%	closed by armed Israeli policemen
2	قامت الشرطة الإسرائيلية بإغلاقه	48%	shutdown by Israeli police

TM Preliminary EX. 08/05/2018 2:2

AWO VS.08.05.18.docx.sdlxliff [Translation]

Figure 4.1.5: An example of UpLIFT retrievals

It could be said that the Trados Studio algorithm system's measurement of the similarity between a TM segment containing an addition and the input segment is closer than for the one containing a move operation. Hence, this study used the results obtained when UpLIFT was off. However, these results provided the motivation to compare the segment retrieval that included a move operation with those including a three-operation edit (see section 4.2).

To summarise, a reordering operation can indeed affect the reuse of previous TM segments, especially when the sentence is short.

4.1.2.6 Retrieval of TM systems: similarities and differences

The outcomes provided by the five CAT tools show that their TM matching algorithms shared some similarities but apparently they used different methods to calculate the matches.

In terms of the similarities, the TM systems shared as a whole the fact that the matching scores decreased as the length of the segments decreased, whether the retrieval depended on the number of words (DVX; some scores of memoQ; and OmegaT) or the total number of characters (some scores of memoQ; and Memsource) in each segment. The similarity of the decrease in scores produced by the TM systems is exemplified in Figure 4.1.6 (below), which illustrates an example of a one-word move.

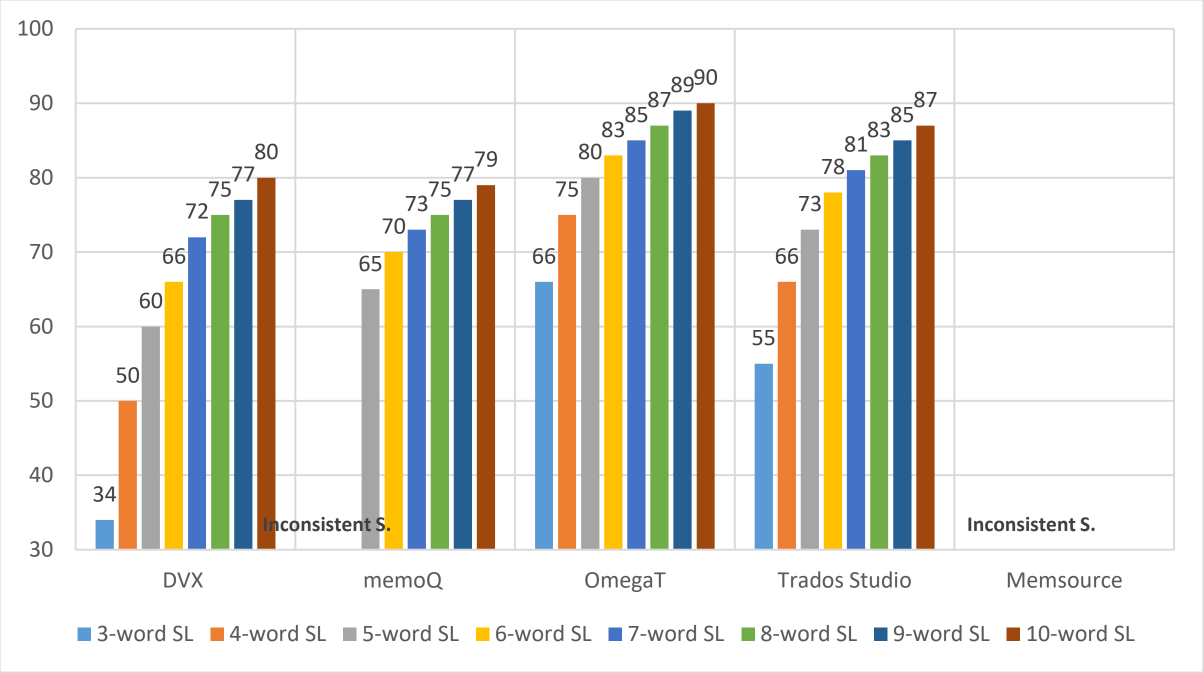


Figure 4.1.6: The four TM retrievals of 3-to-10-word segments with a one-word move[52]

The figure above clearly shows that the matching scores increase if a segment is longer, while they decrease if a segment is shorter. In Memsource, the short segments also produced low matches, although the scores themselves were inconsistent.

In terms of the differences, each TM algorithm used its internal matching mechanism to compute the matches, although these all appeared to be based on the string of surface forms. The differences in the scores produced by the TM systems are exemplified in Figure 4.1.7 (below), which illustrates the retrieval of a ten-word routine with a different unit move.

[52] Due to producing inconsistent scores, Memsource' fuzzy scores were excluded

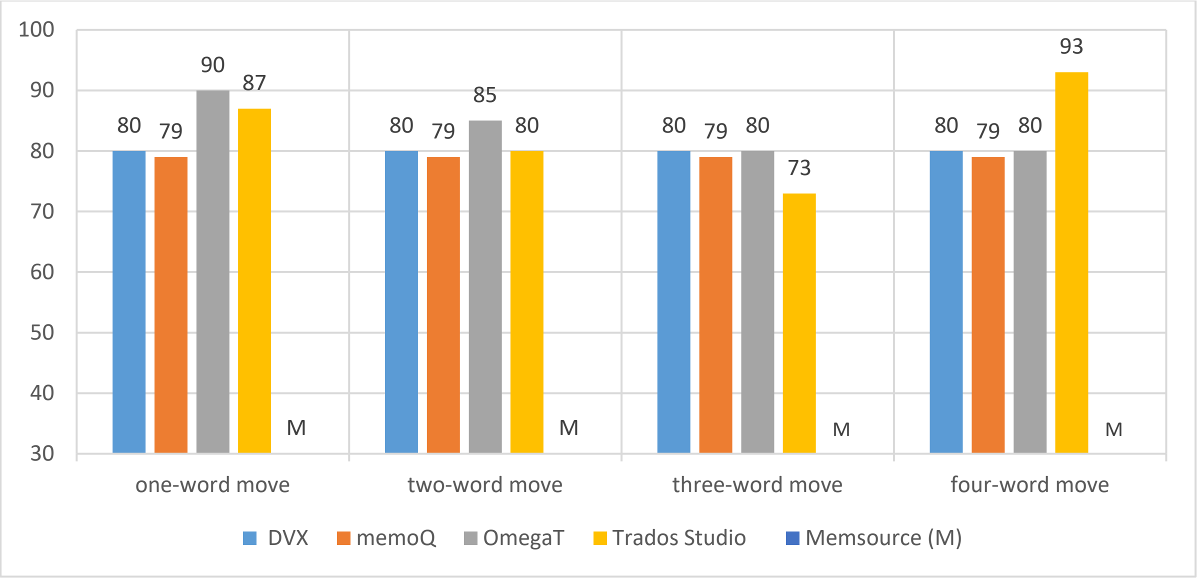


Figure 4.1.7: The TM retrieval of 10-word segments with a different unit move[53]

As Figure 4.1.7 shows, the TM systems used different algorithms to calculate matches, including Memsource, which always produced inconsistent scores. This is evidence that the different tools have different ways of handling such unit moves, although none is completely satisfactory.

4.1.3 Discussion of results

Based on the results of the experiment, this study has concluded that only longer routine segments containing a move (reordering operation) are likely to be presented as TM proposals, while short segments which include a reordering operation do not benefit from the use of any of the TM tools tested. A comparative assessment of the retrieval of low-scoring matches was accomplished by using the length of each segment and the unit-move string as independent variables. Potential reasons for the retrieval of low-scoring matches and the strings of units as independent follow.

[53] Memsource inconsistent scores (M)

A possible explanation for the production of low-scoring matches is that the TM systems' algorithms did not recognise the move intervention as such. It appears that they used a procedure of calculating strings of surface forms. In short segments, reversing words one and two in a three-word segment, for example, gave the following results: DVX provided a 34% match; Trados Studio, a 54% match; and OmegaT, a 66% match. This may be explained by the fact that the estimation of DVX's algorithm was approximately ⅔ non-similar, ⅓ identical. Likewise, Trados Studio's algorithm's assessment was roughly ⅔ non-similar, ⅓ identical. Regarding OmegaT, the algorithm inversely calculated the scores (⅔ identical, ⅓ non-similar). In longer segments, the reversal of words one and two in a nine-word segment produced the following scores: DVX provided a 77% match; Trados Studio, an 85% match; and OmegaT, an 89% match. Thus, it appears that OmegaT regarded a one-word move in a three-word sentence as an intervention on ⅓ of the sentence length, while DVX and Trados Studio regarded such a move as an intervention on ⅔ of the sentence length. The implications for Arabic users of OmegaT is that only those segments of four words or longer that include a one-word move will show as TM proposals.

In terms of memoQ's scores for the three- and four-word segments, the matches seem to decrease as the total number of characters decreases regardless of how many words the segment contains. For instance, segments of 16, 19 or 22 characters provided matches of 62%, 67% and 71%, respectively. Regarding Memsource's scores, the fuzzy scores appeared to rely on the number of characters. This outcome renders their TM retrieval systems unhelpful when retrieving Arabic segments containing a sub-segment move.

With regard to the unit-move string, it appears that some algorithms considered the space between words as a discrete value, while others did not. This resulted in a variation of scores for the same length of segments. Thus, a unit that included a two-word string was regarded as a disconnected string as there was one space between the words, and it was treated as two separate words (i.e. two edit operations). Likewise, in a unit that included a three-word string, the TM algorithms calculated that there were two spaces. This may be the reason why Trados Studio and OmegaT treated the move of multiple-word units differently from single-word moves – a longer unit move attracted a heavier penalty. In a six-word segment, for example, Trados Studio produced a 77%, 66% and 54% match, depending on whether it contained a one-word move 1WM, 2WM or 3WM, respectively, while OmegaT gave a match of 83%, 75%, and 66%, respectively. In contrast, DVX produced a single match score of 66% for either a 1WM, 2WM or 3WM in a six-word segment.

Further, Trados Studio dealt with segments with a four-word unit move in a different way, resulting in a very high match. Trados seemed to deal statistically with a four-word unit move as one chunk. In some ways, this was similar to the method used by OmegaT, which performed well with a one-word move. However, the implementation of the one-chunk treatment (i.e. four-word move) in Trados outperformed the implementation of the one-chunk treatment (i.e. one-word move) in OmegaT. This may be explained by the fact that the calculation of Trados Studio is based on 30% (the lowest match threshold it accepts), while OmegaT's calculation is based on 0% (the system allows 0% to be set as the lowest match threshold).

To see the effect of changes in translation threshold according to the suggestion of the Wolff et al study(2016), the threshold was decreased from 70% to 65%. This allowed increase suggestions retrieved from the TM. As expected, with a lower threshold more suggestions were provided. In OmegaT for example, the matches of three-to-ten-word routines including a one-word move were returned above the translation threshold, ranging from 66% to 90%.

To summarise, all the TM matching algorithms failed to recognise the different word order. The TM systems' matches obtained seem rather inappropriate, since the meanings of the segments are identical. It is therefore not useful for the TM-equipped translator to translate segments from scratch while identical source segments with translation unit are found in the TM database. An appropriate fuzzy matching would have been, either in the band of nearly exact match (ranging from 95% to 99%), or with a standard word order' penalty. If this existed, it could be applied only to Arabic and other free word-order languages.

This outcome is in line with the results of the study conducted by (Baldwin 2009), which highlighted the effect of Japanese word-order variation on the matching mechanism. The current study has provided experimental evidence in Arabic-to-English translation, gathered from the scores supplied by five CAT applications, showing that TM matching metrics are not good at handling the phenomena of the flexible word order.

Generally speaking, the retrieval of a multiple-word unit as an undifferentiated block, as well as the treatment of the move operation as a single, not multiple, intervention, could pave the way for TM developers to design a mechanism that would identify a sub-segment move, allowing them to improve TM recall matching for segments with move operations. The study suggests a standard penalty between 1% to 5% that would apply to free word-order languages.

4.1.4 Evaluation of results

4.1.4.1 Recall and precision

The fundamental measures of the performance quality of TM retrieval mechanisms are recall and precision. The results of the experiment clearly show that the recall of short segments containing a move operation is lower than it is for long segments, but the usability of these recalled segments in terms of translation is very high. Hence, the precision is higher than the recall.

4.1.4.2 Lost usability opportunity

In this study, 'usability' refers to the extent to which a user can achieve a certain goal with a given translation unit. The experiment's findings show that although the key function of a TM system is to handle repetition, the translator may miss out on the potential re-use of their previous translation in cases containing a move operation. The translator could be forgiven for expecting that in such cases a TM algorithm would retrieve all segments (both short and long) with a high match value, due to their identical semantics, but different order, with the TM source. In contrast to this expectation, however, it appears that a translator working with short segments will not be shown a good match and will therefore be deprived of one of the major benefits of TM – consistency of translation. As a result, they would be forced to re-translate the segment from scratch, effectively creating a new translation.

From the perspective of a project manager (whose remit includes such tasks and responsibilities as managing project time and delivering price quotations), such an outcome could negatively affect a text's preparation for translation. It would also have an economic impact: instead of paying a lower price for the translation of source segments that are found in the TM database, which are generally offered in the TM translation proposal window, clients would be charged more, as the translator would have to translate the text manually. In such cases, neither the translator nor the project manager, nor ultimately the client, would benefit from the TM function.

Another weakness surfaces when the TM is asked to handle the move of a consecutive multi-word unit. It might be expected that a TM system would deal with a sub-segment move as an undifferentiated block since the string is moved as a single chunk in the test segments. The results reveal, however, that some TM systems handle the multi-word-string move as if the text were a mixture of discrete words and a block. It was found that systems such as OmegaT and

Trados Studio have a TM similarity algorithm that considers a consecutive multi-word move as a string of discrete words, which does not help when computing a high match for a segment containing such a sub-segment move. In contrast, the same two systems performed very well when OmegaT retrieved segments with a one-word move and when Trados Studio treated a four-word move as one chunk. This is another reason that may explain why TM systems calculate low matches for short segments. Conversely, by handling a consecutive multi-word-string move as an undifferentiated block, as DVX does, the TM system could calculate a higher match for a segment which includes a sub-segment move.

4.1.5 Conclusion

The conclusion that can be drawn from the above experiment is that all the TM matching metrics tested displayed an inability to recognise the move operation, negatively affecting the retrieval of short segments and returning inappropriately low scores. These scores appeared to be based on the string of surface forms and the internal matching mechanism of each system's algorithm. Move strings of different lengths were treated either as multiple-word units or as blocks. Consequently, short segments of Arabic that contained a sub-segment move scored lower, while the longer segments scored higher. As a result, the translators would not be made aware of these proposals; they would be forced to re-translate the segment from scratch, effectively creating a new translation. The study suggests that the retrieval of a multiple-word unit as an undifferentiated block, as well as the treatment of the move operation as a single, not multiple, intervention, could pave the way for TM developers to design a mechanism that would identify a sub-segment move, allowing them to improve TM recall matching for segments with move operations. Alternatively, translators would reduce the translation threshold.

Further work is needed to compare the fuzzy matches produced by the TM similarity scores for Arabic segments that contain a move operation but have an identical meaning with those produced for the intervention of three edit operations. The following section explores whether a move operation is dealt with as an edit operation or not.

Section B:

4.2 A comparison of translation memory retrieval for segments including an edit operation with segments including a move operation: Arabic into English

Section B compares the matches retrieved from segments that include an edit operation with the matches retrieved from those with a move operation. It begins with an introduction to the investigation (4.2.1), describes the experimental setup (4.2.2), summarises the findings (4.2.3), then discusses (4.2.4) and evaluates (4.2.5) the results. Finally, it presents the conclusions drawn from the study (4.2.6).

4.2.1 Introduction

TM systems also offer proposals of segments that contain some measure of dissimilarity between the TM sources and the input: if a close match is found, it is displayed in the translation window, meaning that translator can use three types of edit operations to adjust the proposals: addition, deletion or substitution (Bulté 2018). However, there is little information about the size of penalty applied to each type of edit operation (i.e. the distance), raising the question of whether all systems apply the same penalties. According to Somers (2003), the distance is normalised into a score depending on the length of the strings.

The results in the previous section show that heavy penalties are imposed when there is a move operation (reordering), especially in shorter routines, even if the segments are semantically identical. This finding motivated the following investigation into the retrieval of segments that include a different dissimilarity such as an edit operation (i.e. deletion, addition or substitution).

In this experiment, four source segments (i-iv) were translated into a target language with the help of the TM source: segments (i), (ii) and (iii) have a one-edit operation at the beginning of the segment, while segment (iv) has a move operation. Table 4.2 (below) displays an example of the four similar input segments including a different dissimilarity (addition, deletion, substitution or move operation), and it would be expected to produce a closer TM proposal.

TM translation unit		**Input**	**Difference between the TM source and input**
أمريكي يستغل ثغرة قانونية في موقع تويتر Transliteration: *amrikiin yustaghalu thughrat qanuniat fi mawqie tuytr* English translation: An American exploits a legal loophole on Twitter.	i	شاب أمريكي يستغل ثغرة قانونية في موقع تويتر	The TM proposal needs to be edited by adding a one-word string: the word شاب / *shabun/ 'an adult'* is not found in the TM proposal
	ii	~~أمريكي~~ يستغل ثغرة قانونية في موقع تويتر	The TM proposal needs to be edited by deleting a one-word string: the word of أمريكي / *amrikiin / 'An American'* is an extra word in the TM proposal
	iii	ياباني ~~أمريكي~~ يستغل ثغرة قانونية في موقع تويتر	The TM proposal needs to be edited by replacing one word with the same string: the word أمريكي is in the TM source string while the word ياباني /yabaniin/ 'Japanese' is in the input string.
	iv	~~أمريكي~~ يستغل أمريكي ثغرة قانونية في موقع تويتر	The TM proposal needs to be edited be moving a one-word string: the position of the word يستغل. exchanges with the position of the word أمريكي .

Table 4.2: An example of generating four similar segments against a TM source in Arabic[54]

[54] Track Changes was used for the intervention.

Theoretically, each segment has a different type of dissimilarity in comparison to the TM source. In Table 4.2, the TM source أمريكي يستغل ثغرة قانونية في موقع تويتر / *amrikiin yustaghalu thughrat qanuniat fi mawqie tuytr* / differs from the four test segment as follows:

1) Arabic segment (i) includes one extra word (شاب / *shab*) compared with the TM source, and does not correspond with the TM source in meaning and length (it is longer). As a result, the potential TM proposal needs a one-word addition.
2) Arabic segment (ii) has a one-word omission (أمريكي / *amrikiin*) compared with the TM source, and does not correspond to the TM source in meaning and length (it is shorter). As a result, the potential TM proposal needs one-word deletion.
3) Arabic segment (iii) includes a one-word difference with the TM source, using أمريكي / *amrikiin* instead of ياباني /yabaniin/ 'Japanese', and corresponds to the TM source in length only, while the meaning is slightly different. As a result, the potential TM proposal needs a one-word substitution.
4) Arabic segment (iv) has the same semantic content as the source words and length as the TM source but includes a one-word move, reversing the order of the words أمريكي / *amrikiin* and يستغل / *yustaghalu*.

The question here is whether the TM algorithm gives the higher fuzzy score to the segment with a move operation or the segment with an edit operation, and whether the move operation is treated similarly to one of the three edit operations or in a different way. In order to answer this question, this section compares the retrieved matches provided for the three-operation edit with the matches obtained for the move operation.

4.2.2 Experimental setup

This experiment was carried out using the MeedanMemory corpus, with a subset of 15 Arabic source segments (ranging between five and nine words in length) taken from the test content of the experiment in section A; 3- and 4-word routines were excluded from this experiment since they provided inconsistent scores in memoQ. The three types of intervention (adding, removing, and substituting a one-word string) were applied to the original segments in addition to the move operation. The intervention targeted the first word in each segment:

1) A one-word string was *added* at the beginning of segments 5-10 words in length.
2) A one-word string was *removed* from the beginning of segments 5-10 words in length.

3) A one-word string was *replaced* by a different one at the beginning of segments 5-10 words in length.
4) The position of word one was *reversed* with that of word two at the beginning of segments 5-10 words in length.

It is worth noting that the edit-operation intervention sometimes distorted the meaning of a segment, but this was ignored as the aim here was to look at the TM similarity measurements.

These interventions generated four similar segments for the original segment – a total of 60 test segments (see Appendix Four for the test segments B). The test segments were then imported as a translation document into the five selected CAT tools, while the corpus was imported as a TM. Finally, the pre-translation function was used to review the retrieved matches, under a 30% value as the minimum translation threshold.

The findings associated with each operation through a new match value were assessed in order to ascertain the fuzzy match assigned to each type of intervention.

4.2.3 Findings

The TM algorithms provided different matches for segments with four different types of dissimilarity. The results obtained from testing the five selected TM systems are provided below. Column charts are used to represent the retrieved scores, in which the four different types of dissimilarity are represented on the 'x' axis and the fuzzy matches on the 'y' axis. Further, the TM systems are presented in alphabetical order.

4.2.3.1 DVX output

When the TM of DVX was tested, the segments with the one-word move provided the lowest match, while those with edit operations provided higher matches. The overall scores decreased gradually as the number of words decreased, whichever the type of intervention. The matches of five-to-nine-word routines ranged from 60% to 90%. Figure 4.2.1 (below) shows the retrieved matches that each segment length (SL) was assigned across the four different types of dissimilarity.

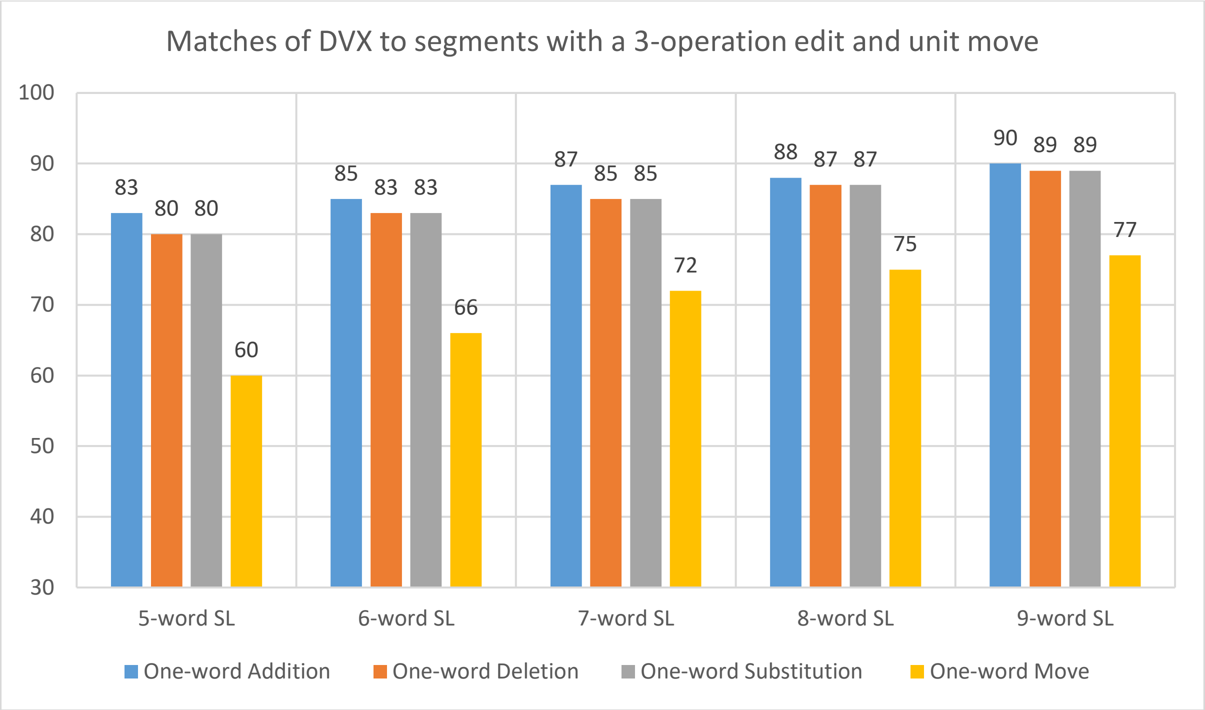


Figure 4.2.1: DVX's matching scores for retrieving segments with four different types of dissimilarity

The figure above clearly shows that the lowest retrieved matching was given to segments with a one-word move, while both a one-word deletion and a one-word substitution matched equally, and a one-word insertion scored the highest match. In the retrieval of five-word segments, for example, the percentage score for moving a one-word string was 60% (i.e. a 40% penalty), while either removing or substituting a one-word string scored an 80% match (i.e. a 20% penalty) and inserting a one-word string scored an 83% match (i.e. a 17% penalty). This means that the move operation was penalised heavily in comparison with the other three one-word operations. Also, the matches retrieved for the three-operation edit were very close to each other, with the addition operation assigned the smallest penalty, while the percentage of matches retrieved for the move operation was lower.

It is evident that these results strongly suggest that the algorithm of DVX penalised the dissimilarity caused by the move operation more heavily than the dissimilarity caused by an edit operation. As a result, the TM's users may see proposals that need an edit operation in higher fuzzy matches than those that have an identical meaning but contain a word move.

4.2.3.2 memoQ output

The matching scores obtained by memoQ also gave the lowest matches to the one-word move, while the retrieved segments including an edit operation received better matches. The matches dropped steadily as the segment length became shorter. The matches of five-to-nine-word routines ranged from 65% to 89%. Figure 4.2.2 (below) displays the retrieved matches assigned to each segment length (SL) according to the four types of dissimilarity.

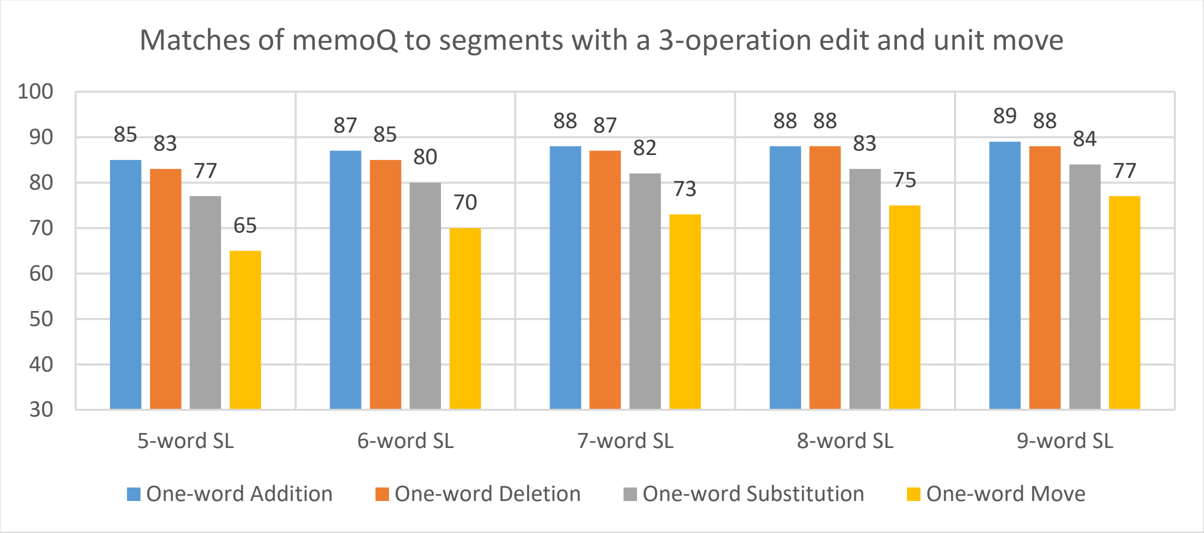


Figure 4.2.2: memoQ's matching scores for retrieving segments with four different types of dissimilarity

Figure 4.2.2 shows that the retrieved segments which included a move operation were assigned the lowest matches, followed by those with a substitution and then those with a deletion operation, while an insertion operation was assigned the highest match. For example, with the five-word segment, the score for moving a one-word string was 65% (i.e. a 35% penalty), while the scores for either adding, removing or substituting a one-word string were 83%, 85% and 77% (i.e. a 17%, 15% and 23% penalty), respectively. This means that memoQ also imposed heavier penalties on the move operation than the three-operation edit. Furthermore, the matches provided for the three-operation edit were relatively close to each other. The addition intervention received the minimum penalty, while the match of the move operation was assigned the heaviest penalty.

As the memoQ similarity algorithm penalises the dissimilarity caused by a move operation more heavily than that caused by the three-operation edit, if proposals with an edit operation are found, they will be offered above proposals with a move operation, which will consequently sink to the bottom of the list of matches.

4.2.3.3 Memsource output

The matching scores of Memsource were derived in a different way, leading to inconsistent scores in the test results. The matches varied according to the type of intervention and were based on the total number of characters in the segment, regardless of how many words the segment contained. Due to the inconsistency of the matches, the key results were summarised by sorting the length of the test segments according to the rule of the longest first. Table 4.3 (below) shows the longest segment in each intervention type, accompanied by its Memsource score.

String length and matching	**One-word move**	**One-word addition**	**One-word deletion**	**One-word substitution**
Longest string by character	61 character	66 character	54 character	61 character
Matching score	84%	92%	88%	90%
Shortest string by character	31 character	36 character	26 character	31 character
Matching score	71%	86%	83%	86%

Table 4.3: Examples for the longest and shortest string in each intervention type with their Memsource matches

The matching values in the table above reveal that the move operation was assigned the heaviest penalties as both the longest and shortest examples show. With a deletion operation, which potentially comprised a shorter string, a segment of 54 characters containing a one-string deletion was given an 88% match, whereas a segment of 61 characters including a one-string move was assigned a lower match, 84%. A similar result was noticed with the shortest string: a segment of 26 characters with a one-string deletion was given an 83% match, whereas a segment of 31 characters including a one-string move was given a lower match, 71%. With the substitution operation, which potentially shared the same string, a 90% match was given for a segment of 61 characters when substituting a one-string unit but only an 84% match when moving a one-string unit. Similarly, with the shortest string, a segment of 31 characters, substituting a one-string unit provided a match of 86% but moving a one-string unit provided a match of only 71%. Among all the types of dissimilarity, it appears that the dissimilarity in the addition operation was given the lightest penalty but the highest match. It also appears that Memsource penalised the dissimilarity caused by a move operation heavily compared with its treatment of the three-operation edit.

4.2.3.4 OmegaT output

Viewing the OmegaT TM results, it can be seen that the operation of moving a one-word string was treated comparably to deleting a one-word string – both operations occupied the lowest level of matching – while the operation of substituting a one-word string or adding a one-word string obtained higher matches. Further, the matching scores decreased as the length of the segments decreased, whichever the intervention. The matches of five-to-nine-word routines ranged from 80% to 94%. Figure 4.2.3 (below) presents the retrieved matches provided for each segment length (SL) according to the four types of dissimilarity.

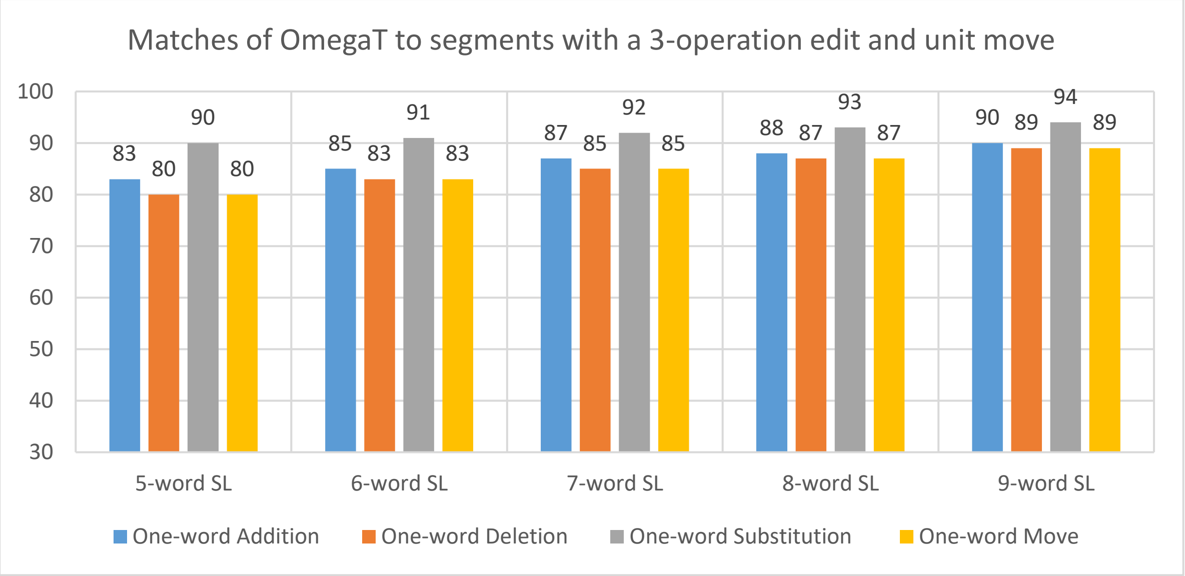


Figure 4.2.3: OmegaT's scores for retrieving segments including four different types of dissimilarity

Figure 4.2.3 shows that a one-word move and deletion were assigned the lowest scores, and a one-word addition achieved higher matches, while a one-word substitution achieved the highest match. In the five-word segments, for example, the score for either moving or deleting a one-word string was an 80% match (i.e. a 20% penalty) for each, while the scores for adding and substituting a one-word string were 83% and 90% (i.e. a 17% and 10% penalty), respectively. This suggests that the system dealt with the move operation in the same way as one type of edit operation: the matches given to a segment with a move operation fell into the same range as the matches given to a segment with an edit operation, although the deletion operation achieved the lowest match.

Thus, the four different types of dissimilarity were given close matches, although the move operation was assigned the lowest. The TM proposals that needed a one-word substitution matched the highest, followed by those that included in an extra one-word string.

4.2.3.5 Trados Studio 2019 output

Among the matching scores produced by Trados Studio, the lowest scores were those for the one-word move, while the highest matches were given to segments with a three-operation edit. The matching values decreased gradually as the number of the words decreased, whichever the intervention type. The matches of five-to-nine-word routines ranged from 73% to 93%. Figure

4.2.4 (below) shows the matches retrieved for each segment length (SL) according to the four types of dissimilarity.

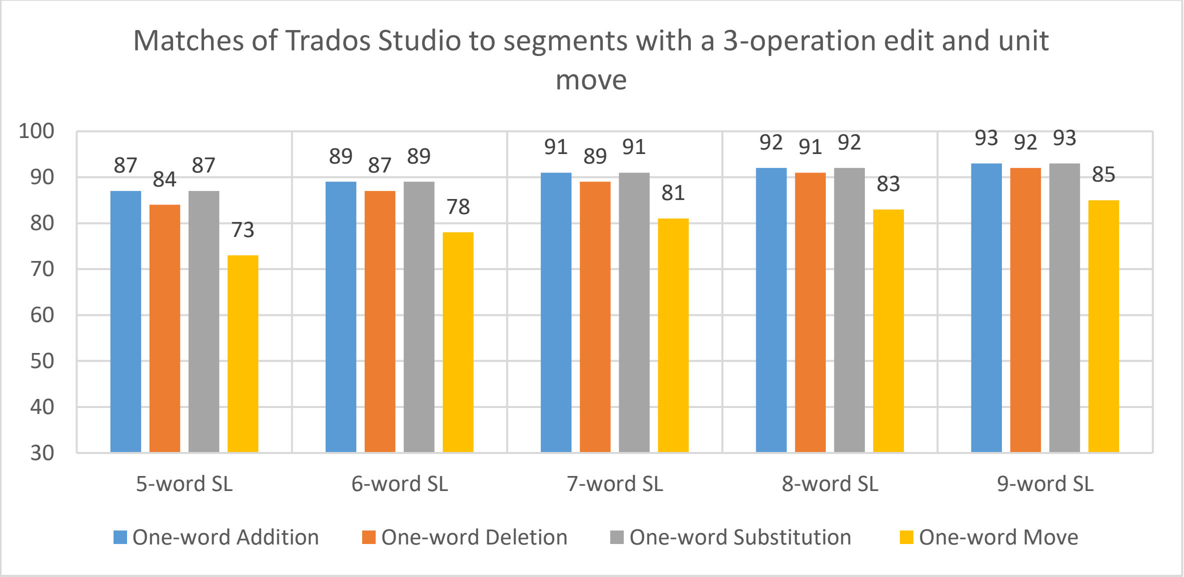


Figure 4.2.4: Trados Studio's matches for retrieving segments including four different types of dissimilarity

The figure above shows that the worst match was given for a segment retrieval containing a move operation, followed by those containing a one-word deletion, while segments that included a one-word addition and substitution provided the highest matches. For example, in the case of five-word segments, a one-word move provided a 73% match (i.e. a 27% penalty), while the scores provided for adding, removing or replacing a one-word string were 87%, 84% and 87% (i.e. a 13%, 16% and 13% penalty), respectively. This means that the move operation acquired the heaviest penalties. The study also noted the close proximity of the scores for the matches of the three-operation edit compared with the match of the move operation, which was assigned the lowest percentage.

The results of Trados Studio appear to suggest that its algorithm penalised the dissimilarity caused by the move operation more heavily than that caused by an edit operation. The lowest match returned to move operations confirmed the Trados Studio's results of the experiment in Chapter Four, section A (see 4.1.2.5).

Generally speaking, the TM algorithm offered proposals of segments with an edit operation in a higher fuzzy match, whereas a move operation, despite the segments' identical meaning, would be assigned a lower match.

4.2.4 Discussion of results

The results demonstrate that all the selected TM systems, except OmegaT, measured the dissimilarity caused by a one-word operation lower than that caused by a three-operation edit. OmegaT, although it dealt with a one-word operation in a similar way to the one-word deletion, achieved the lowest scores. The overall results appear to concur with those in Cormode and Muthukrishnan's (2007) investigation. These researchers explain that the low matching scores caused by a move operation are probably due to the fact that a TM algorithm treats a move operation as a deletion followed by a re-insertion elsewhere. In addition, the edit distance operations applied penalties differently.

Reviewing the overall results, it appears that, apart from the OmegaT scores, the retrieved matches of the three-operation edit were always given a close matching range, while the range of matches for the move operation was lower. This suggests that the TM matching metrics are only able to deal with an intervention of a three-operation edit; they are not set up to look for move operations. This is further evidence that the TM systems have little information regarding the retrieval of a move operation, which correlates with the outcome of the experiment in section A.

It also appears that the consistency of matching scores, excepting those of Memsource, was dependent on the segment's word length (six-to-nine-word sentences), regardless of the type of intervention – i.e. the matching scores decreased as the length of the segments decreased, while the scores increased as the length of the segments increased. In regard to the matches of Memsource, although the scores were inconsistent whichever the type of intervention, the results seemed to follow the same method. This was seen when sorting the results according to the longest segment first (by character): the longest segments were given the highest match and the shortest, the lowest match. These results also correlate with the findings in section A, which suggest that the Memsource measurement is based on the surface forms of segments.

It was also observed that the priority match of the four different dissimilarities differed from one system to another but none of the systems selected the segment with a move operation for the highest score. Under the procedures used in the experiment, DVX, memoQ and Memsource ranked the TM proposal of segments containing an insertion operation at the top, OmegaT gave the segment containing a substitution operation the highest match, while Trados Studio ranked the segment containing either an insertion or substitution operation highest.

4.2.5 Evaluation of results

The results show that the TM systems selected for testing gave the retrieved matches of the segments containing a move operation, such as the reordering phenomenon in Arabic, the lowest fuzzy matching scores. However, the segments that needed at least one edit operation to produce the desired translation achieved higher matches.

4.2.5.1 Recall and precision

The measurement of the recall and precision of these segments shows that the recall of segments containing an edit operation was higher than it was for those with a move operation; however, the usability of these recalled segments in terms of translation was very low. Hence, the results suggest that precision was lower than recall for the retrieval of the three-operation edit but precision was higher than recall for the retrieval of a move operation.

4.2.5.2 Lost priority opportunity

The results obtained by comparing the matches of segments that included a move operation with those that included a three-operation edit show that the TM systems placed matches of the move operation in the lowest bands, while matches of the three-operation edit were assigned a higher band. This meant that the priority match of TM proposals were given to segments containing an edit operation, although the segments that included a move operation were in fact more similar in both surface form and meaning.

As a consequence, these results may have an impact on the fuzzy match grid. This grid is a method for calculating price discounts on fuzzy matches. Typically, translation agencies have a scheme for discounting based on fuzzy match levels: if segments including a move operation are given a discount for fuzzy match bands, the analysis report would show the matches of the move operation in the lower fuzzy band, while the matches of the edit operation would fall in a higher band. This means that clients would not gain much help from the re-use of similar segments containing a move operation.

4.2.6 Conclusions

The results for the experiment investigating the retrieved matches of the four different types of dissimilarity reveal that all the selected TM systems, apart from OmegaT, treated the move operation differently from the three-operation edit. OmegaT treated it as the same as the deletion operation. However, this segment always matched the lowest score whichever the TM system. In addition, the retrieved matches of the three-operation edit were always given a close match range, while the match of the move operation was given a lower level. This is evidence that all the TM matching metrics, apart from OmegaT's, neither recognised the move operation as a specific edit type involving exchange of word positions nor treated it as a three-operation edit.

4.3 Summary of sections A and B

This chapter has investigated the performance of TM systems when retrieving segments that include a move operation (the reordering phenomenon in Arabic). It has also compared the results of the retrieval of segments that include a reordering operation with the retrieval of a three-operation edit. It could be argued that the different ways in which each TM system dealt with the reordering operation had different consequences. The outcome was that if segments contained a reordering operation, such as the reversal of the subject and the verb in the Arabic segments, their retrieval from the TM resulted in very low matches, especially for the shorter segments. This may be because the similarity measurement of the current TM algorithms is based on comparing the identity of the input string with that of the TM sources, yielding a high level of precision but a low level of recall. This was most notable in the results of the experiments in which the systems' criterion of measurement was computed using only the surface forms of the segments.

This outcome represents a weakness in the TM matching metrics that can affect the re-use of the TM proposals. Users may lose the maximum value of matches for lexically and semantically highly similar segments, presenting a real obstacle to the adoption of Arabic as a source text by TM systems since the reordering phenomenon is a crucial characteristic of the language. Arguably, the current TM systems are far from dealing properly with a move operation, supporting this study's hypothesis that TM matching procedures are unable to recognise such an operation. TM developers need to create an effective mechanism for handling a move operation if they are to make the re-use feature more useful. However, as the experimental study was conducted with Arabic-English translation, this claim cannot be generalised to encompass other languages, and further research examining other languages characterised by the flexibility of their word order, such as Greek, is recommended.

4.4 Suggestions for improvement

Given the different mechanisms used by the various TM systems' algorithms, the results of this study suggest that the integration of two aspects of these algorithms would produce better results for retrieving segments containing a move operation.

The OmegaT system measurement succeeded to some degree in retrieving segments with the one-word unit move at a usability level, except for the three-word segments (see Figure 4.1.3). This may be explained by the fact that it treated the one-word unit move (i.e. one chunk) as a one-edit operation rather than two, whereas the same system computed an intervention of the multiple-word unit move as multiple edit operations. Likewise, the other systems computed a move operation as two or more edit operations. If the multiple-word unit move could be treated as a disconnected string – the method used by the DVX and memoQ metrics – we could expect good results, apart from the three-word segments. This is because the calculation is based on using 0% as the minimum match threshold, as in OmegaT.

In a similar way, the Trados Studio system measurement succeeded in retrieving segments with a four-word unit move with high scores because the system dealt with the four-word unit move as one chunk (see Figure 4.1.4). This means that one chunk was treated as a one-edit operation. If a one-chunk mechanism was applied to a one-word, two-word and three-word unit move, the usability could be expected to be very high, especially with the Trados Studio calculation, which is based on a 30% value (not 0% as in OmegaT) as the minimum match threshold.

If the TM developers applied this suggestion of treating a move operation as a one-edit operation and multiple-word unit moves as one chunk, and calculated the numerical matching using 30% as the lowest match threshold, this could potentially increase the fuzzy matches regardless of the length of segment and of the string unit move, thus presenting the translator with a reusability opportunity. For example, the three-word segments, which scored the lowest match in the experiment, could potentially be assigned a 76% or 77% match.

CHAPTER FIVE

TRANSLATION MEMORY RETRIEVAL AND MORPHOLOGY FEATURE: SECTIONS A AND B

5.0 Introduction

Chapter Five evaluates the similarity measurement approach adopted by the TM systems of five different computer-aided translation (CAT) tools in the retrieval of morphological features in Arabic-to-English translation. The chapter is organised as two main sections: section A investigates the performance of TM systems when retrieving segments with inflectional verb-variations, while section B examines the retrieval of segments that include a character-marker omission (Hamza marker).

Section A:

5.1 A comparison of the word similarity measurement in Arabic-English TM segment retrieval with an inflectional affix intervention

Section A takes as its subject of investigation the matching measurement approach of TM systems when retrieving inflectional verb-variation segments in Arabic-to-English translation. The section describes the experimental setup of the investigation (5.1.1), summarises its findings (5.1.2), and then discusses (5.1.3) and evaluates the results (5.1.4). Finally, it presents the conclusions drawn from its findings (5.1.5).

The question the study sets out to answer is whether, when retrieving segments, the TM algorithm measures a combination of the inflectional affixes as a word or as a character intervention. Its hypothesis is that if TM systems have some linguistic knowledge, the penalty will be very low, thereby decreasing the cost of the translation.

5.1.1 Experimental setup

The experiment was carried out using an Arabic-to-English TM: 60 Arabic-to-English aligned segments were extracted from R.TM (see 3.1.5.2) ranging between three and seven words in length (see Appendix Five for the test segments C). Arabic was the source language of the translation units, which included a combination of inflectional affixes, and English was the target language.

The variables applied in the test segments were as follows:

- the routine of segments ranged from three to seven words in length;
- the root form of each verb comprised a three-character root;
- the verb stem was combined with a single character as a prefix or suffix;
- the verb stem was one of four verb-stem templates representing the inflectional verb variations.

In each event, at least three samples were used to verify the results, meaning that the same match was repeated in the three different samples

5.1.1.1 Inflectional affix transformation

Once the test segments had been extracted from the TM, the verb stems of the four templates were transformed from one form into another (i.e. from perfective to imperfective or vice versa) by changing their inflectional affix. The change of character led to a change in the verb tense only; the aspects of the subject remained the same. The rules of transformation applied to the experiment are explained below using the example of the canonical verb فعل (do), which is commonly used by Arabic grammarians to create verb templates.

- Rule 1: The verb template (VT) of the source segment was changed from an imperfective (third person masculine) into a perfective pattern: يفعل (He does)> فعل or فعلَ(He did) . The transformation was made by dropping an initial character يـ (a single-character prefix), or sometimes by adding a diacritic mark on the final-character لَ. However, the insertion of a diacritic mark is optional in Arabic, and it may be omitted from the text.

- Rule 2: In contrast to Rule 1, the verb template was changed from a perfective (third person masculine) into an imperfective pattern, فعل or فعلَ(He did) > يفعل (He does) , by adding an initial-character يـ (a single-character prefix).

- Rule 3: The verb template of the source segment was changed from a perfective (third person feminine) into an imperfective pattern, فعلت (She did) > تفعل (She does), by changing a final character ـت (a single-character suffix) into an initial character تـ (a single-character prefix).

- Rule 4: In contrast to Rule 3, the verb template was changed from an imperfective (third person feminine) into a perfective pattern: تفعل (She does) > فعلت (She did). The change was made by changing an initial character تـ a (single-character prefix) into a final character ـت (a single-character suffix).

Using an Arabic verb conjugator website,[55] the automated ACON application can conjugate the different templates of the Arabic verb by selecting the root and the type.

Table 5.1 (below) illustrates one example of the transformation process in all four cases.

[55] ACON, the Arabic Conjugator - conjugate Arabic verbs online (baykal.be)

R.	Original VT	M.I.	Transformed VT	G.C.
1	**يجمع** المزارع الثمار الناضجة. /yajmae almazarie althimar alnaadijat./ The farmer **collects** ripe fruits.	Dropping prefix	**جمع** المزارع الثمار الناضجة. /jame almazarie althimar alnaadijat./ The farmer **collected** ripe fruits.	Changing sentence tense from present into past (gender> masculine)
2	**قرأ** الطالب في كتابه. /qara altaalib fi kitabih/ The student **read** his book.	Adding prefix	**يقرأ** الطالب في كتابه. /yaqra altaalib fi kitabih/ The student **reads** his book.	Changing sentence tense from past into present (gender> masculine)
3	**طبخت** الأم وجبة الغذاء. /tabkhat al'uma wajabat alghadha'./ Mother **cooked** lunch.	Shifting suffix to prefix	**تطبخ** الأم وجبة الغذاء. /tubikhat al'uma wajabat alghadha'. Mother cooks lunch.	Changing sentence tense from past into present (gender> feminine)
4	**تحذر** الحكومة من احتمال العودة إلى الإغلاق /tahadhar alhukumat min aihtimal aleawdat 'iilaa al'iighlaq/ The government **warns** of a possible return to lockdown	Shifting prefix to suffix	**حذرت** الحكومة من احتمال العودة إلى الإغلاق /hadharat alhukumat min aihtimal aleawdat 'iilaa al'iighlaq/. The government **warned** of a possible return to lockdown	Changing sentence tense from past into present(gender> masculine)

Table 5.1: Transformation of four verb templates in Arabic segments using edit operations

MI: Morphological intervention GC: Grammatical change

Each segment underwent a transformation, which converted linguistically the imperfective pattern of the verbs in the original segments into the perfective pattern, or vice versa, using one type of edit operation. Then, the test segments comprising the document to be translated were run against the TM corpus which included the original segments.

Table 5.1 (above) illustrates the verb templates' original structure, which shows the verb inflections in Arabic, and the transformation of the verb templates (the morphological intervention), revealing the language's rich morphology.

This raises the question of how the TM systems measure the similarity between two source strings: is the similarity measurement based on a word-by-word comparison or is the measurement compared character by character? For example, if two source segments are identical except for a difference in an inflectional affix, does the algorithm measure a combination of the inflectional affixes as a word intervention or a character intervention? If it is dealt with as a character intervention are the types of intervention penalised differently?

The TM matching metrics may compute inflectional verb variations as a word intervention, which means that the algorithm regards the inflected form as either a totally different word, in which case the penalty would be relatively heavy, or a character intervention, in which case the penalty would be based on the type of edit. Hence, it could be expected that the TM matching metrics would have difficulties detecting inflectional affixes, resulting in their omission from the list of TM proposals with high-scoring matches, even though segments with a minor modification were already available in the TM's storage.

5.1.1.2 Translating

Once the test segments were processed, they were imported into the five CAT tools as a translation file, while the translation project in each tool was based on the corpus which had been created as a TM file containing the original segments. The input text, consisting of 60 segments, produced fuzzy matches whose results were then analysed. As the test segments and the TM sources were identical, apart from the difference of an inflectional affix, the most desirable outcome would be for the TM matching metrics to produce a high fuzzy match that would appear at the top of the list of proposals presented to the translator.

To see which fuzzy match was assigned to each type of intervention, the test assessed the findings associated with each operation by a new match value.

5.1.2 Findings

The findings of the investigation reveal the five TM systems' attempts to retrieve matches using different percentages, as described below. It was based on the assumption that the range of high fuzzy matches (85%-95%) or higher are the best matches. Column charts are used to represent the retrieved scores, in which the inflectional affix combinations are represented on the ‘x’ axis and the fuzzy matches on the ‘y’ axis. Further, the TM systems are presented in alphabetical order.

5.1.2.1 DVX scoring

The matches retrieved by DVX were found to occupy a consistent band according to the length of the test segments and whether they contained an inflectional affix intervention (deletion, insertion or substitution). The matching scores decreased in a consistent way as the number of words in the segment decreased, and ranged from 67% to 86%. Figure 5.1.1 (below) illustrates the fuzzy matching scores that each segment length (SL) supplied due to their inflectional affix combination (i.e. the edit distance).

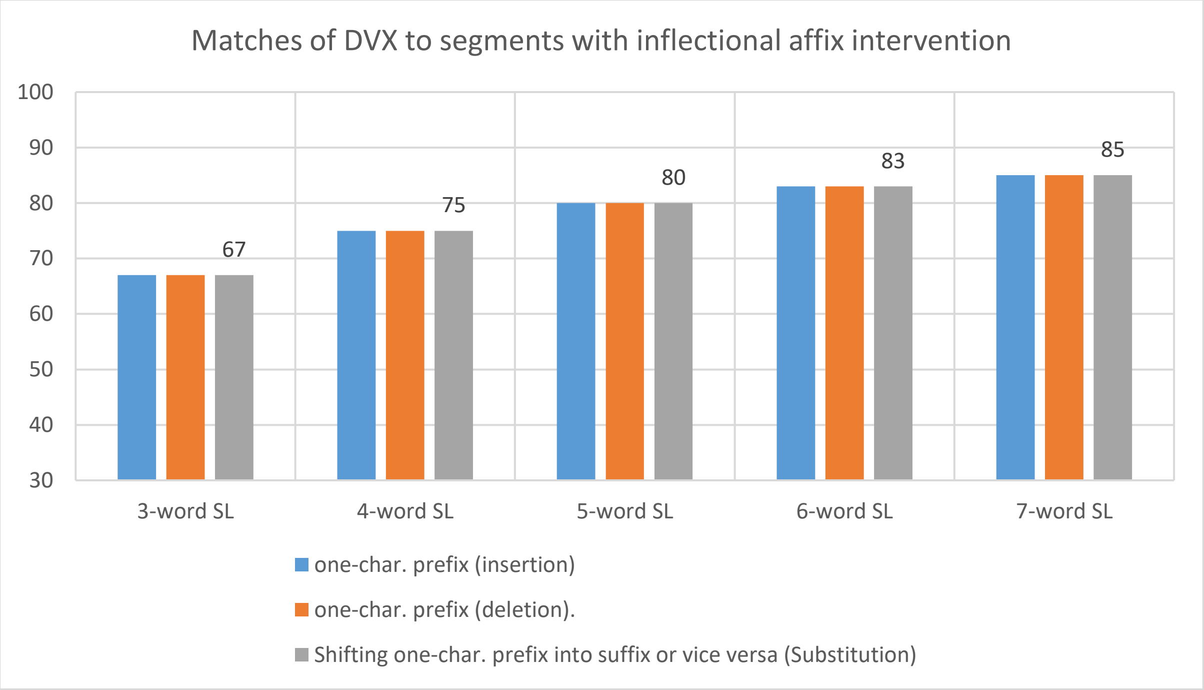


Figure 5.1.1: DVX matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention

The figure above clearly shows that DVX treated the test segments equally regardless of the type of inflectional affixes intervention. Further, the retrieved matches of three-to-seven-word segments were distributed among the different fuzzy bands. For example, 67% provided a low fuzzy score (i.e. a 33% penalty per one-edit operation), while for seven-word segments, 86%

provided a high fuzzy score (i.e. a 16% penalty per one-edit operation, or approximately one word in seven). This means that TM users may not see proposals of high fuzzy matches for short sentences that have just a single character difference.

5.1.2.2 memoQ 9.0 scoring

The matching scores of memoQ were derived from two different ranges: a low match range and a high match range. The five-, six- and seven-word segment routines were in the low fuzzy range, while the three- and four-word segments were given a relatively high fuzzy range whether these segments contained an inflectional affix intervention. The match scores ranged from 77% to 91%. Figure 5.1.2 (below) illustrates the different range of matches for each segment length (SL) provided.

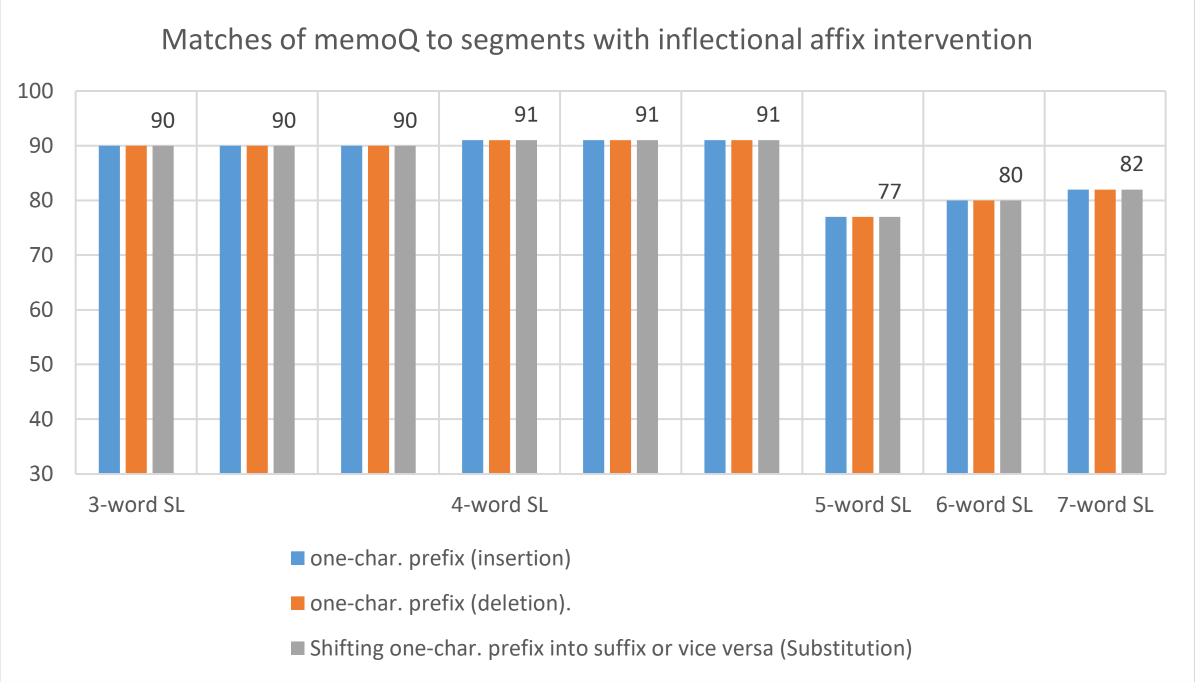


Figure 5.1.2: memoQ matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention

As the figure above shows, the matches of three- and four-word segments with an inflectional affix were retrieved in a high fuzzy match. For example, the three-word and four-word segments were provided with a 90% and 91% match, respectively (i.e. a 10% and 9% penalty). In terms of the segments of five words and above, the scores unexpectedly matched lower regardless of the edit operation. For example, five-word segments provided a match of 77% (i.e. a 23% penalty).

Based on memoQ's results described in Chapter Four, the retrieval of three- and four-word segments was based on the total number of characters, while the retrieval of segments of five words or above was based on the number of words. This may explain the difference in the matching levels: the character-based measurement produced considerably better results. As a result, the short segments would be offered in a high fuzzy band, while longer segments would be scored lower, although in all cases the difference was just a single character.

5.1.2.3 Memsource Cloud scoring

The TM system of Memsource retrieved the test segments in an inconsistent range of scores. The matches appeared to rely in the first place on the total number of segment characters and the string move, and in the second place on the position of the edit operation. Further, the match values decreased as the total number of characters decreased; the length of segments varied from 16 to 49 characters (i.e. both characters and whitespaces), while the match scores varied between 73% and 98%. Due to these scattered scores, the matches illustrated in Figure 5.1.3 are presented as a chart, using a line with markers: the markers represent the inconsistency of scores, while the lines represent the impact of the segment length.

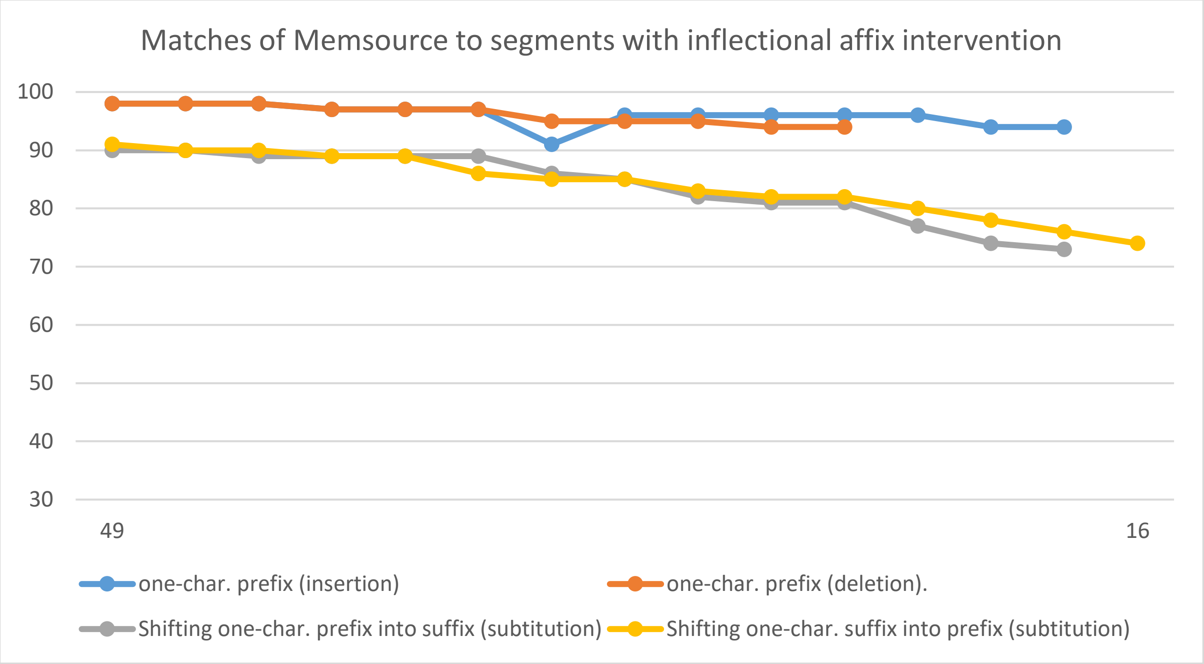


Figure 5.1.3. Memsource matching scores for a segment 49-16 characters long due to changes to an inflectional affix

As Figure 5.1.3 shows, it is obvious that the retrieval of segments with a one-character prefix (insertion or deleletion) were given high percentages, whereas the operation of shifting a one-character prefix into a suffix position, or vice versa, was assigned a lower fuzzy band. In terms of editing a one-character prefix, however, the matches were ranked in the range of nearly exact matches. For example, the matches of segments ranging from 49 to 16 characters, produced by inserting a one-character prefix, ranged from 98% to 94%, whereas segments ranging from 49 to 76 characters. produced by deleting a one-character prefix, also scored between 98% and 94%. Shifting a one-character prefix into a suffix position, or vice versa, produced match scores in the lower fuzzy band. For instance, segments ranging from 46 characters to 18 characters produced scores between 90% and 73% when a one-character prefix was changed into a suffix, whereas segments ranging from 46 characters to 19 characters produced scores between 91% and 74% when a one-character suffix was changed into a prefix.

The explanatory hypothesis is that, on the one hand, a one-character prefix was dealt with as a one-edit operation, while changing a one-character prefix into a suffix, or vice versa, was treated as a two-edit-operation. On the other hand, editing a one-character prefix occurred on the word-initial position, while changing a one-character prefix into a suffix, or vice versa, occurred on the word-initial and word-final positions. This suggests that the matching metrics dealt with the impact of a prefix combination in a different way to that of a suffix combination. As a result, the retrieval of segments with an inflection affix would be offered at a high fuzzy level under specific conditions.

5.1.2.4 OmegaT scoring

The fuzzy matches provided by OmegaT were relatively high; however, they dropped gradually as the segment became shorter, whether it contained a deletion, insertion or substitution operation. The matching scores consistently related to the segments' word length – the scores ranged from 83% to 92%. Figure 5.1.4 (below) shows the matching values for each segment length (SL) according to the editing of an inflectional affix.

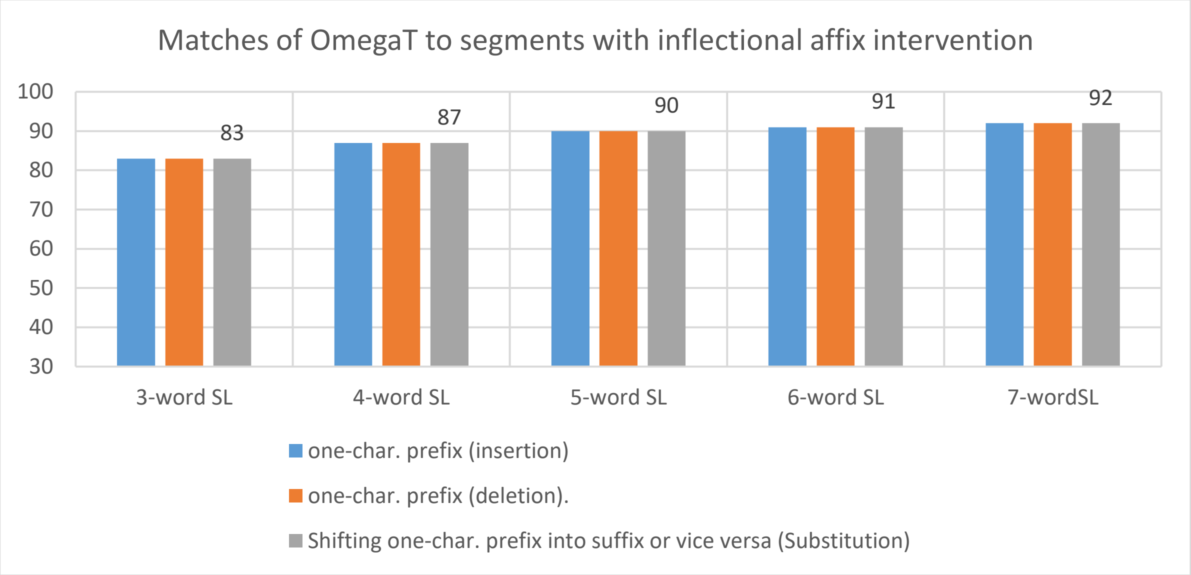


Figure 5.1.4: OmegaT matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention

As Figure 5.1.4 clearly shows, OmegaT's matching metrics dealt with the different ways of editing the inflectional affix in the same fashion, retrieving four- to seven-word segments in a high fuzzy band; only the three-word routine was placed in the middle fuzzy band. This means that OmegaT would retrieve segments with an inflectional affix – except for a three-word routine – in a high fuzzy band, which would be very useful from the perspective of translation costs.

5.1.2.5 Trados Studio 2019 scoring

The matching scores produced by Trados Studio also fell steadily as the segment length became shorter, whether these segments contained a deletion, insertion or substitution operation. The matching values were consistently related to the segment's word length. The match scores ranged from a 78% to 91%. Figure 5.1.5 (below) displays the matching values for the retrieval for each segment length (SL).

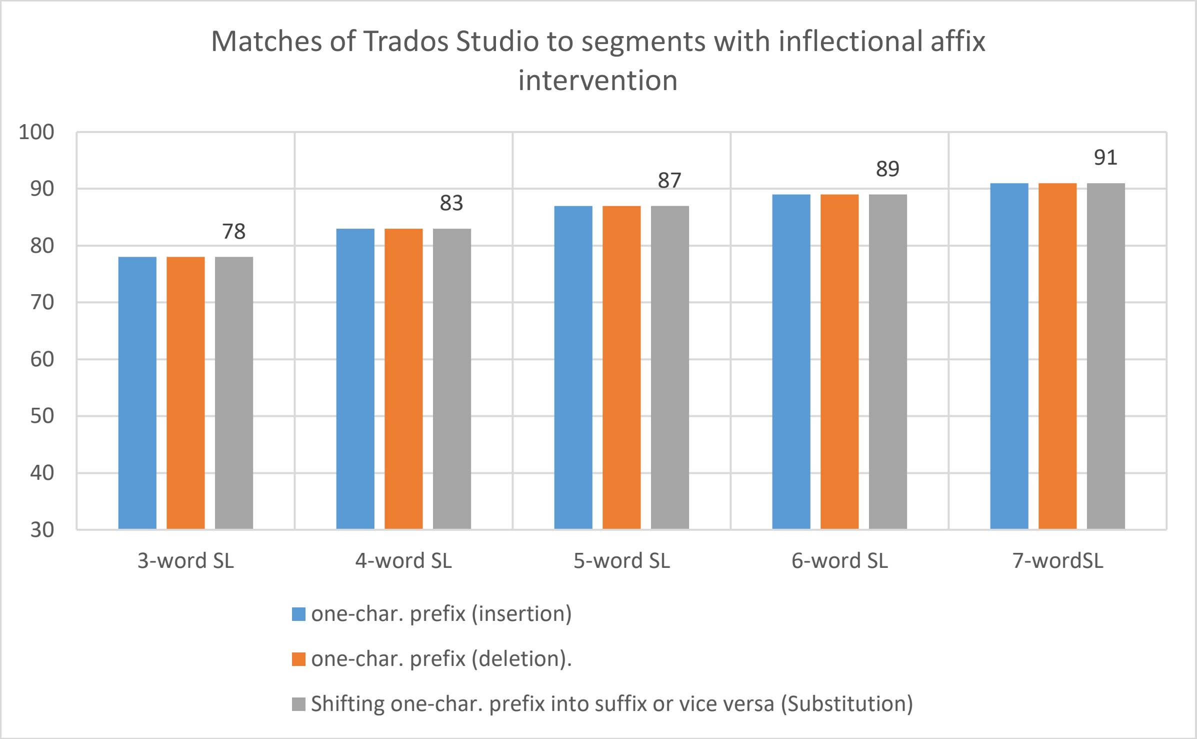


Figure 5.1.5. Trados Studio matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention

It can be seen that Trados Studio dealt with the retrieval of segments with an inflectional affix in the same way regardless of the type of character-edit operation involved. The matches were distributed between middle and high fuzzy bands, where the three- and four-word segments matched 78% and 83%, respectively (i.e. in the middle fuzzy band), and the five- six- and seven-word segments scored in a high fuzzy band. This means that TM users would not see three- and four-word segments with only a one-character difference in the high fuzzy band range.

5.1.2.6 Effect of diacritic marks on TM retrieval

The results showed that the various TM systems differed in their handling of diacritic marks. First, the algorithm of DVX, OmegaT, Trados Studio systems and the scoring of five- to seven-word segments in memoQ, which produced consistent matches according to the segments' word length, did not appear to be influenced by the insertion or removal of diacritic marks – the matches retrieved were the same.

Secondly, the metrics of Memsource and the scoring of three- or four-word segments in memoQ, whose character-based algorithm provided inconsistent values, were affected by a combination of diacritic markers. When calculating segments with and without a diacritic mark using a Levenshtein website,[56] the URL estimated a diacritic marker as a one-edit distance. Hence, a diacritic mark was treated as equal in weight to a one-character intervention in character-based metrics.

5.1.3 Discussion of results

The experiment's findings show that the TM systems treated a combination of inflectional affixes in different ways: the TM matching algorithms dealt with the morphological combination as an intervention on the whole word, as a single character change, or according to the position of the intervention. In all the systems, however, it appears that segment length had a bearing on the results.

These findings prompted a comparative analysis of each TM's retrieval of fuzzy bands. This was accomplished by using the length of each segment and the affix position and type as independent variables.

Turning to the DVX results first, it seems that the TM system's algorithm dealt with the inflectional affix as an intervention on the whole word. To account for this, a procedure calculating the surface form of the strings was used. In five-word segments, for example, DVX provided an 80% match (i.e. a 20% penalty). This may be explained by the fact that the

[56] https://planetcalc.com/1721/

algorithm estimated that a four-word string was identical to a five-word string, while a one-word string was non-similar (i.e. $\frac{4}{SL\ 5} = \frac{80\%}{100}$ identical vs. $\frac{1}{SL\ 5} = \frac{20\%}{100}$ non-similar). This implies that the DVX metrics recorded the edit operation (i.e. the inflectional affix) as an intervention on the whole word, resulting in low scores for segments that have a small number of words and an increase in scoring for longer segments.

The reason behind the OmegaT and Trados Studio results could be that their TM similarity algorithms are not only based on the number of words but also employ a specific mechanism for an individual edit operation (i.e. a single-character intervention) to measure the segments' similarity. In five-word segments, for example, any type of character editing (i.e. insertion, deletion, or substitution) was penalised 10% and 13% in OmegaT and Trados Studio, respectively; however, the matching scores provided were consistently in line with the segment's word length whatever the number of characters, which resulted in decreasing scores for short segments and increasing scores for longer ones. However, a comparison of the matching mechanisms of the two systems shows that OmegaT outperformed Trados Studio; the lowest match was scored 83% by OmegaT and 78% by Trados Studio, whereas the highest scores were 92% and 91% for OmegaT and Trados Studio, respectively.

As for the scores of memoQ, in terms of consistent scores, the system algorithm seems to use an internal mechanism to compute a combination of inflectional affixes in segments of five words or above. The mechanism produced the lowest average scores for the five-, six- and seven-word routines compared with the other systems that provided consistent scores. With a five-word routine, for example, memoQ supplied a 77% match (a 23% penalty) whatever the type of character editing. The penalties imposed by DVX, Trados Studio and OmegaT were 20%, 13% and 10%, respectively. The penalty imposed by memoQ was the heaviest. This means that the similarity algorithms in memoQ, where the measurement was word-based, imposed the heaviest penalty due to the character combination. In terms of the inconsistent matches (i.e. the three- and four-word segments), the matches were retrieved with high percentages despite the short segment length. This may be explained by the fact that the recall was based on the number of characters.

Memsource's matches, which were apparently inconsistently produced according to the number of characters, showed that the retrieval of segments with the insertion or removal of a one-character prefix gave high percentage scores, while the operation of substituting one character produced a lower percentage. It seems that Memsource's retrieval mechanism penalised a prefix combination relatively lightly. This was calculated not according to a linguistic analysis but from the perspective that a prefix combination may cause less damage to the word form than a suffix combination. As a result, in some cases, the TM matching measurement performed well when a one-character prefix (i.e. inflectional affix) was inserted or removed, but not a one-character suffix. To bear in mind, the study used a very short root – a three-character word including a single character combination, the retrieval of a longer base-form including a prefix or suffix combination may be scored differently by TM systems' algorithms.

Overall, the different tools appear to have different routines for handling such inflectional affix interventions. Although none of them is fully satisfactory, especially for short segments, Memsource outperformed the other systems when the intervention of an inflectional affix was a prefix only. The metrics of memoQ penalised the heaviest when the system provided consistent matches. In all the TM systems, the matching scores reduced as the length of the segments decreased but it was seen most clearly in the systems that produced consistent matches.

Arguably, if the TM systems had undertaken a morphological analysis, they would have treated the inflectional affix in quite a different way since they would have recognised the root form.

To summarise, the TM matching measurements failed to recognise inflectional affixes. This outcome is in line with the results of the studies conducted by Macklovitch and Russell (2000) and Planas and Furuse (1999), which found that one of the limitations of TM systems is their inability to recognise inflectional variants when retrieving stored data. The current study has provided further experimental evidence, gathered from the scores supplied by five CAT applications, showing that TM matching metrics are not good at distinguishing morphological combinations.

5.1.4 Evaluation of results

From a usability point of view, the test results show that, although the translator would potentially spend less time and effort editing the inflectional verb-variation segments, they could miss out on seeing those TM proposals because of their low scores. What the users of TM would expect – from a translator's perspective – is that TM algorithms would retrieve inflectional verb-variation segments with a very high match score (i.e. a range of high fuzzy or 85%-95%) since these would need only one edit operation to be identical to the input text. The impact of high fuzzy matches appears in the translation cost. Contrary to this expectation, however, it appears that a translator working with short segments will not be shown a high but a low fuzzy proposal, which may result in the proposals being lost. Hence, the project manager, when preparing a report, may produce inappropriate fuzziness percentages for the translation of a text with a rich morphology including segments with inflectional verb variations, and the price they quote for the translation will consequently be higher than it should be. Table 5.2 shows the bands of fuzzy matches produced for the test segments reported by each TM system.

Fuzzy bands	**% range**	**DVX**	**memoQ**	**Memsource**	**OmegaT**	**Trados Studio**
Nearly exact match	95% - 99	0	0	20	0	0
High fuzzy band	85% - 94	12	24	26	48	36
Middle fuzzy band	75% - 84	36	36	10	12	24
Low fuzzy band	50% - 74	12	0	4	0	0
No match	0 - 49%	0	0	0	0	0
Total	Total	60	60	60	60	60

Table 5.2: Fuzzy match bands as computed by each TM system

Table 5.2 displays the ways in which the TM systems differed in fuzzy-match distribution. OmegaT showed a significantly higher number of matches for the high fuzzy band (85-99%), followed by Memsource, while DVX ended up with a significantly smaller number than the other bands. The fuzzy matches varied in distribution according to the different TM systems:

- OmegaT retrieved only 12 out 60 segments, representing 20%, in a lower fuzzy band. These results appear to be the best.

- Memsource retrieved 14 out of 60 segments, representing 24%, in a lower fuzzy band; however, the high fuzzy scores were mainly produced when the intervention was a prefix.
- Trados Studio retrieved 24 out of 60 segments, representing 40%, in a lower fuzzy band.
- memoQ retrieved 36 out of 60 segments, representing 60%, in a lower fuzzy band.
- DVX retrieved 48 out of 60 segments, representing 80%, in a lower fuzzy band. These results are the worst.

As mentioned above, because the fuzzy match levels play a significant role in the calculation of translation costs, these results would have a definite impact on the discount applied to texts that are rich in morphological combinations. Preventing segments that include an inflection affix from ranking as a high fuzzy match would therefore impact the efficiency, consistency and cost of a translation.

5.1.5 Conclusion

The overall conclusion drawn from the results of testing the retrieval of TM sources for a text that is rich in morphological combinations is that all the selected systems revealed a deficiency when it came to identifying inflectional affixes, although OmegaT and Memsource returned more than three-quarters of segments in the high fuzzy band, and memoQ produced considerably better scores to short segments than longer segments. The overall matching scores appeared to be based purely on the string of surface forms and the internal machinery of each system's algorithm, without any linguistic analysis. If the TM systems had been able to undertake a morphological analysis, their TM retrieval could have been improved by prior morphological reductions such as the recognition of root forms. The outcome shows that an inflectional affix intervention was treated as either an intervention on a whole word or a single character change. Consequently, the high matching of retrieved inflectional verb-variation segments in an Arabic-to-English translation would depend on the segment length and the position of the intervention. Further work is needed to extend the investigation to other morphologically rich languages, different positional affixes and longer string formations such as a noun derivation.

The next section explores whether the five TM systems dealt with a character-marker omission as a character intervention or as a minor difference.

SECTION B

5.2 A comparative evaluation of translation memory in segment retrieval with a character-marker omission: Arabic to English

The comparative evaluation in section B was intended to complement the study in section A. The aim was to evaluate the TM retrieval of segments with morphological features in Arabic-to-English translations by investigating their retrieval of segments including a character-marker omission. The section describes the setup of the experiment (5.2.1); summarises (5.2.2), discusses (5.2.3), and then evaluates the results (5.2.4). Finally, it presents the conclusions drawn from its findings (5.2.5).

The study set out to observe the different ways in which the CAT tools' matching metrics dealt with the omission of the Hamza marker, analysing the performance of the TM retrieval in situations where two segments are exactly the same in terms of syntax and morphology, but one of them omits a character marker. The implication is that, although this omission does not impact the meaning, if it is dealt with as an error, the translator might not be shown a nearly exact match. This potential problem provided the motivation for the following investigation into the impact of the omission of a character marker.

5.2.1 Experimental setup

To achieve the aims of this study, 45 Arabic-to-English aligned segments were extracted from R.TM (see 3.2.5.2) ranging between three and seven words in length, in which the Arabic source contained a variant of Hamza. The sentences were characterised by a Hamza spelling in a word-initial, mid-word, or word-final position in segments comprising between three and seven words. At least three samples were used for each routine. The main steps for creating the test segments are detailed below.

Having extracted the test segments from the TM, the Hamza was removed from its carrier. These interventions only changed the shapes of characters; the intended meaning of the words remained the same. For example, the Hamza was removed from its carrier ئ ,ؤ , إ , أ so that the new character shapes resembled a ى, و, ا . This intervention occurred in the three different word positions: word-initial, mid-word and word-final.

The performance of the TM retrieval was accordingly defined in terms of the absence of the Hamza marker from its carrier. In this case, the effect of these errors can be viewed in the TM's fuzzy matches. Table 5.3 (below) gives an example of a segment before and after the intervention.

Intervention	**Pre-intervention**	**Post-intervention**
Omitting the Hamza marker	**وأنجبت** القدس العديد من الكتاب والشعراء / *wainjabat alquds aledyd min alkitab walshueara* / Several writers and poets were born in Jerusalem!	**و~~أ~~انجبت** القدس العديد من الكتاب والشعراء

Table 5.3: An example of a Hamza-marker omission (Thawabteh 2013)[57]

In the table (above), the example in the pre-intervention column represents the TM sources, while the example in the post-intervention column represents the test segments. The Hamza marker was then omitted from each segment.

Once the test content was processed, it was imported into the five CAT applications as a file comprising 45 Arabic segments (see Appendix Six for the test segments D). The translation project in each application was based on the corpus created as a TM file that included the original segments, while the file containing the test segments was uploaded as a file for translation. Then, a pre-translation function was used to obtain the TM matching scores. As the test segments and the TM sources were identical, apart from the difference of a Hamza marker, the most desirable outcome would be that the TM matching metrics produced scores in the range of nearly exact matches.

[57] Track Changes was used for the intervention

5.2.2 Findings

The retrieved matches show that the TM matching metrics treated an omission of the Hamza marker in different ways. The most important results provided by the five TM systems are described below. In this experiment, the range of nearly exact matches (95%-99%) was assumed to be the most appropriate as the two segments were identical in meaning and surface forms except for an omission of the Hamza marker. Column charts are used to represent the retrieved scores, in which the omission of the Hamza marker is represented on the 'x' axis and the fuzzy matches on the 'y' axis. Further, the TM systems are presented in alphabetical order.

5.2.2.1 DVX scoring

The DVX findings show that the TM scores were consistent, according to the length of the segments, whether these segments contained an omission of Hamza from its carrier in a word-initial (I), mid-word (M) or word-final (F) position. Further, the matches retrieved decreased steadily in a consistent way as the segment length decreased. The match scores ranged from 67% to 86%. Figure 5.2.1 (below) presents the scores that each segment length (SL) was assigned due to the omission of Hamza markers.

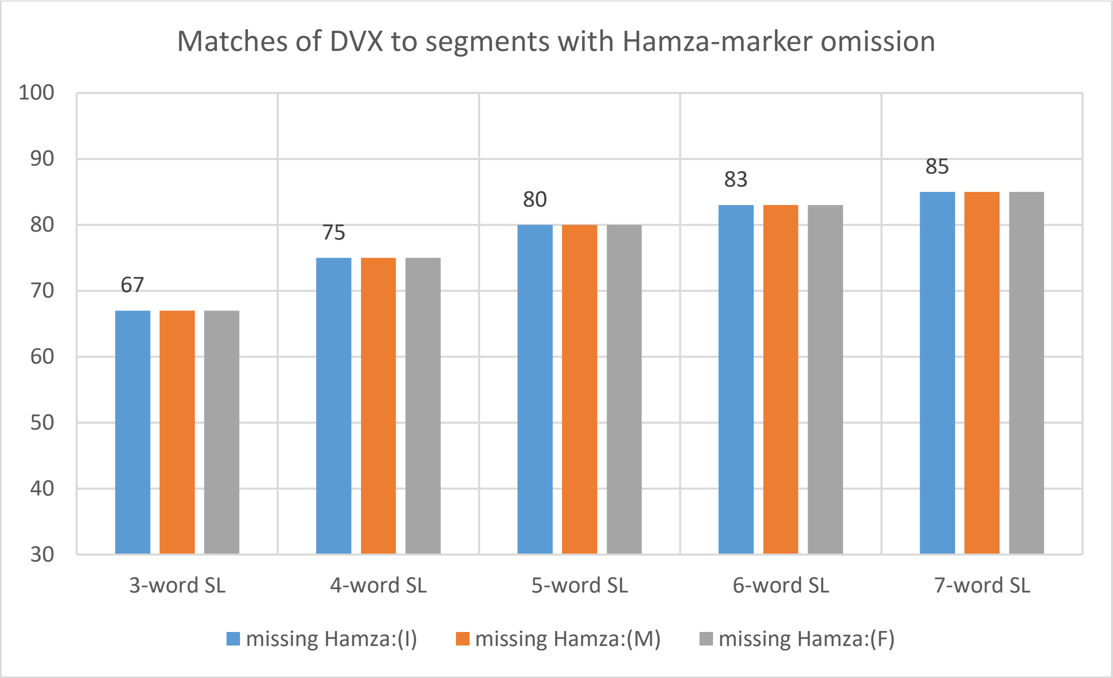


Figure 5.2.1: DVX matching scores to 3-to-7-word SL with a Hamza-marker omission in different positions

Figure 5.2.1 shows that the retrieved matches provided by the Hamza-marker omission achieved lower fuzzy levels; this was increasingly the case with the shorter segments. For example, the DVX matching metrics retrieved the three-word segments without the marker with a 67% match (i.e. a 33% penalty per character-marker omission). This suggests that the matching metrics treated the orthographic error in Arabic as a major one. As a result, the TM proposals that include an omission of the Hamza marker would not be scored as a nearly exact match (95%-99%) but would be penalised extremely heavily, especially the shorter segments.

5.2.2.2 memoQ 9.0 scoring

The scores of memoQ were categorised in two phases. Phase one comprised consistent matching values of segments of five words or above, which (according to the number of words) were relatively low. Phase two comprised the scores of three- and four-word segments, which were inconsistent and produced higher matches. The match scores ranged from 77% to 92%. Figure 5.2.2 (below) demonstrates the values of consistent matches and the values of inconsistent matches assigned to each word segment length (SL).

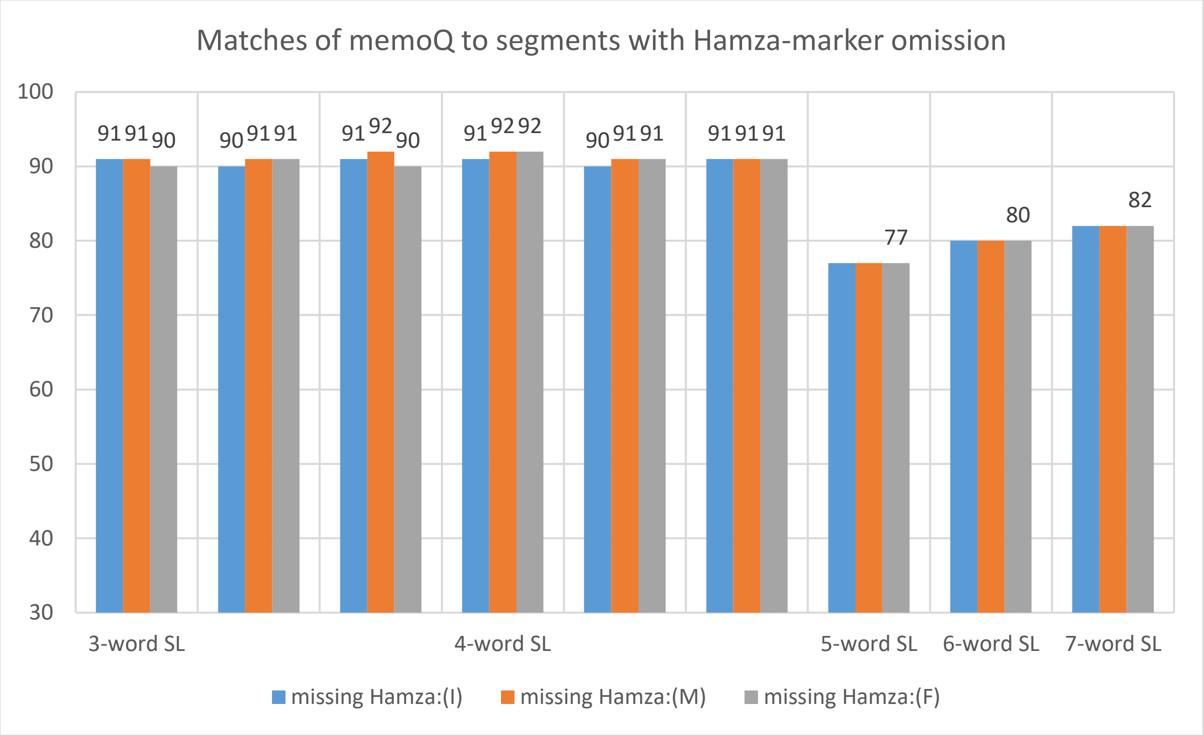


Figure 5.2.2: memoQ matching scores to 3-to-7-word SL with a Hamza-marker omission in different positions

In the figure above, the segments of five, six, or seven words with a character-marker omission were retrieved with relatively low scores. For example, the retrieval of five-word segments matched 77% (i.e. a 23% penalty due to the character-marker omission). This seems to indicate that the TM matching metrics treated the Hamza-marker omission as a major difference. In contrast, the retrieval of three- or four-word segments with the marker omission, which were given inconsistent scores, were retrieved in a higher fuzzy band. For example, segments of 20, 22 or 28 characters including an omission of Hamza marker provided matches of 91%, 91% and 92%, respectively, even though these segments shared the same number of words. This performance was the same for all positions of character markers.

Hence, it appears that memoQ produced consistent matches but with relatively lower percentages than those segment routines with inconsistent matches. This may be explained by the aforementioned hypothesis that when the memoQ system retrieved segments according to the number of characters (i.e. three- and four-word segments) the matches were higher, whereas when it retrieved segments according to the number of words (i.e. segments of five words and above) the matches were lower. Overall, although the short segments were offered higher matches than the longer ones, neither matches would be scored as nearly exact matches.

5.2.2.3 Memsource scoring

The TM algorithm of Memsource retrieved the test segments with inconsistent scores. The scores appeared to depend on the position of the Hamza marker and the character string of the word, where the two components were based on the total number of segment characters. It seems that the retrieval of segments which omitted the Hamza marker in word-initial positions were assigned higher matches than those with the omission in different word positions. Figure 5.2.3 (below) presents the inconsistent scores. The test segments were made up of between three and seven words and were 46 to 10 characters in length, while the scores ranged between 67% and 95%.

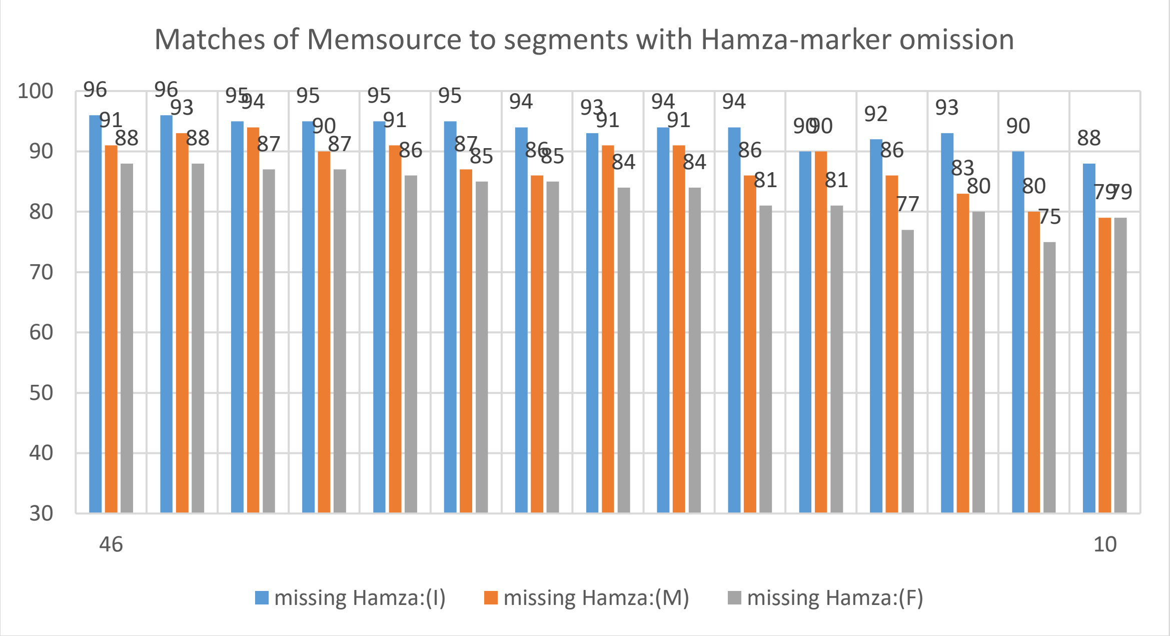


Figure 5.2.3: The matches of Memsource to segments with a Hamza-marker omission in different positions

In Figure 5.2.3, the matches were inconsistent; however, it was noted that as the segment length decreased in terms of characters, the scores decreased. The matches, therefore, seem to depend on the number of characters in the segments. The first interesting observation is that the match scores of segments where the omitted character was the Hamza marker at word-initial position were very high. With seven-word routines, for example, a segment containing 45 characters produced a 96% match for an omitted initial Hamza marker (i.e. 4% penalty), a different segment containing 42 characters produced a 91% match for the omitted marker in mid-word position (i.e. a 9% penalty), while a segment of 46 characters produced an 88% match for the same omission in a word-final position (i.e. a 12% penalty). This suggests that the Memsource matching metrics treated the position of the omitted marker in different ways: if the omitted Hamza was in a word-initial position it imposed a very light penalty, but the penalty changed if the position of the omitted character changed.

This suggests that the Memsource matching measurement considered the omission of a character marker in a word-initial position to be less important than its omission in different word positions. As a result, although the segment would be assigned to a high fuzzy band, especially when the omitted character marker was located in a word-initial position, it was not enough to score as a nearly exact match. Further, when calculating the word with and without a Hamza marker, for example the word وأنجبت / *wainjabat /* in Table 5.3, using a Levenshtein website,[58] the URL estimated the omission of a Hamza marker as a one-edit distance. Hence, a Hamza-marker omission was treated as equal in weight to a one-character intervention in character-based metrics.

5.2.2.4 OmegaT scoring

The OmegaT findings show that the match scores decreased steadily as the number of words decreased, whether these segments contained an omission of Hamza from its carrier in a word-initial, mid-word or word-final position. Further, the matches retrieved decreased steadily in a consistent way as the segment length decreased. The matching scores ranged from 83% to 92%. Figure 5.2.4 (below) displays the gradual decrease of matches assigned to each segment length (SL).

[58] https://planetcalc.com/1721/

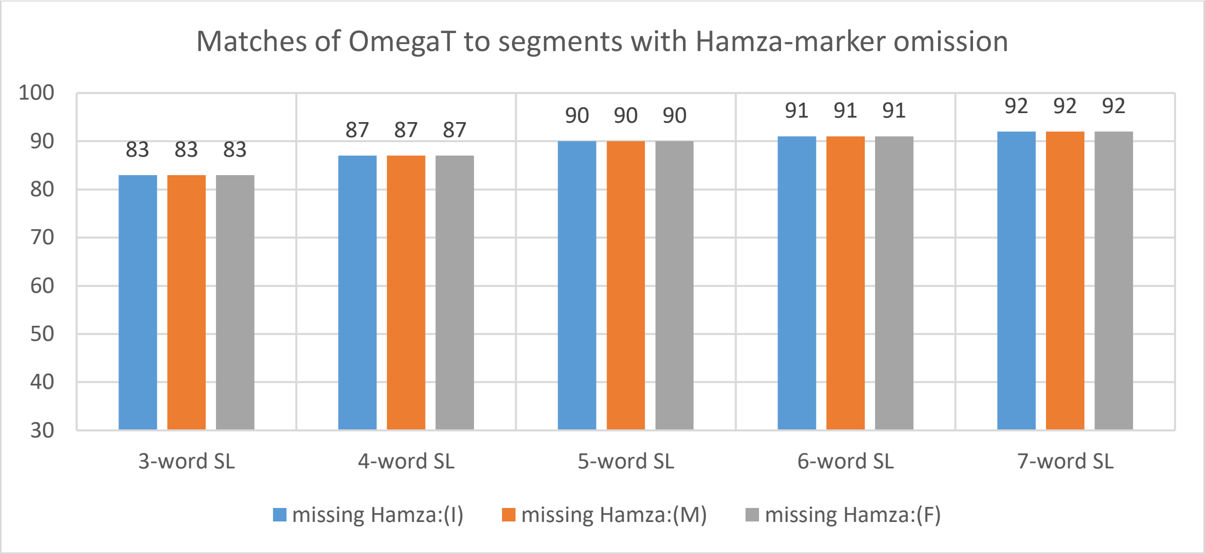


Figure 5.2.4: OmegaT matching scores to 3-to-7-word SL with a Hamza-marker omission in different positions

Figure 5.2.4 shows that the retrieved matches produced for the retrieval of segments with a Hamza-marker omission scored lower fuzzy levels. For example, the retrieval of seven-word segments matched 92% (i.e. an 8% penalty). As a consequence, the segments would not score at nearly exact match level.

5.2.2.5 Trados Studio 2019 scoring

In contrast, the results of Trados Studio reveal that the matches retrieved were very high; it seems that the omission of a Hamza marker was treated as a very minor error regardless of its word position – the scores ranged from 97% to 99%. Figure 5.2.5 (below) presents the retrieved matches that each segment length (SL) was assigned due to the missing Hamza marker.

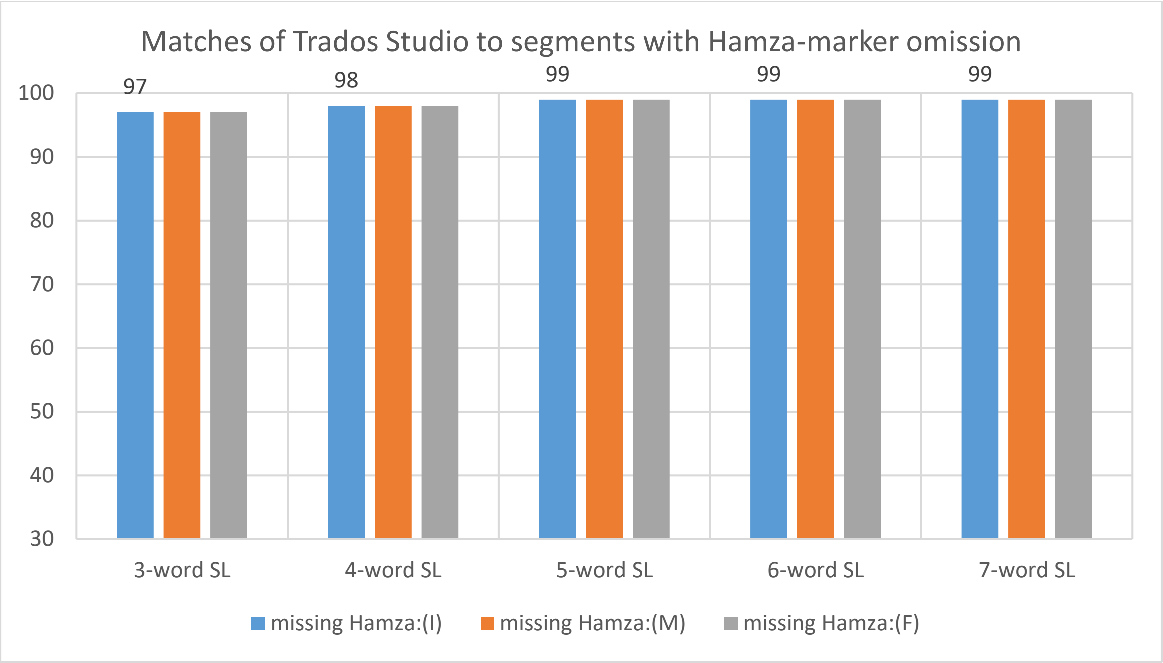


Figure 5.2.5: Trados Studio matching scores for 3-to-7-word SL with a Hamza-marker omission in different positions

The figure above clearly shows that Trados Studio provided nearly exact matches for segments that included the omission of a Hamza marker, even the shortest segments, suggesting that the system succeeded in some way to treat the omission of a Hamza as a very minor difference. In some cases, the Trados Studio measurement mechanism succeeded in detecting the orthographic error of omitting the Hamza marker, and therefore it performed the best out of all the systems in the retrieval of segments with a Hamza-marker omission when translating from Arabic to English.

5.2.3 Discussion of results

The test results reveal that the TM systems treated the retrieval of segments containing a Hamza-marker omission in differing ways. The resultant matches reflected how each system handled such an orthographic error – either as a word intervention or a character intervention; the outlier was Trados Studio.

The matching metrics of DVX appear to treat the Hamza-marker omission as a word intervention. For instance, four-word segments were assigned a 75% match (i.e. a 25% penalty). This may be explained by the TM's estimation that a three-word string was identical to a four-word string, while a one-word string was non-similar (i.e. $\frac{3}{SL\ 4} = \frac{75\%}{100}\ identical\ vs.$

$\frac{1}{SL\ 4} = \frac{25\%}{100}$ non-similar). This means the matching metrics recorded the character-marker omission as an intervention on the whole word, which resulted in lower fuzzy matches.

OmegaT and memoQ both appear to treat a Hamza-marker omission as a character intervention, but they approach it in different ways. Firstly, the matches of OmegaT and some of memoQ's scores – five- to seven-word segments – were assigned consistently according to the number of words in the segment. In five-word segments, for example, 90% and 77% were assigned by OmegaT and memoQ, respectively. However, in a comparison of the matches between the two systems, OmegaT outperformed memoQ. Secondly, Memsource's and some of memoQ's scores – three- and four-word segments – were assigned higher matches, which placed them at nearly exact match level. Memsource's measurement imposed a lighter penalty when the missing marker was located in a word-initial position. High scoring may be explained by the fact that recall was based on the total number of characters.

Finally, Trados Studio's treatment of the omission of the hamza marker was significantly more effective: the matches were ranked at the range of nearly exact matches regardless of segment length. This suggests that the system detects the orthographic errors and handles them as minor differences. As a result, TM proposals which include a character-marker omission would be offered in the nearly exact match band, with a consequent impact on the cost of the translation.

5.2.4 Evaluation of results

One of the main advantages of working with a TM system is that the higher the matches retrieved, the greater the reduction in translation price. This means that the matches retrieved in the band of early exact matches (95%- 99%) attract a very low charge. Table 5.4 (below) presents the levels of fuzzy matches (per segment) the five TM systems produced for the test segments.

	DVX	**memoQ**	**Memsource**	**OmegaT**	**Trados Studio**
95%-99%	0	0	6	0	45
85%-94%	9	18	30	39	0
75%-84%	27	27	30	6	0
50%-74%	9	0	9	0	0
No-matches	0	0	0	0	0
Total	45	45	45	45	45

Table 5.4: The distribution of test segments according to the level of fuzzy matches using the five TM systems

As the table above shows, Trados Studio ranked 45 out of 45 segments, representing 100%, in the highest band of fuzzy matches (the nearly exact match band). This was followed by Memsource, which matched six segments, representing 8%, at nearly exact match level. The retrieved matches of the other systems scored lower. Hence, Trados Studio would undoubtedly offer the best price for clients, whereas the other systems would mean higher prices; this is solely due to the difference in calculations regarding the Hamza-marker omission.

5.2.5 Conclusion

As can be seen above, it seems clear that the matching metrics of the various TM systems treated the omission of the Hamza marker in Arabic texts in different ways. Trados Studio outperformed the other systems in that it seemed to recognise the variant Hamza and thus treated its omission as a minor difference, providing matches in the nearly exact matchband. Meanwhile, those matching scores of Memsource that were assigned the nearly exact match level appeared to depend on the word-initial position and the total number of segment characters.

5.3 Summary: TM retrieval of Arabic linguistic features

The following section presents an analysis of the results obtained from the four experiments described in Chapters Four and Five. These compared the performance of five TM systems when retrieving segments containing an Arabic linguistic feature in order to produce an English translation. The findings showed that, whereas each TM system seemed to use its own method to compute the matching scores, all of them appeared to base these scores on the string of surface forms. This summary of the results may help to provide an answer to the main question raised by this research (see section 1.4): which translation tool can best handle Arabic linguistic features when translating between Arabic and English?

Section A in Chapter Four investigated the ways in which the five TM systems retrieved segments that included a reordering operation (i.e. a syntactic structure), while section B analysed the ways in which they retrieved segments that included a three-operation edit. Section A in Chapter Five evaluated the performance of the systems when retrieving segments containing an inflectional affix (i.e. a morphology inflection), while section B examined the retrieval of a segment with a Hamza-marker omission (i.e. an orthography feature).

Overall, the findings show that the TM systems' output could be classified as two different groups. The first includes systems which provided a consistent banding of their fuzzy matches according to the number of words in the test segment. This group consists of DVX, OmegaT, Trados Studio and some of memoQ's scores (i.e. segments of five words or above). The second group, which consists of Memsource and some of memoQ's scores (i.e. three-word and four-word segments), supplied inconsistent scores according to the total number of characters in the test segment. In terms of the reordering operation and the morphological inflection, the matching scores in the two groups appeared to be based purely on the string of surface forms; the matches fell steadily as the segment length became shorter. In terms of orthography, excepting Trados Studio, the other four systems employed a similar method. Trados Studio, however, dealt with the missing Hamza markers in a different way: it seemed to recognise the variant Hamza and treated the omission accordingly as a minor difference.

Chapter Four concluded that the TM matching metrics scored the segment retrieval that included a reordering operation in lower fuzzy bands than those provided by the three-operation edit. Chapter Five concluded that the matching metrics dealt with the segment retrieval containing a one-character inflection and the omission of the Hamza marker as either an intervention on a whole word or a single character change.

The results of the four experiments were subjected to analysis in order to reach a conclusion regarding TM retrieval. Segments of five, six and seven words were selected – the length of segments used as the content of the four experiments. The comparison consisted of the segment retrieval with a one-word move, one-word addition, one-word deletion and one-word substitution from Chapter Four; and the segment retrieval with a one-character edit and a case-marker omission from Chapter Five. The comparison is presented below, system by system.

Table 5.5 (below) presents the DVX matches that the different segment lengths achieved due to each linguistic feature.

Intervention	**5-word SL**	**6-word SL**	**7-word SL**
One-word move	60%	66%	72%
One-word addition	83%	85%	87%
One-word deletion	80%	83%	85%
One-word substitution	80%	83%	85%
One-character inflection	80%	83%	85%
Character-marker omission	80%	83%	85%

Table 5.5: A comparison of DVX matches due to the different linguistic features

Table 5.5 clearly shows that, out of all the features (a three-operation edit, a character inflection, a character-marker omission and a move operation), the move operation scored the lowest. Further, the character-marker omission was equated to a one-character inflection and to a whole-word substitution or deletion. This suggests that the matching metrics of DVX did not treat the segment retrieval containing Arabic linguistic features linguistically but as edit operations.

Table 5.6 (below) displays the memoQ matches that the different segment lengths achieved due to each linguistic feature.

Intervention	5-word SL	6-word SL	7-word SL
One-word move	65%	70%	73%
One-word addition	85%	87%	88%
One-word deletion	83%	85%	87%
One-word substitution	77%	80%	82%
One-character inflection	77%	80%	82%
Character-marker omission	77%	80%	82%

Table 5.6: A comparison of memoQ matches due to the different linguistic features

The table above shows that the segment retrieval that included a move operation scored the lowest match. Further, the same match was given to segments with the omission of a character-marker, one-character inflection and one-word substitution. This suggests that memoQ's matching algorithm dealt with segments that included Arabic linguistic features as edit operations; no linguistic information seems to have been involved.

Table 5.7 (below) illustrates OmegaT's matching scores for the different segment lengths due to each linguistic feature.

Intervention	5-word SL	6-word SL	7-word SL
One-word move	80%	83%	85%
One-word addition	83%	85%	87%
One-word deletion	80%	83%	85%
One-word substitution	90%	91%	92%
One-character inflection	90%	91%	92%
Character-marker omission	90%	91%	92%

Table 5.7: A comparison of OmegaT matches due to the different linguistic features

As Table 5.7 shows, the segment retrieval that included a move operation was dealt with in a similar way to a deletion operation; however, these retrieved scores matched the lowest.

Further, the character-marker omission was equated to a one-character inflection, and also to a whole-word substitution. This suggests that the OmegaT system's matching metrics treated the segment retrieval including Arabic linguistic features as edit operations, not linguistically.

Table 5.8 (below) shows the Trados Studio matches that the different segment lengths achieved due to each linguistic feature.

Intervention	**5-word SL**	**6-word SL**	**7-word SL**
One-word move	73%	78%	81%
One-word addition	87%	89%	91%
One-word deletion	84%	87%	89%
One-word substitution	87%	89%	91%
One-character inflection	87%	89%	91%
Character-marker omission	99%	99%	99%

Table 5.8: A comparison of Trados Studio matches due to the different linguistic features

Table 5.8 shows that, in Trados Studio, the match of a move operation was also scored the lowest. Further, the match produced for a one-character inflection was similar to that for a one-word substitution, suggesting that a one-character inflection was equated to a whole-word substitution. Interestingly, Trados Studio produced scores at the level of a nearly exact match for the omission of the Hamza marker, suggesting that although it treated the addition operation and one-character inflection as a whole-word substitution, it seemed to recognise the orthographic error and imposed a very low penalty for the Hamza-marker omission. Hence, it could be argued that Trados Studio's treatment of orthographic errors may be built on an Arabic-specific filter routine since the lightness of the penalty indicates that the system treated the missing characters as minor differences.

However, with Memsource, the matches were inconsistent regardless of the type of intervention. It seems that retrieval was based on the number of characters. In terms of a reordering operation, the system provided the lowest match in comparison to the three-operation edit (see Chapter Four, Table 4.3). Regarding morphological inflection, Memsource provided very high matches for the retrieval of segments that included a one-character prefix edit (see Figure 5.1.3); segments with the omission of the Hamza marker (but only in the word-initial position) also provided high matches (see Figure 5.2.3). This suggests that the Memsource system assigned high scoring according to the position of the intervention, not according to a linguistic perspective.

Given the results summarised above, it can be said that the matching metrics of the five TM systems – apart (possibly) from Trados Studio in relation to the character-marker omission – do not appear to have any linguistic basis. Trados Studio treated the omission of the marker as a minor difference. In contrast, the matching metrics of the DVX, memoQ, OmegaT and also Trados Studio systems, in which the recall was based on the number of words in a segment, show that a one-character inflection was equated with a whole-word substitution. This was the same for the segment with a missing marker in DVX, memoQ and OmegaT, but not Trados Studio. This may be due to the fact that, in a European language, changing a single character often completely changes the meaning of the word.

In terms of usability, these results imply that semantically highly repetitive texts are not enough on their own if the TM is to be put to good use; the segments also have to be in the same word order. The current TM matching metrics tend to be based on surface string and edit distance only, and failure to develop the matching measurement could result in a significantly lower usability of TM matches and higher costs.

To conclude, although not one of the TM systems tested performed to a satisfactory standard with the three Arabic linguistic features under investigation, OmegaT outperformed the other systems in terms of a reordering operation because at least the system measured the reordering operation as similar to a deletion operation and therefore did not assign heavier penalties (as the other systems did). In terms of the morphological inflection, OmegaT also outperformed the other systems – the system retrieved about 80% of the segments that included a one-character inflection at a high fuzzy level (85%-94%), while the other systems matched the segments at lower fuzzy levels. In terms of the orthographic errors, Trados Studio outperformed the other systems as it treated the omission of a Hamza marker as a very minor difference, and as a result, provided scores at the level of nearly exact matches (95%-99%).

CHAPTER SIX

LINGUISTIC ERROR CLASSIFICATION OF MACHINE TRANSLATION OUTPUT

6.0 Introduction

Chapter Six is a transitional chapter between the part of this thesis concerned with the evaluation of TM function and the part investigating the quality of MT output. The chapter evaluates Google NMT, as a representative of NMT systems, using a subset of the test data as used in translation memory retrieval Section 6.1 reviews the findings of the TM retrieval tests in Chapters Four and Five. Following this, section 6.2 describes this experiment's evaluation method and setup, section 6.3 presents the results and section 6.4 summarises the conclusions. Finally, section 6.5 explains how this experimental work could be extended.

6.1 Review of the findings of TM retrieval

Chapters Four and Five investigated the retrieval capabilities of TM systems when dealing with Arabic language segments that include a reordering operation, a morphological inflection and orthographic errors. The results showed that although the systems' similarity measurement retrieved long segments containing the reordering operations with usable scores (70% or above), they failed to retrieve shorter sentences at the same level. Based on these results, it was concluded that the potential for the re-use of high-similarity translations in TM is not high due to weaknesses in the TM matching measurement. In the case of morphological inflection, the matching metrics equated an inflectional affix with a whole-word substitution, which resulted in the provision of lower fuzzy matches for segments that differed only in the inflectional affix. In the case of orthographic errors, all the systems tested, apart from Trados Studio, equated an omitted character marker with an inflectional affix, which in turn was equated with a whole-word substitution. This prevented segments containing orthographic errors from ranking as a nearly exact match, potentially impacting the efficiency, consistency and cost of the translation.

The decision to use the same test data to evaluate the performance of MT in dealing with these three main features was taken with these results in mind. This sort of evaluation should enable a comparison between the two types of translation tools, providing translators with valuable information about which tool would offer them the more relevant help.

Although in the previous tests, the language direction was Arabic to English, this experiment investigates both directions to enable a further comparison. The aim was to highlight the errors and the lack of errors in the MT output when producing Arabic<>English translation (i.e. Arabic to English and English to Arabic).

6.2 Experiment and evaluation

6.2.1 Evaluation method

The method of evaluating the MT quality was based on a subset of the MQM error typology (Lommel et al., 2014),[59] which has four main error classifications: accuracy, language, terminology and style. The errors are also assigned four severity levels: critical, major, minor and neutral. Penalties are applied to each error and level of severity.

Two of the above categories were selected for this experiment: accuracy and language. The experiment applied four subcategories of the accuracy category: mistranslation (incorrect interpretation of source text), addition (unnecessary elements in the translation not originally present in the source text), omission (essential element in the source text missing in the translation) and non-translation. The three subcategories of the language category were: grammatical errors, syntactic errors, spelling and punctuation errors. The terminology category was disregarded since the test data, which was originally extracted from the MeedanMemory corpus, consisted of a general type of data. Furthermore, this category should only be applied if a terminology source or glossary is provided, which is not used in our research. Style was also disregarded since the test segments belonged to a range of styles; however, the test segments were short, simple and comprehensible even without any context.

Two levels of severity were selected for the experiment: major errors and minor errors. Major errors applied when a significant change was seen in the meaning of the translated sentences (i.e. the errors were visible in the MT content). Minor errors applied when there was no loss of meaning but a decrease in quality, fluency or clarity. The critical error was disregarded since this level of errors applied when there was major loss of meaning which was not found in the simple kind of test data. Regarding the penalties, two degrees of penalty were applied: a two-

[59] http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf

point penalty imposed due to a major error and a one-point penalty imposed due to a minor error Lommel et al., (2015).

6.2.2 Data and translating

As mentioned earlier, this study used the same content as the tests evaluating TM retrieval. As they contained short, simple segments, only one was chosen to represent the four tests: test suite A. This consisted of 95 (Arabic and English) translation units originally extracted from MeedanMemory (ranging from 3- to 10-word sentences). Two tests were created from its content: test one comprised the 95 segments in Arabic, while test two contained the 95 segments in English (see Appendix Seven for the test segments A).

Next, test one was translated from Arabic into English and test two from English into Arabic, using Google Translate - Google was selected as it is the best-known and most widely used MT system.[60]

6.2.3 Evaluation procedure

In order to be representative of all types of manipulations, a subset of 25 out of the 95 translations from each test was selected for error annotations. That means that error samples of all different manipulations are represented in the tests. When translating into Arabic, the MT system translated 167 words, and when translating into English, it translated 228 words.

The two translations were inserted into two spreadsheet files segment by segment to make the error annotations. The researcher flagged the errors, then, the results were reviewed by two different people:

In terms of the Arabic output, the error annotations were reviewed by an MT student whose first language is Arabic with very advanced English. Regarding the English output, the error annotations were reviewed by an English studies PhD student at Swansea University who is a native speaker of English.

[60] The tests were translated in August 2019: https://translate.google.co.uk/

The errors were counted and classified according to the subcategories mentioned above: mistranslated, non-translated, additional or missing words were classified as errors of accuracy; grammatical errors, misspellings or errors of syntactic structure were classified as errors of language.

Table 6.1 (below) show the subcategories of errors, severity levels and penalty degrees used in the experiment.

<table>
<tr><th>Error category</th><th colspan="2">Subcategories</th><th>Severity level</th><th>Penalty</th></tr>
<tr><td rowspan="4">Accuracy</td><td colspan="2">Mistranslation</td><td>Major error</td><td>2-point penalty</td></tr>
<tr><td colspan="2">Non-translation</td><td>Major error</td><td>2-point penalty</td></tr>
<tr><td colspan="2">Addition</td><td>Major error</td><td>2-point penalty</td></tr>
<tr><td colspan="2">Omission</td><td>Major error</td><td>2-point penalty</td></tr>
<tr><td rowspan="5">Language</td><td colspan="2">Grammatical errors</td><td>Major error</td><td>2-point penalty</td></tr>
<tr><td rowspan="2">Syntactic errors</td><td>subject-verb-object</td><td>Minor error</td><td>1-point penalty</td></tr>
<tr><td>modified term before modifier (e.g. adjective precedes noun)</td><td>Major error</td><td>2-point penalty</td></tr>
<tr><td colspan="2">Misspelling</td><td>Minor error</td><td>1-point penalty</td></tr>
<tr><td colspan="2">Punctuation errors</td><td>Minor error</td><td>1-point penalty</td></tr>
</table>

Table 6.1: Subcategories, severity levels and penalties degree used in the experiment (Lommel et al., 2014)

The errors identification was based on the pre-established categorization, the severity levels and the penalty systems.

6.3 Findings

The translations provided by the MT system show that the system produced a number of syntactic and morphological errors. The results of the experiment are summarised below.

6.3.1 English output of MT

The results displayed in Table 6.2 (below) show that translating from Arabic into English using MT produced perfect word order in English – no errors were made. Further, the morphology produced was very good.

Error categories	**Error sub-categories**	**NWE**	**NSE**	**Linguistics features**	**Rate of errors**	**Rate of success**
Errors of accuracy	**Mis.W.**	5	*	Morphology	5.26%	94.74%
	Non.T.W	0	*			
	Addition	0	*			
	Omission	3	*			
Errors of language	**G.E**	4	*			
	S.E	0	*			
	W.O.S	*	No errors	Word order	0	100%
	Total	**12**	**0**			

Table 6.2: The rates of error and success for MT output in English

NWE: Number of word-level errors (... out of 167)

NSE: Number of segment-level errors (... out of 25)

MIS.W: Mistranslated words

NON.T.W: Non-translated words

G.E: Grammatical errors

S.E: Spelling errors

W.O.S: Word order structure

As Table 6.2 shows, it is clear that MT produced high-quality translations from Arabic into English. The translated segments were free from errors of non-translation, addition, or spelling, as well word order; however, a few errors were seen in the other subcategories, the most frequent being grammatical. Examples of these can be seen below.

a) **Mistranslations**: MT incorrectly translated some information from Arabic into English. For example:

   Source text (ST):

   صحيفة الأخبار المصرية **تقدم** معطيات حول القمة العربية في الدوحة.

   Transliteration: *sahifat al'akhbar almisriat* ***taqadam*** *mueatiyaat hawl alqimat alearabiat fi aldawha.*

   MT output: The Egyptian newspaper Al-Akhbar **presents** data on the Arab summit in Doha.

   Back translation: صحيفة الأخبار المصرية **تعرض** بيانات عن القمة العربية في الدوحة.

In the above example, MT mistranslated the word تقدم /*taqadam/* as ‘presents’; it would be correct if it was translated as ‘provides’. In the back translation, the translation of the verb ‘presents’ expressed different meaning.

b) **Omissions**: the MT English output missed some information contained in the Arabic source. For example:

   ST:

   تحرر قوات النيتو بأفغانستان **صحفي** نيويورك تايمز البريطاني

   Transliteration: *tuharir quwwat alniytu bi'afghanistan* ***suhufia*** *niuyurk taymz albritanii*

   MT output: The NATO forces in Afghanistan liberate the British *New York Times.*

   Back translation: *:* قوات الناتو في أفغانستان تحرر صحيفة نيويورك تايمز البريطانية.

The MT's English output did not contain the meaning of the word صحفي / *suhufia*; it was supposed to be translated as 'journalist'. In the back translation, the meaning of the word 'journalist' was missing.

c) **Grammatical errors**: MT committed some grammatical errors. For example:

ST:

يستهدف تفجير سفارة فرنسا بنواكشوط

Transliteration: ***yastahdif*** *tafjir sifarat faransa bnwakshut*

MT output: A bombing **targeting** the French embassy in Nouakchott.

The MT made a grammatical error: the English target did not contain the correct form; it would have been better to use (A bombing targets).

6.3.2 Arabic output of MT

The results illustrated in Table 6.3 (below) show that MT did not produce Arabic's natural word order – VSO. However, it provided good translations in terms of morphology.

Error categories	**Error sub-categories**	**NWE**	**NSE**	**Linguistic features**	**Rate of errors**	**Rate of success**
Error of accuracy	Mistranslation	7	*	**Morphology**	**9.58%**	**90.42%**
	Non-translation	2	*			
	Addition	0	*			
	Omission	3	*			
Errors of language	Grammatical errors	6	*			
	Spelling errors	0	*			
	Word-order structure	*	18	**Word order**	**72%**	**28%**
	Total	16	18			

Table 6.3: Error and success statistics for the MT output in Arabic

NWE: Number of word-level errors (... out of 228)

NSE: Number of segment-level errors (... out of 25)

Table 6.3 shows that MT produced 18 out of 25 segments that were not in the natural word order. However, its performance was better in terms of morphology: it committed only a few errors, almost all of which were mistranslated words. Some examples are shown below.

a) **Word order**: the results showed that MT produced the Arabic word order as SVO (subject-verb-object); however, the natural word order of Arabic is VSO. The example below displays the MT system’s erroneous preferential decision:

ST: **A Muslim woman sues** a judge in a headscarf incident.

MT output:

امرأة مسلمة تقاضي قاض في حجاب.

Transliteration: ***aimra'at musalamat taqadi*** *qad fi hijabin*

Back translation: A Muslim woman sues a judge in a veil.

Although this example is grammatically correct, the word order would be better if it were VSO, with the emphasis focused on the action of judging. As a result, the translation is incoherent.

b) **Mistranslation:** MT translated some information from the English source into Arabic incorrectly. For example:

ST: The **dismissed** Palestinian government denies the relationship of resistance factions with rockets fired from Gaza.

MT output:

الحكومة الفلسطينية **المبعدة** تنكر علاقات فصائل المقاومة بالصواريخ التي تطلق من غزة.

Transliteration: *alhukumat alfilastiniat* ***almubeadat*** *tnkr ealaqat fasayil almuqawamat bialsawarikh alty tutliq min ghazat*

Back translation: The **expelled** Palestinian government denies ties to the resistance factions with rockets fired from Gaza.

In the example above, the meaning of the word ‘dismissed’ in the English source was translated as المبعدة / *almubeadat* (back translation, ‘expelled’). A better translation would have been المقالة / *almuqala*.

c) Non-translation: the Arabic sentences included some untranslated words; they were left in English. For example:

ST: **T-Mobile** in Germany bans the use of **Skype** on **iPhone**.

MT output: **T-Mobile** في ألمانيا يحظر استخدام **Skype** على **iPhone**.

Transliteration: *T-Mobile fi 'almania yahzur aistikhdam Skype ealaa iPhone*

Back translation: T-Mobile in Germany prohibits using Skype on iPhone.

The proper nouns of the English words, ‘T-Mobile’, ‘Skype’ and ‘iPhone’ were not translated into Arabic; the transliteration technique should have been used to transfer these words into Arabic.

d) **Grammatical errors**: MT committed some grammatical errors which caused incorrect information. For example:

ST: **A researcher** uses his thoughts to update Twitter.

MT output:

يستخدم **الباحث** أفكاره لتحديث تويتر

Transliteration: *yustakhdam* ***albahith*** *'afkarah litahdith tuyitir*

Back translation: The researcher uses his thoughts to update Twitter.

In the example above, MT translated the English phrase ‘a researcher’, which uses the indefinite article, into Arabic as الباحث , using the definite article ال /*al*. It should have been translated into the equivalent in Arabic, باحث / *bahith*, without any article. In the back translation, it is translated into ‘the researcher’, changing the meaning.

6.3.3 MT syntactic errors vs morphological errors

An holistic analysis of these results shows that MT produced correct translations in terms of morphology in both translation directions, as well as the English word order. However, it often failed to produce the natural word order of Arabic (VSO). Pie charts are used to represent the rate of errors.

Figure 6.1 (below) displays the overall percentage when the output was in English. The morphology percentage was calculated according to the number of translated words (228 words), while the syntactic structure score was computed according to the number of translated segments (25 segments).

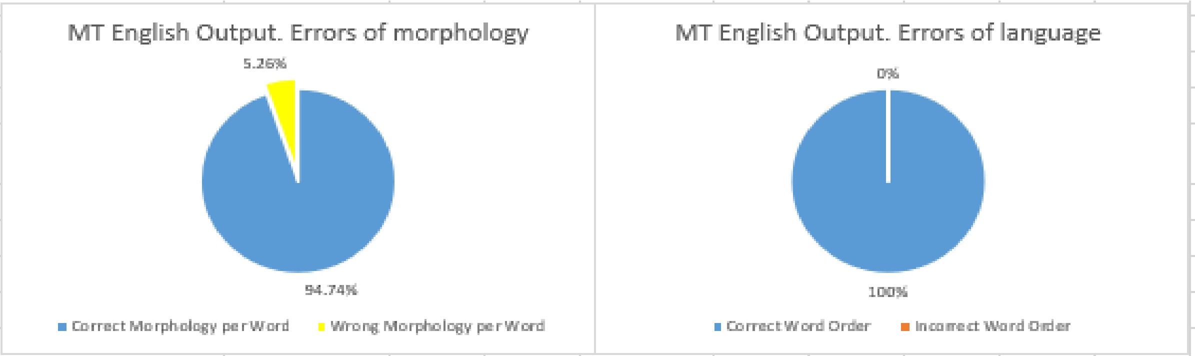


Figure 6.1: Analysis of the syntax and morphology of the MT production when the output was English

In Figure 6.1, the pie charts clearly exhibit that the quality of the MT translation when translating from Arabic to English was very high. MT produced a 100% correct translation in terms of word order, while it produced around a 95% correct translation of the English morphology.

Figure 6.2 (below) shows the overall statistics for the Arabic output. The number of translated words was 167, while the number of translated segments was 25.

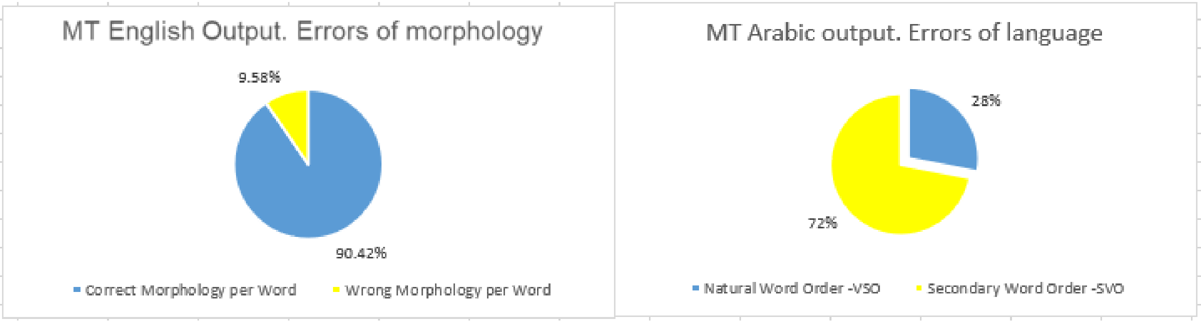


Figure 6.2: Analysis of the syntax and morphology of the MT production when the output was Arabic

In the figure above, it can be clearly seen that MT transferred less than a third of the correct word-order structure: 18 out of 25 segments were not produced in the natural word order, whereas the tool produced a better average regarding morphology, representing 90%.

The results of the MT system suggest that, in terms of morphology, good translations were produced in Arabic-to-English translation; however, MT rarely reproduced Arabic's natural word order.

6.4 Summary

Testing the same data using TM and MT tools enabled a comparison of the tools' output in terms of their usefulness to translators. The results mentioned above show that, in terms of morphology, good translations were produced in both translation directions; however, MT rarely reproduced Arabic's natural word order.

The NMT system made few morphological errors when translating a text, whether this was Arabic into English or vice versa, while the TM systems appeared deficient when it came to identifying inflectional affixes in the retrieval of Arabic texts. This suggests that MT can deal with morphology effectively, while TM cannot.

Regarding word order, MT produced a number of errors when trying to recreate the natural word order of Arabic (VSO), while the similarity measurement of TM, although it retrieved long sentences that included a reordering operation in usability matching values, failed to do the same for short sentences. This suggests that both translation tools found the different syntactic rules of the two languages problematic.

To conclude, MT generated satisfactory translations in terms of morphology; however, both MT and TM systems struggled to deal properly with the phenomenon of Arabic's flexible word order.

6.5 Extending the evaluation

This study has evaluated very specific aspects of MT. It could be further enhanced by extending the evaluation of the quality of MT output in the following ways. First, the test content here comprised relatively short, simple segments (three-to-ten words); it would be useful to test longer segments consisting of natural-length sentences. Secondly, the test segments were extracted from open-resource data (MeedanMemory), and on occasion the quality of the English translation was not good. In such cases, it would be worthwhile using data from commercial (confidential) resources. Thirdly, the tests were translated via one NMT system (Google Translate); it would be useful to use a multiple set of NMT systems, in order to compare the results. Fourthly, it would also be interesting to evaluate MT output using the

TAUS two-parameter scale of adequacy and fluency, and also applying automatic evaluation metrics.

As mentioned previously, part of the aim of this study was to evaluate the MT output using a combination of automatic metrics and human judgment, thus allowing a comparison between the results of the evaluation criteria. The study's advances in these directions are described in the following chapter.

CHAPTER SEVEN

COMPARATIVE EVALUATION OF THE OUTPUT QUALITY OF NEURAL MACHINE TRANSLATION SYSTEMS: ARABIC<>ENGLISH TRANSLATION

7.0 Introduction

Chapter Seven presents an investigation, using a combination of human evaluation (applying adequacy and fluency rankings) and the BLEU and hLEPOR automated metric, into the quality of translations produced by a set of NMT systems for a selection of Arabic<>English language pairs. The chapter is organised as follows: section 7.1 describes the findings, section 7.2 discusses the results, and section 7.3 presents the conclusions.

The investigation used the Application Programming Interface (API) of four NMT systems in Python to translate two sets of Arabic and English samples. The basic API model is similar to visiting a free service of Google Translate, Bing Microsoft, etc; however, an API is a way to programmatically interact with a separate software application component. Google APIs for example, allow the translator to connect the code to the whole range of Google services, such as Google Translate.

7.1 Findings

The results of human evaluation of the quality of Arabic<>English MT systems are provided below. The answers to the demographic questions are presented as descriptive variables. The chapter begins with some detailed demographic information about the participants in terms of their experience and background in evaluating MT output using Adequacy and Fluency, then examined the human evaluators' ratings of the translations according to an Adequacy and Fluency scale, followed by the BLEU and hLEPOR scores. Finally, it summarised the results of the two evaluation methods. Column charts are used to represent the evaluators' ratings. Further, the MT systems are presented in alphabetical order.

7.1.1 Demographic questions

This study was conducted in the Department of Modern Languages, Translation and Interpreting, Swansea University. The total of participants was 15 postgraduate / graduate students who agreed to participate in the investigation. Ten respondents with first language Arabic and an advanced knowledge of English ranked the Arabic translations in terms of

Adequacy and fluency, and the English translations in terms of Adequacy only. Five respondents who are native speakers of English ranked the English translations in terms of fluency.

Before beginning the tests, the Arabic-English translators were asked to supply the following information:

- Do you use MT systems to translate (in either direction) between Arabic and English? (Yes: 90%; No: 10%)
- Do you usually edit the MT systems' output? (Yes: 80%; No: 20%)
- Have you previously been involved in evaluating the output quality of MT in terms of Adequacy and Fluency? (Yes: 40%; No: 60%)

The results of the demographic questionnaire show that 10 participants were qualified translators according to the definition included in the questionnaire, 90% said that they use the MT systems, while 80% answered that they use post-editing to improve the quality output of the MT systems. Only 40% of the participants had previously been involved in Adequacy and Fluency rating. The demographic information related to the participants' experience and their background shows that the participant had less practice in rating Adequacy and Fluency.

7.1.2 Arabic-to-English NMT translations

Four Arabic source sentences (questions 1-4) were translated into English using four different NMT systems (Bing Microsoft Translator (hereafter referred to as Bing M), Google, Lilt and Yandex).[61] The Arabic source and the different versions of the translation are given below.

Question 1:

وأضاف اعتقد ان العلاقة وصلت الى مستوى لم نشهده من حيث الصدق والصراحة مع مسئولين كوريين

ديمقراطيين اجتمعنا معهم خلال الايام الاربعة الماضية.

[61] The tests were translated in March 2020

Transliteration:*[62] *wa'adaf 'aetaqid 'ana alealaqat wasalat 'iilaa mustawaaan lm nashhaduh min hayth alsidq walsarahat mae masyuwlin kuriiyn dimuqratiiyn aijtamaena maeahum khilal al'ayam alarbet almadia*

Bing M. output: 'I think the relationship has reached a level that we have not seen in terms of honesty and honesty with Democratic Korean officials we have met with over the past four days,' he said.

Google output: 'I think the relationship has reached a level that we have not seen in terms of honesty and frankness with the DPRK officials we have met with over the past four days,' he added.

Lilt output: 'I think the relationship has reached a level that we have not witnessed in terms of sincerity and sincerity with democratic Korean officials we have met in the past four days.'

Yandex output: Added 'I think the relationship has reached the level of what we see in terms of honesty and frankness with officials of Korea the Democrats we met with them during the last four days.'

The reference human translations can be found in Appendix (Eight).

The evaluators were asked to rate the Adequacy and Fluency of the four English translations of the Arabic source (above) – a sentence consisting of 23 words – on a scale of 1 to 4.

a) Adequacy:

Figure 7.1 (below) illustrates how the evaluators (ten Arabic-English translators) ranked the four MT systems for Adequacy: the ranking reflects the degree to which the English translations expressed the meaning of the Arabic source.

[62] The transliteration of this and all the following Arabic sentences was accomplished using Google Transliteration.

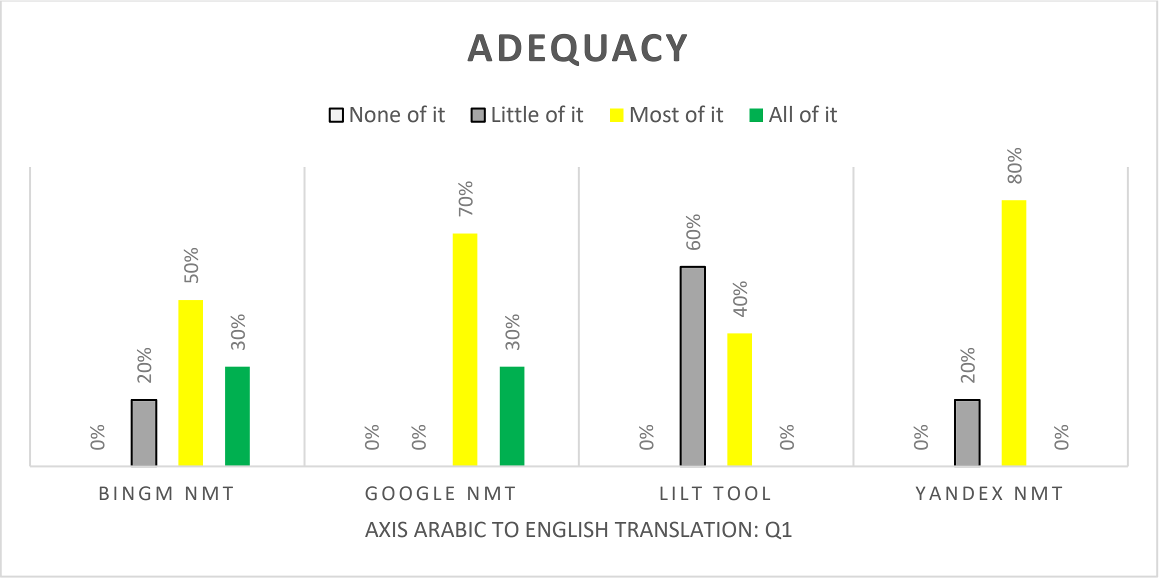


Figure 7.1: Ranking levels of Adequacy for the four MT systems: Arabic to English (Q1)

In the above chart, the evaluators' combined scores, represented by the coloured (yellow and green) bars, suggest that the English translations conveyed a high level of meaning: Google ranked top (100%), followed by Yandex and Bing M. (both 80%). The combined scores in the shaded (light and dark-grey) bars suggest the informativity of the translations was poor, with Lilt ranking the worst (60%).

The results show that none of the evaluators thought Yandex conveyed all of the meaning of the source sentence, while over a quarter thought that Bing M. and Google conveyed the whole meaning. The scores increased when the respondents were asked if the MT systems conveyed most of the meaning: over three-quarters thought that Yandex, slightly less than three-quarters thought that Google, half thought that Bing M. and over a third thought that Lilt conveyed most of it. In terms of the systems that expressed a poor level of meaning, slightly less than two-thirds thought that Lilt conveyed little of the meaning, whereas two (out of ten) respondents thought that Bing M. and the same percentage thought that Yandex conveyed little of it.

b) Fluency:

Figure 7.2 (below) illustrates how the evaluators (five native speakers of English) ranked the four MT systems' English translations in terms of Fluency: the ranking reflects the extent to which the English translations were perceived as grammatically well-formed, free of errors in spelling, idiomatic terms, titles and names, and conformed to the sort of language a native speaker would naturally use.

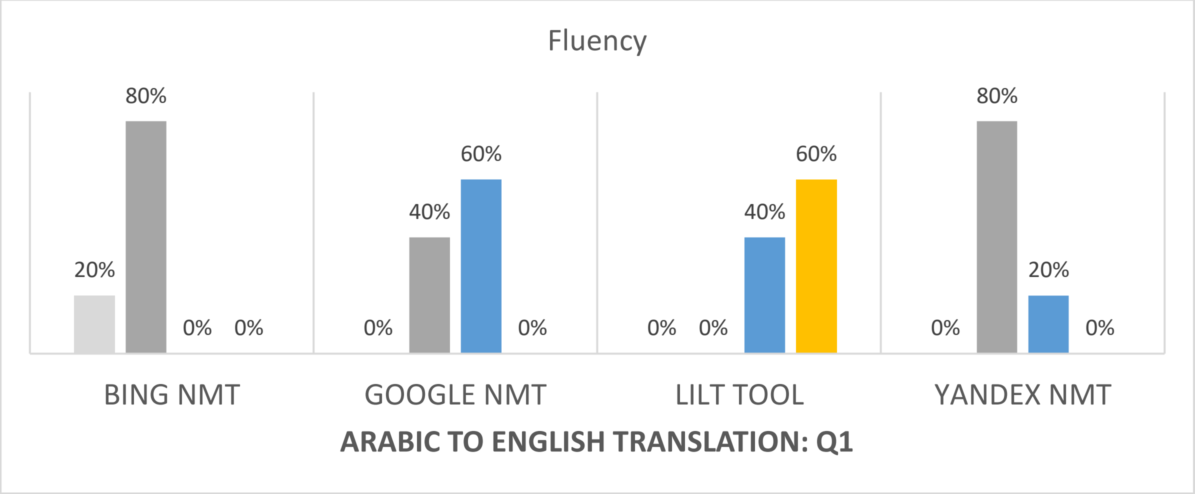


Figure 7.2: Ranking levels of Fluency for the four MT systems: Arabic to English (Q1)

In Figure 7.2, the combined scores of the coloured bars represent a high level of Fluency: Lilt ranked top (100%) in this respect, followed by Google (60%). The combined scores of the shaded bars represent a poor level of Fluency: Bing M. was rated as extremely poor (100%).

The results show that none of the respondents thought that Bing M., Google or Yandex produced stylistically flawless translations, but three (out of five) respondents judged Lilt's output to be flawless. The scores increased when the respondents were asked if the MT systems conveyed a good translation: three judged Google's, two judged Lilt's, and one judged Yandex's translations as good. However, four thought that Bing M. and Yandex, and two considered that Google produced disfluent translations. One respondent judged Bing M.'s translation to be incomprehensible, whereas no one thought that the translations produced by Google, Lilt and Yandex were incomprehensible.

Table 7.1 presents an average of the mean scores for the four MT systems' output in Question 1 (Q1) in terms of Adequacy and Fluency, and the scores of BLEU and hLEPOR. The mean score of human judgement is computed out of four since the MT output was distributed over four levels, while the scores of the metrics are ranges between 1.0 and 0.0: a perfect match is a score of 1.0, whereas a perfect mismatch is a score of 0.0.

Parameter	Bing M. NMT	Google NMT	Lilt Tool	Yandex NMT
Adequacy	3.1	3.3	2.4	2.8
Fluency	1.8	2.6	3.6	2.2
BLEU scores[63]	0.5076	0.4966	0.5280	0.3775
hLEPOR scores[64]	0.6451	0.6737	0.6737	0.6338

Table 7.1: Average of the mean scores of the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q1)

The table shows that the mean Adequacy scores ranged between 2.4 and 3.3, while the mean Fluency scores ranged between 1.8 and 3.6. Google's translation led those of the other systems in terms of Adequacy, while Lilt's led in terms of Fluency.

i) Google's mean score was higher than the other three systems' (3.3 out of 4) for Adequacy, Bing M. was in second place with 3.1, Yandex was third with 2.8, while the Lilt tool came bottom with 2.4.

ii) Lilt exchanged places with Google at the top as regards Fluency: Lilt was assigned an average Fluency score of 3.6 and Google was assigned 2.6, while Yandex and Bing M. scored 2.2 and 1.8, respectively.

The slightly higher scores for Adequacy in Bing M., Google and Yandex, however, suggest that MT systems might be slightly better at producing adequate translations than they are fluent ones. In contrast to the other three MT systems, however, the Lilt tool was assigned a relatively higher score for Fluency than for Adequacy.

[63] Reference translations come from LDC-2004T18 (Linguistic Data Consortium) corpus

[64] Reference translations come from LDC-2004T18 (Linguistic Data Consortium) corpus

The inter-rater reliability was poor with -0.15 for Adequacy and was moderate with 0.59 for Fluency (Exact Fleiss' kappa). Therefore. impossible to draw clear conclusions, but there is a tendency for lower agreement among a larger number of raters to be seen as best.

The same translations were evaluated using the BLEU and hLEPOR automated metric. The table shows that the BLEU scores ranged between 0.3775 and 0.528, while the hLEPOR scores ranged between 0.6338 and 0.6737. The average scores of hLEPOR outperformed BLEU's, whichever the MT system.

i) hLEPOR scored the four systems' translations above 63%: the Google and Lilt outputs provided the highest score with 0.6737 each, followed by Bing M. with 0.6451 and finally Yandex with 0.6338.

ii) BLEU scored the output of the four systems at less than 53%: Lilt's output was also assigned a higher score at 0.5280 (in hLEPOR, its score was 0.6737), Bing M. was in second place with 0.5076, Google was third with 0.4966, while Yandex came bottom with 0.3775.

Google was rated the best in terms of both Adequacy and Fluency. Google Translate, alongside the Lilt tool, achieved the highest hLEPOR score, while Lilt scored the highest in BLEU.

Question 2:

واوضح هذا الاكتشاف ان بعض الحيوانات الثديية كبيرة الحجم التى عاشت فى هذا العصر قد تكون
.آكلة للحوم ولديها الشجاعة الكافية لمنافسة الديناصورات على الغذاء ومكان المعيشة

Transliteration: *wa'awdah hdha alaiktishaf 'ana bed alhayawanat althadiiyat kabirat alhajm alta eashat fa hdha aleasr qad takun akilatan lilhawm waladayha alshajaeat alkafiat limunafasat aldynaswrat ealaa alghidha' wamakan almaeisha*

Bing M. output*:* Some large mammals that lived in this age may be carnivorous and have the courage to compete with dinosaurs for food and living space, the discovery said.

Google output*:* This discovery indicated that some large-sized mammals that lived in this age may be carnivores and have the courage to compete with dinosaurs for food and the place of living.

Lilt output*:* The discovery said some large mammal animals that lived in this era may be a meat eater and have the courage to compete with dinosaurs for food and living space.

Yandex output*:* He explained this discovery to some of the mammals of great size who lived in this era may be carnivorous and have the courage to compete with dinosaurs for food and knowledge.

Again, the evaluators rated the Adequacy and Fluency of the four different English translations above – a sentence consisting of 27 words – using a scale of 1 to 4.

a) Adequacy:

Figure 7.3 (below) displays how the evaluators rated the extent to which the meaning of the Arabic sentence in Q2 was represented by the English translations produced by the four MT systems.

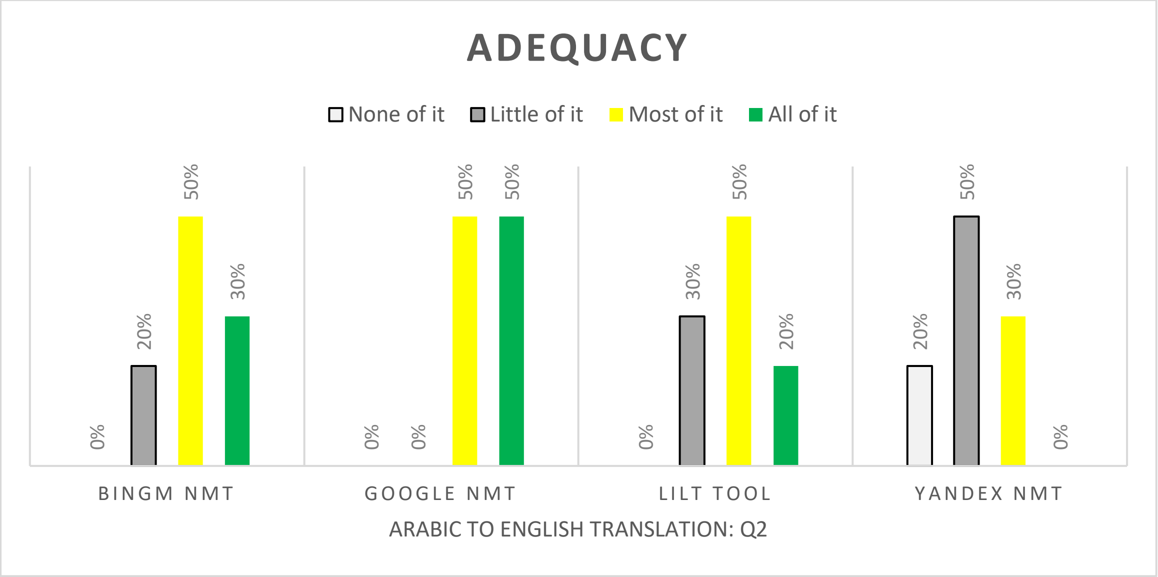


Figure 7.3: Ranking levels of Adequacy for the four MT systems: Arabic to English (Q2)

In the above chart, the combined scores of the coloured bars place the Google translation at the top (100%) and Bing M. as very high (80%), while the combined scores of the shaded bars rank the level of Yandex's translation as Low (70%).

In terms of Adequacy, the results show that none of the evaluators thought that Yandex conveyed all of the meaning of the source sentence, while half judged that Google, three (out of ten) respondents judged that Bing M. and two judged that Lilt failed to convey the full meaning. Around half of the evaluators thought that Bing M., Google and Lilt conveyed most of the meaning, while over a quarter thought that Yandex did. However, half thought that Yandex conveyed little of the meaning, whereas over a third thought that Lilt and over a quarter that Bing M. also conveyed little of the sense of the Arabic sentence. No one thought that Bing M., Google or Lilt conveyed none of the meaning, while two thought that Yandex failed to convey any of the meaning.

b) Fluency:

Figure 7.4 (below) displays the evaluators' estimation of the Fluency of the four MT systems' English translations.

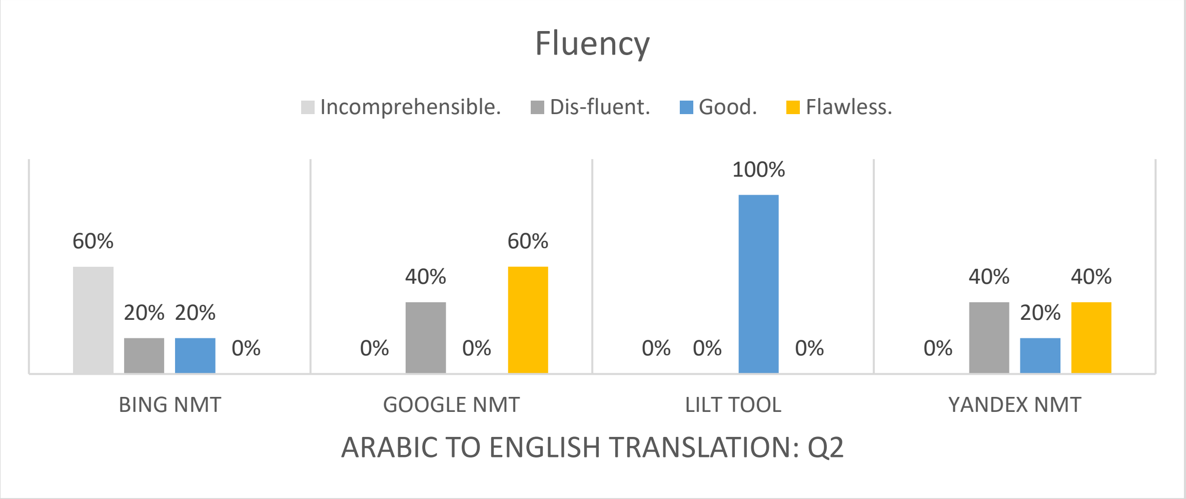


Figure 7.4: Ranking levels of Fluency for the four MT systems: Arabic to English (Q2)

In the above chart, the combined scores of the coloured bars suggest that Lilt came out top (100%), while the combined scores of the shaded bars rank Bing M. as high (80%).

In terms of Fluency, the results show that none of the respondents thought that Bing M. or Lilt produced a stylistically flawless translation, but three (out of five) respondents judged Google's and two respondents considered Yandex's translations to be fluent. However, the five respondents all thought that Lilt produced a good translation, while one judged Bing M.'s and one considered Yandex's as good. In terms of the systems that were thought to produce translations that exhibited a poor level of Fluency, two respondents considered that Google, two that Yandex, and one that Bing M. produced disfluent translations. Three respondents judged that Bing M. produced an incomprehensible translation, whereas no one thought that the translations produced by Google, Lilt or Yandex were incomprehensible.

Table 7.2 (below) presents the average of the mean scores for the Adequacy and Fluency, in addition to the BLEU and hLEPOR scores, of the four MT systems' output for Q2.

Parameter	**Bing M. NMT**	**Google NMT**	**Lilt Tool**	**Yandex NMT**
Adequacy	3.1	3.5	2.9	2.1
Fluency	1.6	3.2	3	3
BLEU scores	0.6012	0.5891	0.4457	0.3532
hLEPOR scores	0.6643	0.5944	0.6360	0.5244

Table 7.2: Average of the mean scores for the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q2)

Table 7.2 shows that the mean Adequacy scores ranged from 2.1 to 3.5, while the mean Fluency scores ranged from 1.6 to 3. 2. Google outperformed the other systems for both criteria:

i) In terms of Adequacy, Google's mean score was the highest (3.5), Bing M. was in second position (3.1), Lilt was third (2.9) and Yandex came bottom (2.1).

ii) Google's mean score was also assigned a higher average Fluency score (3.2), followed by Lilt and Yandex with 3 each, and finally Bing M. with 1.6.

The ranking of Bing M. and Google was slightly higher for Adequacy than for Fluency, whereas the ranking of Lilt and Yandex was slightly higher for Fluency than Adequacy.

The inter-rater reliability was poor with -0.03 for Adequacy and was moderate with 0.55 for Fluency.

Regarding the scores of metrics, the table shows that the values of BLEU ranged between 0.3532 and 0.6012, while hLEPOR values ranged between 0.5244 and 0.6643. The average scores of hLEPOR outperformed BLEU's scores, whichever the system.

i) hLEPOR scored the outputs of the four systems above 52%: the score of Bing M. was higher than the other three systems (0.6643), Lilt was in second place (0.6360), Google was third (0.5944), while the Yandex tool came bottom (0.5244).

ii) BLEU scored the output of Bing M. and Google relatively higher than that for Lilt and Yandex: Bing M. was also assigned a higher average score with 0.6012, followed by Google with 0.5891 (in hLEPOR, the Bing and Google outputs scored 0.6643 and 0.5944, respectively). The output of Lilt and Yandex scored very low in BLEU: Lilt scored 0.4457 and Yandex, 0.3532.

Google was rated the best for Adequacy and Fluency, while Bing M. scored the highest in both hLEPOR and BLEU.

Question 3:

وقال ان اوغندا لا يمكن ان تتفاوض. اننا اصغر من ان نتفاوض انك لا تستطيع ان تكون ضعيفا وتتفاوض.

Transliteration: *waqal 'ana 'uwghanda la ymkn 'an ttfawd. 'anana 'asghar min 'an natafawad 'iinak la tastatie 'an takun daeifanaan watatafawad*

Bing M. output*:* Uganda cannot negotiate. We are too small to negotiate that you cannot be weak and negotiate.

Google output*:* He said that Uganda could not negotiate. We are too young to negotiate that you cannot be weak and negotiate.

Lilt output*:* He said Uganda could not negotiate. We are too small to negotiate that you cannot be weak and negotiated.

Yandex output*:* He said that Uganda could not negotiate. I'm too young to negotiate you can't be weak and negotiating.

The evaluators rated the four different English translations according to their levels of Adequacy and Fluency – a sentence consisting of 19 words – using a 1-4 scale.

a) Adequacy:

Figure 7.5 (below) demonstrates how the evaluators ranked the Adequacy of the four MT systems' output.

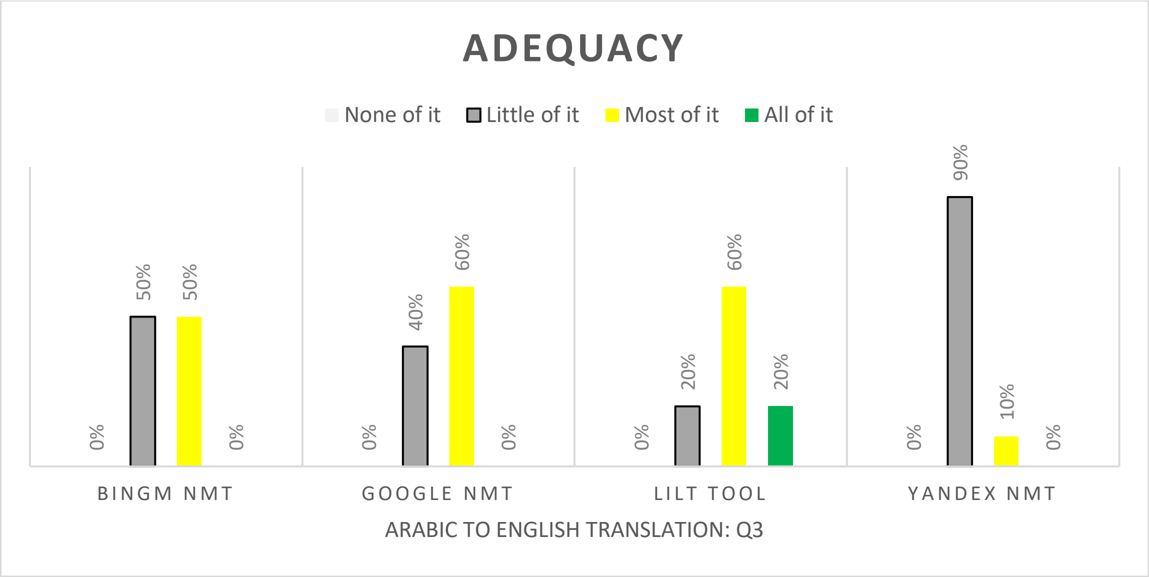


Figure 7.5: Ranking levels of Adequacy for the four MT systems: Arabic to English (Q3)

In the above chart, the combined scores of the coloured bars show that Lilt was ranked highly (80%), while around 90% of the evaluators found Yandex's output poor.

When judging which MT system rendered an adequate translation of the Arabic source, it appears that the evaluators thought that neither Bing M., Google nor Yandex conveyed all of the meaning, while two (out of ten) respondents thought that Lilt conveyed it all. Obviously, the scores increased when the systems were judged on whether they conveyed most of the sense of the Arabic sentence: slightly less than two-thirds of the respondents thought that Google and Lilt, and half thought that Bing M. conveyed most of it, while only one respondent thought that Yandex did. Regarding the systems that produced a less-than-adequate translation, the majority of the respondents nine (out of ten) thought that Yandex conveyed little of the meaning, while half thought that Bing M., over a third thought that Google, and slightly less than a quarter thought that Lilt did. However, no one thought that any of the MT systems conveyed none of the sense of the source sentence.

b) Fluency:

Figure 7.6 (below) demonstrates how the evaluators ranked the four MT systems' output in terms of Fluency.

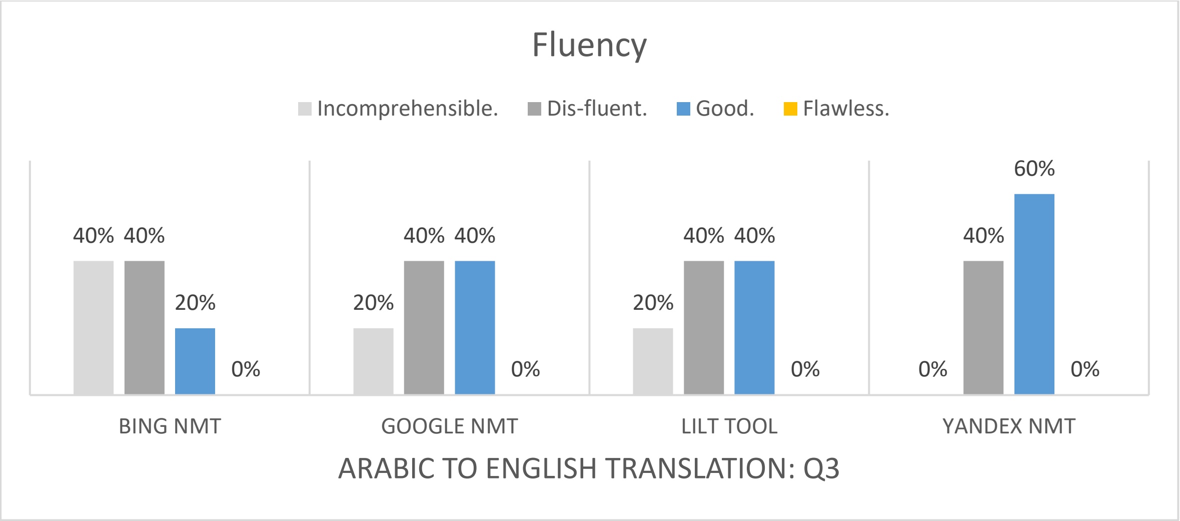


Figure 7.6: Ranking levels of Fluency for the four MT systems: Arabic to English (Q3)

In the above chart, the combined scores of the coloured bars show that 60% thought Yandex was good, while Bing M.'s was ranked as disfluent.

The results show that no one judged that the four systems provided a stylistically flawless translation, while three (out of five) respondents judged that Yandex's, two considered Google's, two considered Lilt's, and one considered Bing M.'s output as good. Regarding a poor level of Fluency, each MT system was judged by two respondents as producing a disfluent translation, whereas the same number of respondents judged Bing M.'s output, one judged Google's and one judged Lilt's as incomprehensible. The average of the mean scores for the Fluency and Adequacy of the MT systems' output for Q3, in addition to the BLEU and hLEPOR scores, is presented in Table 7.3 (below):

The average of the mean scores for the Fluency and Adequacy, in addition to the BLEU and hLEPOR scores, of the MT systems' output for Q3 is presented in Table 7.3 (below):

Parameter	Bing M. NMT	Google NMT	Lilt Tool	Yandex NMT
Adequacy	2.5	2.6	3.0	2.1
Fluency	1.8	2.2	2.2	2.6
BLEU scores	0.3876	0.3484	0.4303	0.3484
hLEPOR scores	0.6250	0.7692	0.7843	0.7692

Table 7.3: Average of the mean scores for the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q3)

As this table shows, the mean Adequacy scores ranged from 2.1 to 3.0, while the mean Fluency scores ranged from 1.8 to 2.6. Lilt's translation led those of the other systems in terms of Adequacy, while Yandex's led in terms of Fluency:

i) In terms of Adequacy, the Lilt mean score was rated first – 3.0 out of 4, Google was in second place with 2.6, Bing M. was third with 2.5, and Yandex was in final place with 2.1.

ii) Yandex was assigned an average Fluency score 2.6, followed by Google and Lilt with 2.2 each, and finally Bing M. with 1.8.

The ranking of Bing M. and Google and Lilt was slightly higher for Adequacy than Fluency, whereas Yandex was assigned a relatively higher score for Fluency than Adequacy.

The inter-rater reliability was in slight agreement with 0.04 for Adequacy and was fair with 0.21 for Fluency.

With regard to the automatic evaluation, the BLEU scores ranged between 0.3484 and 0.4303, while the hLEPOR scores ranged between 0.6250 and 0.7843. hLEPOR's scores outperformed BLEU's, whichever the system.

i) hLEPOR scored the outputs of the four systems above 62%: Lilt's score was very high (0.7843), followed by the outputs of Google and Yandex with 0.7692 each, and finally Bing M. (0.6250).

ii) BLEU scored the systems' output less than 43%: Lilt's was also assigned a higher-than-average BLEU score (0.4303) (in hLEPOR, Lilt scored 0.7843), followed by Bing M. with 0.3876, and finally Google and Yandex with 0.3484 each.

This suggests that the low scoring was mainly caused by incorrect translation of the Arabic. However, Lilt was rated the best in terms of Adequacy and Fluency, and was also awarded the highest scores by both hLEPOR and BLEU.

Question 4:

يذكر ان اخر مأساة اسفرت عن خسائر جسيمة فى صفوف السويديين وقعت عام 1994 عندما غرق مركب فى بحر البلطيق, الامر الذى ادى الى غرق 892 شخصا بينهم 551 سويديا.

Transliteration: *yudhkar 'ana akhir masat 'asfarat ean khasayir jasimat fa sufuf alsuwidiiyn waqaeat eam 1994 eindama gharaq markab fa bahr albltyq, al'amr aldha 'adaa 'iilaa gharaq 892 shakhsaan baynahum 551 suidiaan*

Bing M. output*:* The last tragedy, which caused heavy damage to Swedes, occurred in 1994 when a boat sank in the Baltic Sea, killing 892 people, including 551 soy.

Google output*:* It is noteworthy that the last tragedy resulted in massive losses among the Swedes, which occurred in 1994 when a boat sank in the Baltic Sea, which resulted in the drowning of 892 people, including 551 Sue Idia.

Lilt output*:* The latest tragedy caused heavy losses among the Swedes in 1994 when a boat drowned in the Baltic Sea, which sank 892 people, including 551.

Yandex output*:* Recall that another tragedy resulted in heavy losses in the ranks of the Swedes occurred in 1994 when the boat sank in the Baltic Sea, which led to the sinking of 892 people, including 551 Su-Lydian.

The four different English translations for Q4 were rated according to their Adequacy and Fluency on a sentence consisting of 30 words using a scale of 1 to 4.

a) Adequacy

Figure 7.7 (below) illustrates the evaluators' rating of the Adequacy of the four systems' output.

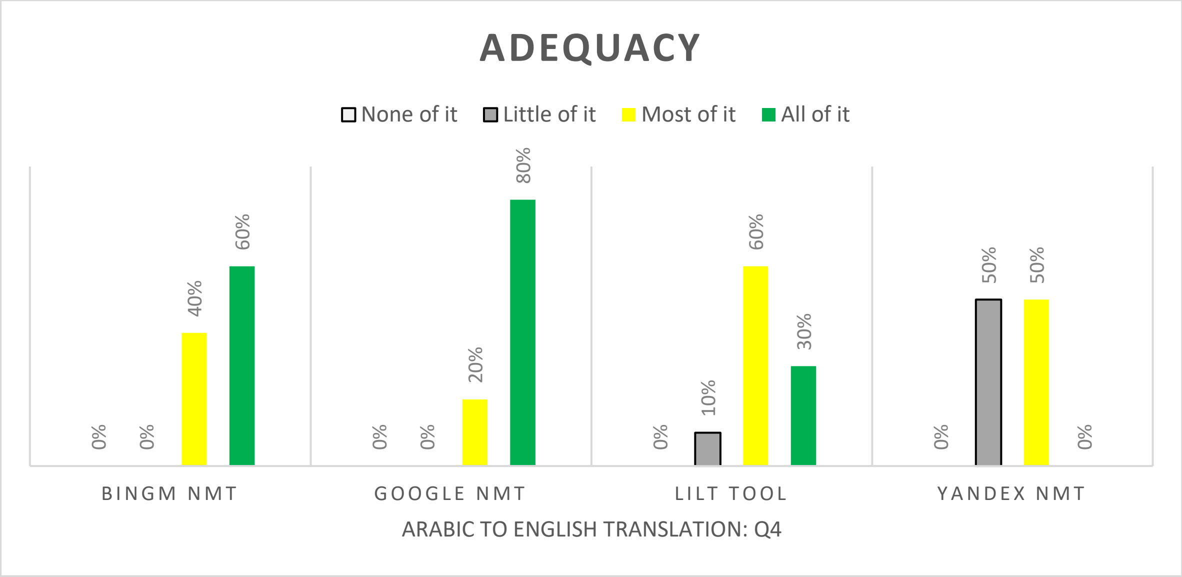


Figure 7.7: Ranking levels of Adequacy for the four MT systems: Arabic to English (Q4)

In the above chart, the combined scores of the coloured bars show that Google and Bing M. were ranked at the top (100% each), followed by Lilt (90%), with Yandex performed relatively low (50%).

In terms of which systems expressed the meaning with a high level of Adequacy, the results show that although none of the respondents thought that Yandex conveyed all of the meaning, over three-quarters judged that Google conveyed it all, while slightly less two-third thought that Bing M. and over a quarter thought that Lilt did. Lilt and Yandex were rated better in conveying most of the meaning: slightly less than two-thirds of the evaluators thought that Lilt and half thought the Yandex conveyed most of the meaning, whereas over a third thought that Bing M. and two (out of ten) respondents thought that Google expressed most of it. However, for those systems regarded as delivering a poor level of Adequacy, half thought Yandex conveyed little of the meaning, whereas only one respondent considered that Lilt's translation was inadequate. No one thought that any one of the systems conveyed none of the sense of the source sentence.

b) Fluency:

Figure 7.8 (below) shows how the evaluators rated the Fluency of the four systems' output.

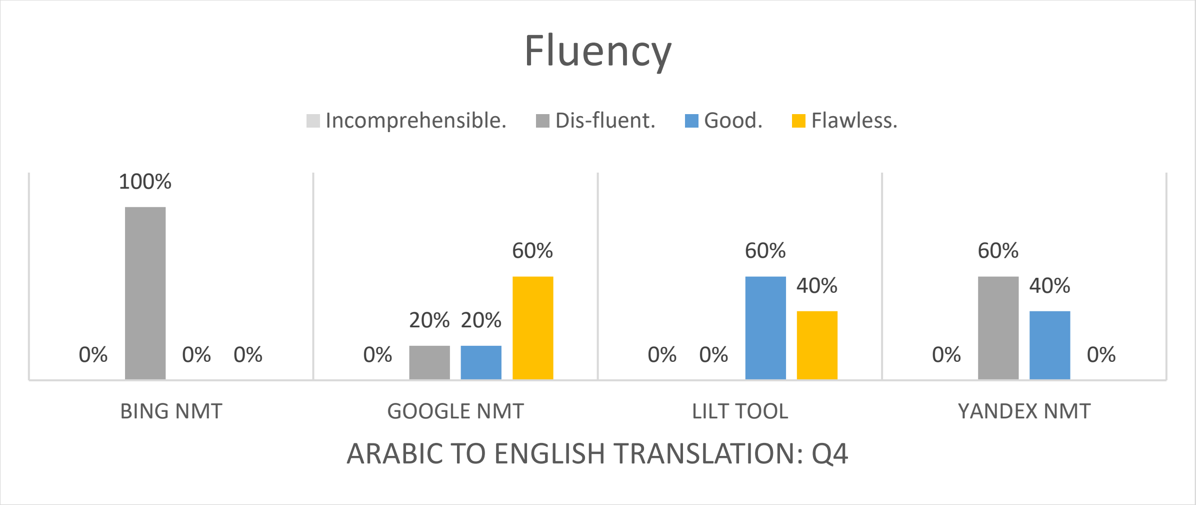


Figure 7.8: Ranking levels of Fluency for the four MT systems: Arabic to English (Q4)

In the above chart, the combined scores of the coloured bars show that Lilt was ranked top (100%), followed by Google (80%), indicating that the evaluators considered these systems' translations as relatively fluent, while all the evaluators found Bing M.'s output disfluent.

In terms of the systems that provided a high level of Fluency, the results show that although none of the evaluators judged that Bing M. and Yandex produced a stylistically flawless translation, three (out of five) respondents judged that Google and two judged that Lilt produced flawless translations. Three respondents considered that Lilt, two that Yandex and one that Google produced good translations. In terms of those systems that provided a poor level of Fluency, all the respondents considered that Bing M. produced a disfluent translation, whereas three respondents thought that Yandex's and one thought that Google's translation was disfluent. However, no one considered that Lilt produced a disfluent or incomprehensible translation.

Table 7.4 (below) presents the average of the mean scores for the four MT systems' output for Q4 in terms of Adequacy and Fluency, in addition to the scores of BLEU and hLEPOR:

Parameter	Bing M. NMT	Google NMT	Lilt Tool	Yandex NMT
Adequacy	3.6	3.8	3.1	2.5
Fluency	2	3.4	3.4	2.4
BLEU scores	0.6135	0.5774	0.4585	0.4887
hLEPOR scores	0.5095	0.6748	0.5095	0.4938

Table 7.4: Average of the mean scores for the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores(Q4)

As Table 7.4 shows, the mean Adequacy scores ranged from 2.5 to 3.8, while the mean Fluency scores ranged from 2 to 3.4. Google's translation led those of the other systems in terms of Adequacy, while Google shared the lead with Lilt over the other systems in terms of Fluency.

i) Google's mean score was very high for Adequacy (3.8), Bing M. was second with 3.6, Lilt was third with 3.1, and finally Yandex was in last place with 2.5.
ii) Google and Lilt were rated first with a mean score of 3.4, while Yandex and Bing M. were rated 2.4 and 2, respectively.

The ranking of Bing M. and Google and Lilt was slightly higher for Adequacy than Fluency, whereas Yandex was assigned a relatively higher score for Fluency than Adequacy.

It seems that the ranking of all four systems were slightly higher for Adequacy than for Fluency. This suggests that MT systems might be slightly better at producing adequate translations than they are fluent ones.

The inter-rater reliability was slight agreement with 0.06 for Adequacy, and was substantial agreement with 0.65 for Fluency which was the most reliable.

With respect to the metric scores, BLEU's ranged between 0.4585 and 0.6135, while the hLEPOR scores ranged between 0.4938 and 0.6748. hLEPOR scored the output of Google, Lilt and Yandex higher, while BLEU scored the output of Bing M. higher.

i) hLEPOR scored the output of Google relatively high with 0.6748, followed by Bing M. and Lilt with 0.5095 each, and finally Yandex with 0.4938.

ii) BLEU scored the output of Bing M. the best (0.6135) (in hLEPOR, Bing's output scored 0.5095), Google was in second place (0.5774), Yandex was third (0.4887), while the Lilt tool came bottom (0.4585).

Google was rated the best in terms of Adequacy, and scored the highest in hLEPOR, while Bing M. was rated the best for Fluency, and scored the highest in BLEU.

Summary

Table 7.5 contains an overview of the mean scores for all four sentences and the four MT systems' Arabic-to-English translations (A = Adequacy, F = Fluency).

NMT system	**SQ1**	**SQ2**	**SQ3**	**SQ4**	**Overall score**	**BLEU score**	**hLEPOR score**
Bing M.	A:3.1 F:1.8	A:3.1 F:1.6	A:2.5 F:1.8	A:3.6 F:2	A:3.07 F:1.8	**0.5274**	0.6109
Google	A:3.3 F:2.6	A:3.5 F:3.2	A:2.6 F:2.2	A:3.8 F:3.4	**A:3.30** F:2.85	0.5028	**0.6780**
Lilt Tool	A:2.4 F:3.6	A:2.9 F:3	A:3.0 F:2.2	A:3.1 F:3.4	A:2.85 **F:3.05**	0.4656	0.6508
Yandex	A:2.8 F:2.2	A:2.1 F:3	A:2.1 F:2.6	A:2.5 F:2.4	A:2.37 F:2.55	0.3919	0.6053

Table 7.5: Overall scores for Adequacy and Fluency, and the BLEU and hLEPOR scores, for the four NMT systems (Arabic-to-English translation)

In summary, when ranking the systems' English translations of Arabic sentences, the participants' preferences varied from system to system. However, the overall mean scores reveal that the most adequate translations were produced by Google, while the most fluent translations were produced by Lilt. More specifically, the overall mean scores gathered from the data in Table 7.5 show that, in terms of Adequacy, Google scored best (an average of 3.30), followed by Bing M. (an average of 3.07), with third place occupied by Lilt (an average of 2.85). At the bottom was Yandex with an average of 2.37. Lilt exchanged places with Google at the top as regards Fluency: Lilt was assigned first place (an average of 3.05), followed by Google (an average of 2.85), then Yandex (an average of 2.55) and finally Bing M. (an average of 1.8). Overall, the average scores suggest that the Bing M.'s Google's and Yandex';s translations from Arabic to English expressed a slightly higher level of Adequacy than Fluency. In contrast to the other three MT systems, however, the Lilt tool was assigned a relatively higher score for Fluency than for Adequacy.

In terms of the automatic evaluation, hLEPOR computed Google as the best (0.6780), followed by the Lilt tool (0.6508), while Bing NMT was placed third (0.6109 and Yandex was in last place (0.6053). BLEU computed Bing NMT as the best (0.5274), followed by Google (0.5028), Lilt was placed third (0.4656), while Yandex gave the poorest result (0.3919). The highest value scored by hLEPOR was 0.6780, while the highest BLEU score was 0.5274.

In the Arabic-to-English translation, there was some correlation in the results: Table 7.5 shows that the evaluators judged Google's translation to be the best in terms of Adequacy (3.30 out of 4) and Fluency (3.12 out of 4). In hLEPOR, Google scored 0.6780 out of 1.0, while Bing M. achieved the best BLEU score of 0.5274 out of 1.0. The overall performance score was high – above 50% – whichever the evaluation method.

7.1.3 English-to-Arabic NMT translations

Four English source sentences (Q5-8) were translated into Arabic using the same four MT systems.[65] The source sentence and the different versions of the translation are displayed below.

Question 5:

De Villepin also said that the number of practising Muslims in France is in about the same range as for other religions, in other words less than 10% out of five million people.

[65] The tests were translated in March 2020

Bing M. output:

وقال دو فيلبان أيضاً إن عدد المسلمين الممارسين في فرنسا هو في نفس النطاق تقريباً كما هو الحال بالنسبة للأديان الأخرى، وبعبارة أخرى أقل من 10% من أصل خمسة ملايين شخص.

Back translation: De Villepin also said that the number of Muslim practitioners in France is about the same as other religions, in other words less than 10% out of five million people.

Google output:

قال دو فيلبان أيضًا أن عدد المسلمين المتدينين في فرنسا في نفس النطاق تقريبًا كما هو الحال في الديانات الأخرى ، أي أقل من 10٪ من أصل خمسة ملايين شخص.

Back translation: De Villepin also said that the number of religious Muslims in France is in roughly the same range as in other religions, less than 10% of the five million people.

Lilt output:

وقال دي فيلبين أيضاً أن عدد المسلمين الذين يمارسون مهنة في فرنسا في نفس النطاق الذي يتدربون عليه في الديانات الأخرى، وبعبارة أخرى أقل من 10 في المائة من خمسة ملايين شخص.

Back translation: De Villepin also said that the number of Muslims practicing a profession in France is in the same range as they train in other religions, in other words less than 10 percent of the five million people.

Yandex output:

وقال دي فيلبان أيضا إن عدد المسلمين الممارسين في فرنسا هو في نفس النطاق تقريبا بالنسبة للديانات الأخرى ، وبعبارة أخرى أقل من 10 ٪ من أصل خمسة ملايين شخص.

Back translation: De Villepin also said that the number of practicing Muslims in France is in roughly the same range as for other religions, in other words less than 10% out of five million people.

The four different versions of Arabic translations were rated by the evaluators on a sentence consisting of 33 words using a scale of 1-4 for Adequacy and Fluency.

a) Adequacy:

Figure 7.9 (below) illustrates how the translations of the four MT systems were ranked by the respondents in terms of Adequacy.

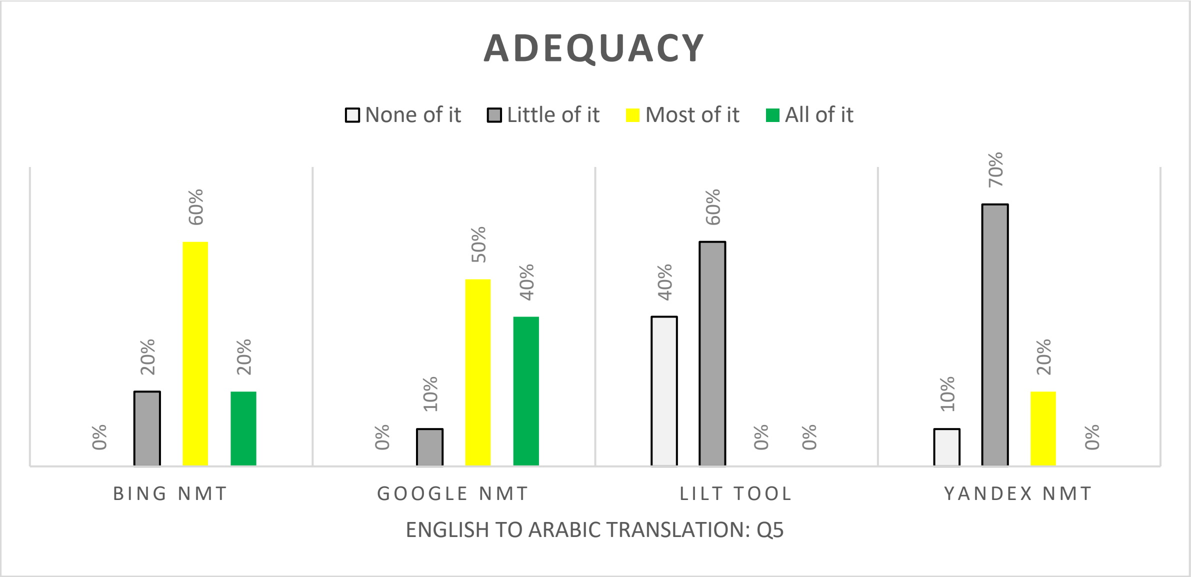


Figure 7.9: Ranking levels of Adequacy in the four MT systems: English to Arabic (Q5)

In the above chart, the combined scores of the coloured bars rank Google at the top (90%), followed by Bing M. (80%). In contrast, Lilt was given the lowest level (i.e. poor Adequacy), followed by Yandex.

In terms of the translations that conveyed the meaning of the source sentence highly adequately, the results show that although no one thought that Lilt's and Yandex's output expressed all of the meaning, over a third of the respondents considered that Google's and two (out of ten) respondents that Bing M.'s did. The same percentage thought that Yandex expressed most of the meaning, while over half judged that Bing M. and half thought that Google expressed it most adequately, but none considered that Lilt did. Referring to a poor level of Adequacy, over two thirds of the evaluators judged that Yandex and slightly less than two-thirds judged that Lilt expressed little of the meaning, whereas around two thought that Bing M.'s translations and one respondent thought that Google's also produced little of the meaning. Furthermore, one respondent thought that the translation produced by Yandex was incomprehensible, while slightly less than half thought that Lilt's was equally poor.

b) Fluency

Figure 7.10 (below) illustrates how the four Arabic translations in Q5 were ranked for Fluency.

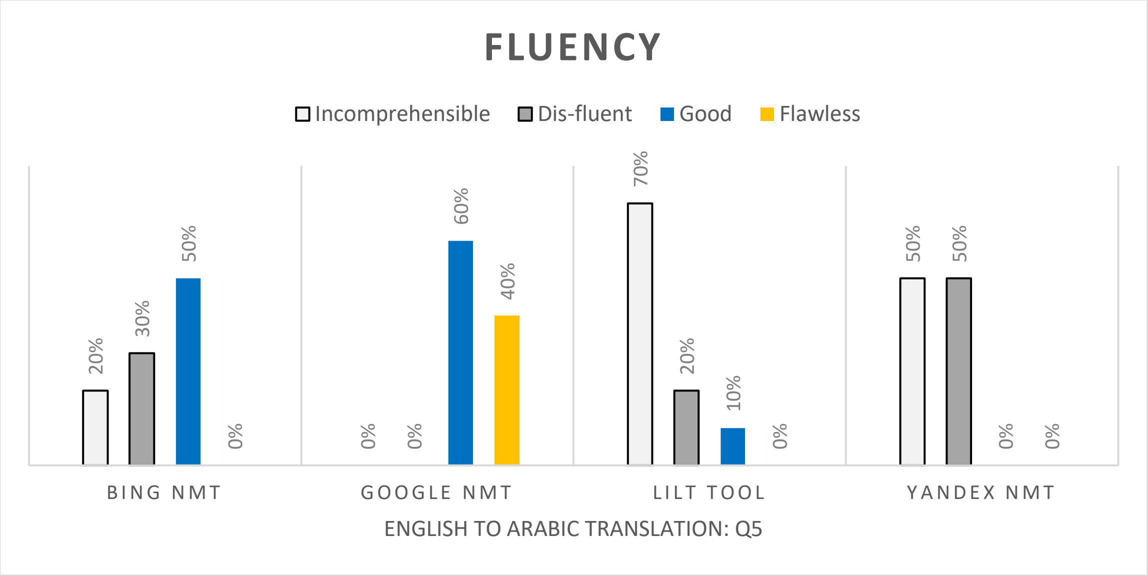


Figure 7.10: Ranking levels of Fluency in the four MT systems: English to Arabic (Q5)

The above chart shows that Google's output was also ranked at the top (100%) for Fluency, while Lilt's and Yandex's was ranked as disfluent.

More specifically, the results show that although none of the evaluators judged Bing M., Lilt and Yandex as providing stylistically flawless translations, over one-third judged Google's as flawless. Also, no one judged Yandex as providing a good translation, but only one respondent thought that Lilt, half thought that Bing M. and over half thought that Google provided translations with a good degree of Fluency. In terms of the systems whose output had a poor level of Fluency, half of the respondents judged the translation by Yandex as disfluent, while three judged Bing M.'s and two Lilt's as disfluent. Similarly, two judged that Bing M. provided an incomprehensible translation, while nearly two-thirds judged Lilt's and half judged Yandex's as incomprehensible.

The average of the mean scores for the four MT systems' output for Q5 in terms of Fluency and Adequacy, and the scores of BLEU and hLEPOR, is presented in Table 7.6 (below). The mean scores of the human judgements were also calculated out of four as the systems' output was distributed over four levels, while a perfect match of the automatic metrics is a score of 1.0 and a perfect mismatch, a score of 0.0.

Parameter	Bing M. NMT	Google NMT	Lilt Tool	Yandex NMT
Adequacy	3.0	3.3	1.6	2.1
Fluency	2.3	3.4	1.4	1.5
BLEU scores	0.1843	0.2238	0.2570	0.2664
hLEPOR scores	0.4528	0.4166	0.3383	0.4942

Table 7.6: Average of the mean scores for the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q5)

Table 7.6 shows that the mean Adequacy scores ranged from 1.6 to 3.3, and the mean Fluency scores from 1.4 to 3.4. Google was ahead of the other systems in both Adequacy and Fluency.

i) In terms of Adequacy, Google's mean score was rated highest (3.3), Bing M. was second (3.0) and Yandex third (2.1). The poorest performing system was Lilt (1.6).

ii) Google was also assigned the highest Fluency score with 3.4, followed by Bing M. with 2.3, Yandex with 1.5, and finally Lilt with 1.4. It can be seen that Bing M., Lilt and Yandex achieved slightly higher scores for Adequacy than for Fluency, while Google was assigned a relatively higher score for Fluency than for Adequacy.

The slightly higher scores for Adequacy in Bing M., Lilt and Yandex suggest that MT systems might be slightly better at producing adequate translations than they are fluent ones. In contrast to the other three MT systems, however, Google was assigned a relatively higher score for Fluency than for Adequacy.

The inter-rater reliability was slight agreement with 0.17 for Adequacy and with 0.097 for Fluency.

In terms of the automatic evaluation, both metrics scored the output of the four systems very low – less than 50% – although the average hLEPOR scores outperformed those of BLEU whichever the system. The BLEU scores ranged between 0.1843 and 0.2664, while hLEPOR's ranged between 0.3383 and 0.4942.

i) hLEPOR scored Yandex’s output higher than those of the other three systems, at 0.4942, while Bing M. was assigned second place with 0.4528, Google was third with 0.4166, and the Lilt tool came last with 0.3383.

ii) BLEU also scored Yandex’s output the highest at 0.2664 (in hLEPOR, the highest score was 0.4942), followed by Lilt (0.2570), Google (0.2238) and Bing M. (0.1843).

This suggests that the low scoring was mainly caused by the failure to produce correct translations in Arabic. However, Google was rated the best in terms of both Adequacy and Fluency, while Yandex scored the highest in hLEPOR and BLEU.

Question 6:

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Bing M. output:

بدا Orejuela هادئة جدا كما كان يجري نقله إلى الطائرة الأمريكية التي من شأنها أن تأخذه إلى ميامي في
ولاية فلوريدا.

Back translation: Orejuela looked very quiet as he was being transported to an American plane that would take him to Miami, Florida.

Google output:

بدا Orejuela هادئًا تمامًا لأنه كان يقود إلى الطائرة الأمريكية التي ستأخذه إلى ميامي في فلوريدا.

Back translation: Orejuela seemed completely calm as he was driving to the American plane that would take him to Miami, Florida.

Lilt output:

(بدا هادئاً تماماً بينما كان يقود إلى الطائرة الأمريكية التي ستأخذه إلى (ميامي) في (فلوريدا Orejuela

Back translation: Orejuela looked completely calm while driving to the American plane that would take him to Miami, Florida.

Yandex output*:*

أوريخويلا بدت هادئة تماما كما كان يجري أدى إلى الطائرة الأمريكية التي من شأنها أن تأخذه إلى ميامي في
ولاية فلوريدا.

Back translation: Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida

The evaluators were asked to rate the Adequacy and Fluency of the four English translations of the Arabic source (above) – a sentence consisting of 21 words – on a scale of 1 to 4.

It was observed that the proper noun 'Orejuela' was left untranslated in English in Bing M.'s, Google's and Lilt's versions. Further, the English word jumped to the left-hand side in both Bing M.'s and Google's Arabic output, in addition to the Arabic word that precedes it. These issues would definitely have had an effect on the evaluators' responses. Their ratings of the Adequacy and Fluency of the four different translations are presented below.

a) Adequacy

Figure 7.11 (below) displays the respondents' evaluation of the extent to which the meaning of the English source was expressed by the different systems' translations.

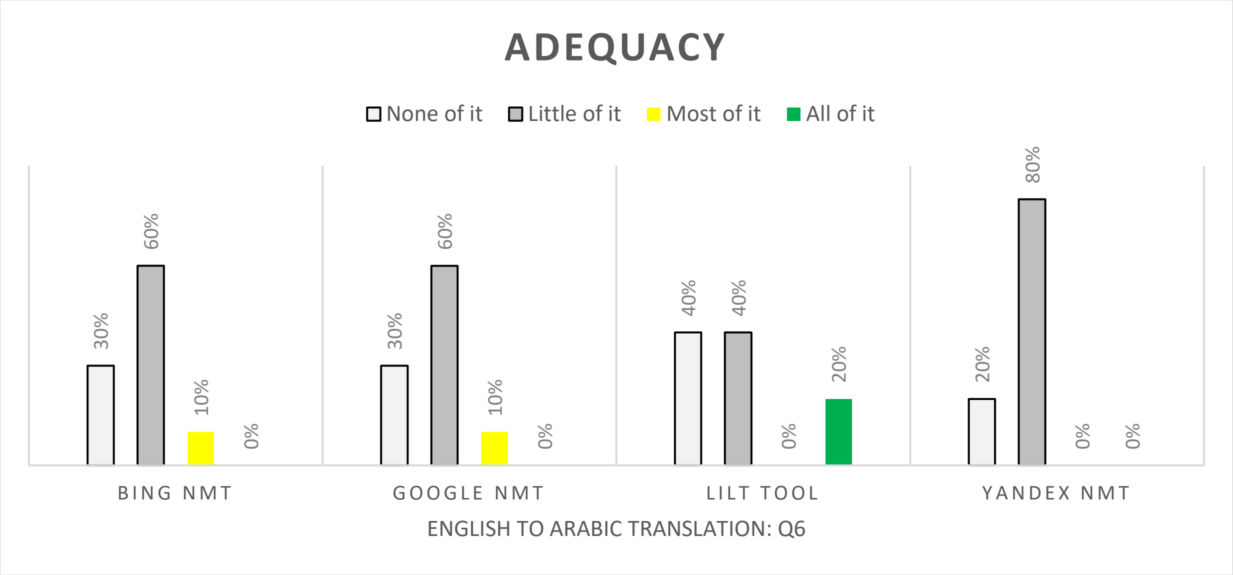


Figure 7.11: Ranking levels of Adequacy in the four MT systems: English to Arabic (Q6)

The above chart reflects how the evaluators assessed the Adequacy levels of the translations when certain words were left untranslated. The combined scores of the shaded bars show a poor level of Adequacy for all the systems – Yandex was judged 100% poor.

The results show that no one thought that the output of Bing M., Google or Yandex expressed all of the meaning, while only two (out of ten) respondents thought that Lilt expressed it all. The low rating of Bing M. and Google could be rationalised by the fact that the non-translation of the proper noun led to further issues which impacted the target's meaning. However, one respondent of the evaluators thought that Bing M. and the same ratio thought that Google expressed most of the meaning, while none thought that Yandex or Lilt did. On the other hand, over three-quarters thought that Yandex, over half thought that Bing M. and Google, and four thought that Lilt expressed little of the meaning. The same percentage judged that Lilt and over a quarter judged that Bing M., Google and Yandex expressed none of the sense of the source text.

There was noticeable disagreement among the evaluators concerning Lilt. Also, around 20% found that Bing M. expressed all of the meaning, while 40% found its translation expressed none of the meaning. This may be explained by the impact of the non-translated proper noun, which caused the word to jump in the target text.

b) Fluency

Figure 7.12 (below) displays the evaluators' rating of the Fluency of the Arabic translations produced by the four MT systems.

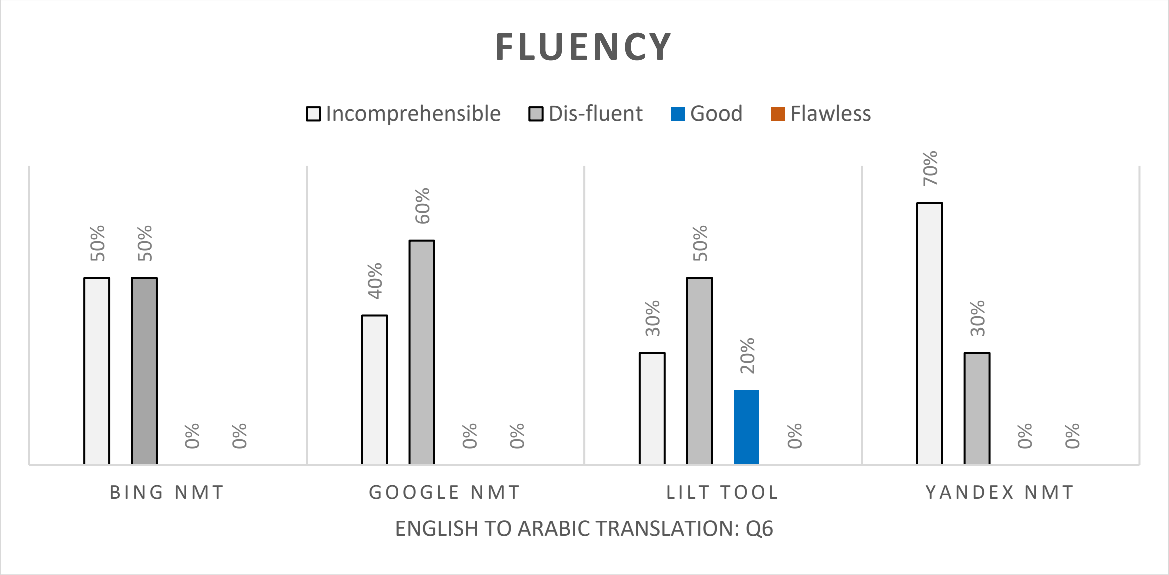


Figure 7.12: Ranking levels of Fluency in the four MT systems: English to Arabic (Q6)

In a similar way to the figures for the Adequacy ratings, Figure 7.12 reflects the fact that the evaluators' responses regarding the Fluency of the MT output appear to have been influenced by the non-translated word and its impact on the quality of the Arabic target text. The combined scores of the shaded bars indicate a poor level of Fluency for all the systems.

The results show that no one judged that the four systems provided a stylistically flawless translation. Except for two respondents who found that Lilt provided a good level of Fluency, none of the evaluators thought that Bing M., Google or Yandex produced good translations in terms of Fluency. On the other hand, over half judged that Google and half that Bing M. and Lilt provided disfluent translations, whereas over a quarter judged that Yandex's output was disfluent. The same percentage judged Lilt's translation as incomprehensible, while over two-thirds thought that Yandex's, half that Bing M.'s, over a third that Google's were incomprehensible.

Table 7.7 presents the average of the mean scores for the four MT systems' translations of Q6 in terms of Adequacy and Fluency, in addition to the BLEU and hLEPOR scores.

Parameter	Bing M. NMT	Google NMT	Lilt Tool	Yandex NMT
Adequacy	1.8	1.8	2.0	1.8
Fluency	1.5	1.6	2.1	1.3
BLEU scores	0.3404	0.2382	0.2881	0.2979
hLEPOR scores	0.5389	0.3592	0.2409	0.2994

Table 7.7: Average of the mean scores for the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q6)

Table 7.7 shows that the mean Adequacy scores ranged from 1.8 to 2.0, while the mean Fluency scores ranged from 1.3 to 2.1. Lilt beat the other systems in the two criteria.

i) In terms of Adequacy, Lilt was rated first with a mean score of 2.0, while Bing, M. Google and Yandex shared second place with 1.8 each.

ii) Lilt's mean score for Fluency was also the best (2.1), while Google, Bing M. and Yandex were rated 1.6, 1.5 and 1.3, respectively.

Obviously, the Arabic translations of Q6 received a very low rating – except for Lilt, which performed slightly better – although the English source sentence does not seem especially difficult. It is suggested that the low rating was mainly caused by the non-translation of the proper noun 'Orejuela', which led to a number of consequences in the Arabic script, such as the word jumping to the left-hand side. However, further errors were also committed by Bing M. (mistranslation of gender) and Google and Lilt (mistranslation of the passive voice).

The inter-rater reliability was poor with -0.0203 for Adequacy and with 0.0206 for Fluency.

Regarding the metric scores, those of BLEU ranged between 0.2382 and 0.3404, while hLEPOR's ranged between 0.2409 and 0.5389. However, hLEPOR's average scores outperformed BLEU's, whichever the system.

i) hLEPOR scored the output of Bing M. higher (0.5389) than those of the other three systems. Google scored 0.3592, Yandex, 0.2994 and the Lilt tool, 0.2409.

ii) BLEU also scored the Bing M. output the highest with 0.3404 (in hLEPOR, the highest score was 0.5389), while Yandex, the Lilt tool and Google were rated 0.2979, 0.2881 and 0.2382, respectively.

This suggests that the low scoring was mainly caused by producing incorrect translations in Arabic. However, Lilt was rated the best in terms of Adequacy and Fluency, while Bing M. scored the highest in both hLEPOR and BLEU.

Question 7:

The statement said that the Arab taking a position,[66] especially its intellectuals and thinkers, is a historical responsibility whose surrender is a surrender of the right to expression and a betrayal of every message and value, spiritually and culturally.

Bing M. output:

وقال البيان ان اتخاذ العرب لموقف وخاصة مثقفه ومفكريه مسؤولية تاريخية ان استسلامها هو تنازل عن حق
. التعبير وخيانة لكل رسالة وقيمة روحيا وثقافيا

Back translation: The statement said that the Arab taking a position, especially intellectuals and intellectuals, has a historical responsibility that its surrender is an abdication of the right of expression and a betrayal of every message and spiritual and cultural value.

Google output*:*

وقال البيان إن العرب الذين يتخذون موقفا ، وخاصة المثقفين والمفكرين ، مسؤولية تاريخية ، واستسلامها هو
.تنازل عن الحق في التعبير وخيانة لكل رسالة وقيمة روحيا وثقافيا

[66] Although this is not grammatically correct English, we cannot intervene in the corpus sentence.

Back translation: The statement said that the Arabs who take a position, especially the intellectuals and thinkers, are a historical responsibility, and their surrender is a surrender of the right to expression and a betrayal of every spiritual and cultural message and value.

Lilt output*:*

وقال البيان إن العرب الذين يتخذون موقفا، ولا سيما المفكرين والمفكرين، مسؤولية تاريخية، يستسلمون الحق في
التعبير وخيانة لكل رسالة وقيمتها وروحيا وثقافيا.

Back translation: The statement said that Arabs who take a position, especially thinkers and thinkers, are a historical responsibility, surrender the right to expression and betray every message and its value, spiritually and culturally.

Yandex output*:*

وقال البيان ان العربية اتخاذ موقف ، وخاصة المثقفين والمفكرين هي المسؤولية التاريخية التي الاستسلام هو
الاستسلام للحق في التعبير و الخيانة من كل رسالة قيمة روحيا وثقافيا.

Back translation: The statement said that the Arab take a position, especially intellectuals and thinkers are historical responsibility that surrender is surrender to the right to express and betrayal of every message valued spiritually and culturally.

The evaluators rated the Adequacy and Fluency of the four different Arabic translations – a sentence consisting of 39 words – using a scale of 1-4.

a) Adequacy

Figure 7.13 (below) demonstrates how the evaluators ranked the Adequacy of four MT systems' output in Arabic.

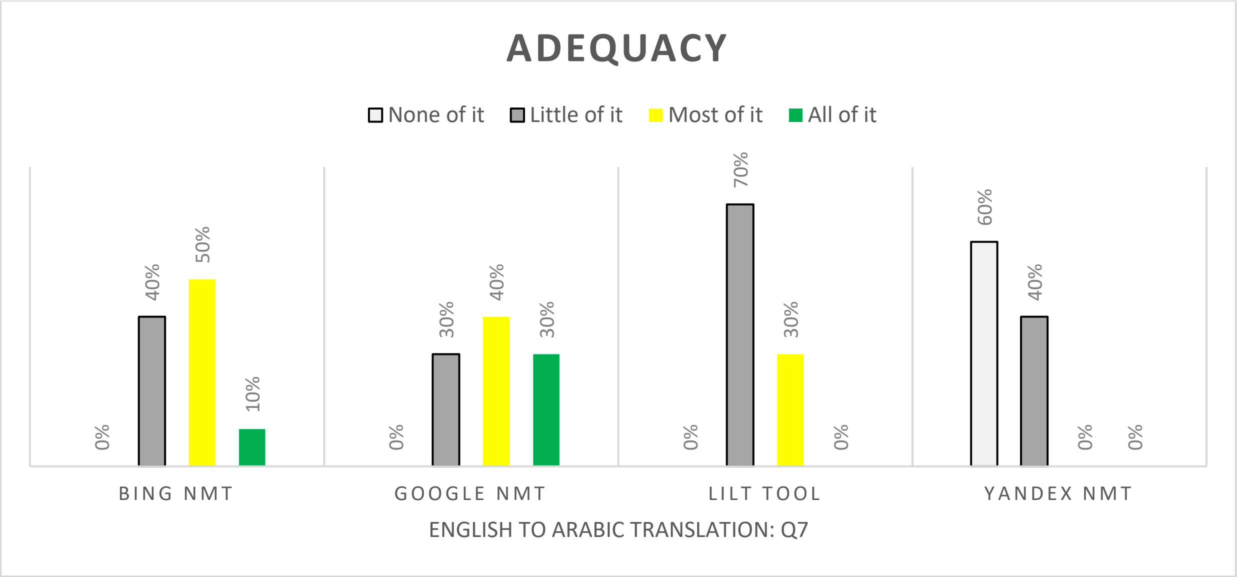


Figure 7.13: Ranking levels of Adequacy in the four MT systems: English to Arabic (Q7)

The above chart shows that the total score of the coloured bars ranked Google as high – around 70% – in contrast to Yandex, which was found to be poor (100%).

In terms of a high level of Adequacy, the results show that none of the evaluators thought that Lilt's or Yandex's output expressed all of the original meaning, while over a quarter thought that Google and one respondent that Bing M. expressed it all. Over a quarter thought that Lilt, a third that Google and half that Bing M. expressed most of it. In terms of a poor level of Adequacy, over two-thirds judged that Lilt, over a third considered that Bing and Yandex, and over a quarter thought that Google expressed little of the meaning. In contrast, although no one thought that Bing M, Google and Lilt expressed none of the meaning, nearly two-thirds of the respondents thought that Yandex failed to convey any of the meaning of the source text.

b) Fluency

Figure 7.14 (below) demonstrates how the evaluators ranked the Fluency of the four MT systems' Arabic output.

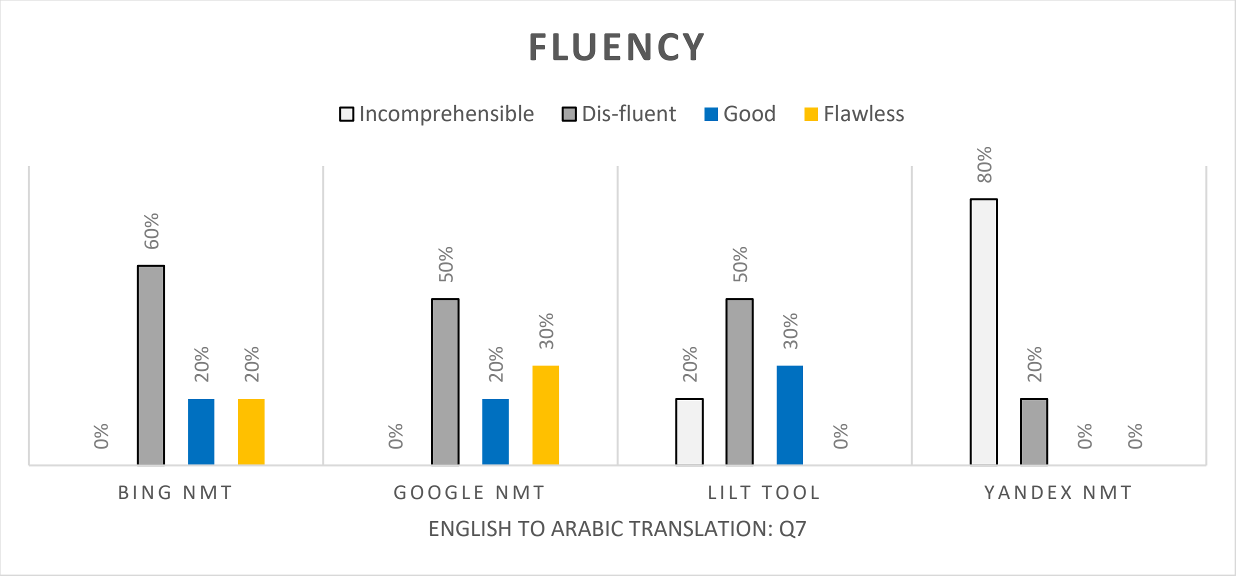


Figure 7.14: Ranking levels of Fluency in the four MT systems: English to Arabic (Q7)

The above chart shows that the combined scores of the coloured bars were lower than the shaded bars, which indicates that all four translations lacked Fluency.

In terms of a high level of Fluency, none of the evaluators judged that either Lilt or Yandex provided stylistically flawless translations. However, three (out of ten) respondents thought that Google and twothat Bing M. provided a flawless output, while two rated Bing M.'s and Google's level of Fluency as good, and over a quarter also judged Lilt's as good. Regarding a poor level of Fluency, over half of the respondents thought that Bing M., half thought that Google and Lilt, and 20% judged that Yandex provided disfluent translations. Two respondents judged Lilt's and over three-quarters judged Yandex's as incomprehensible.

The average of the mean scores for the four MT systems regarding the Fluency and Adequacy, in addition to the scores of BLEU and hLEPOR, of the translations of Q7 is presented in Table 7.8 (below).

Parameter	Bing M. NMT	Google NMT	Lilt Tool	Yandex NMT
Adequacy	2.7	3.0	2.3	1.4
Fluency	2.2	2.8	2.1	1.2
BLEU scores	0.2544	0.2363	0.2259	0.3085
hLEPOR scores	0.4332	0.4285	0.2909	0.3571

Table 7.8: Average of the mean scores for the four systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q7)

As can be seen in Table 7.8, the mean Adequacy scores ranged from 1.4 to 3.0, while the mean Fluency scores ranged from 1.2 to 2.8. Google led the other systems in the two criteria.

i) In terms of Adequacy, Google's mean score was the best at 3.0, Bing M. was second with 2.7, Lilt was in third place with 2.3, and Yandex was bottom with 1.4.

ii) Google's mean score was also assigned first place for Fluency (2.8), followed by Bing M. (2.2), Lilt (2.1) and Yandex, with a very low score (1.2). The table shows that the evaluators provided slightly higher scores for Adequacy than for Fluency.

Overall, the evaluators' ratings of Adequacy and Fluency for Q7, indicating that Bing M. and Google might perform slightly better than Lilt and Yandex. It is suggested that the low rating was mainly caused by the mistranslation of the phrases 'intellectuals and thinkers' and '…whose surrender is a surrender of…'.

The inter-rater reliability was slight agreement with 0.05 for Adequacy and with 0.08 for Fluency.

With regard to the automatic evaluation, the BLEU scores ranged between 0.2259 and 0.3085, while hLEPOR's scores ranged between 0.2909 and 0.4332. Although, the average scores of the two metrics were very low – less than 50% – the hLEPOR scores were higher than BLEU's, whichever the system evaluated.

i) hLEPOR scored the output of Bing M. the highest with 0.4332, followed by Google (0.4285), Yandex (0.3571) and finally the Lilt tool (0.2909).

ii) BLEU scored the output of Yandex the highest with 0.3085 (in hLEPOR, the highest score was 0.4332). Bing M. was second with 0.2544, Google was in third place with 0.2363, and the Lilt tool was last with 0.2259.

This suggests that the low scoring was mainly caused by incorrect translations in Arabic. Google was rated the best in terms of both Adequacy and Fluency, while Bing M. scored the highest in hLEPOR and Yandex was placed first by BLEU.

Question 8:

Kenteris (31 years old) and Thanou (29 years old) missed a drug test on August 12 on the eve of the Olympic Games in Athens, saying they had been in a motorcycle traffic accident.

Bing M. output:

كان كانينيليس (31 عاما) وثانو (29 عاما) تغيبا عن اختبار المخدرات في 12 اب/اغسطس عشية دورة الالعاب
الاولمبية في اثينا، مؤكدين انهما تعرضا لحادث سير على الدراجات النارية.

Back translation: Cannellis, 31, and Thano, 29, missed a drug test on August 12 on the eve of the Athens Olympics, saying they had been involved in a motorcycle accident.

Google output:

غاب Kenteris (31 عامًا) و Thanou (29 عامًا) عن اختبار مخدرات في 12 أغسطس عشية الألعاب الأولمبية
في أثينا ، قائلين إنهما تعرضا لحادث مروري على دراجة نارية.

Back translation: Kenteris (31) and Thanou (29) missed a drug test on August 12 on the eve of the Athens Olympics, saying they had a traffic accident on a motorcycle

Lilt output:

فقد كان Kenteris (31 سنة) and (29 عاما) اختصار للمخدرات في 12 أغسطس عشية الألعاب الأولمبية
في أثينا، قائلاً أنهما كانا في حادث حركة الدراجات النارية.

Back translation: Kenteris (31) and (29) had an abbreviation for drugs on August 12 on the eve of the Athens Olympic Games, saying they were in a motorcycle traffic accident.

Yandex output:

Kenteris (31 عاما) Thanou (29 عاما) غاب عن اختبار المخدرات في 12 آب عشية الألعاب الأولمبية في
أثينا ، قائلا أنها كانت في دراجة نارية في حادث مروري.

Back translation: Kenteris (31) Thanou (29) missed a drug test on August 12 on the eve of the Olympic Games in Athens, saying that she was on a motorcycle in a traffic accident.

The four different Arabic translations were rated in terms of their Adequacy and Fluency – a sentence consisting of 34 words – using a 1-4 scale.

It was noticed that, as in Q6, the English proper nouns were not translated by Google, Lilt or Yandex. Also, the words jumped to the left-hand side in the translated sentences, in addition to the preceding Arabic word. The Adequacy and Fluency ratings for the four different translations are presented below.

a) Adequacy

Figure 7.15 (below) illustrates how the respondents rated the Arabic translations' Adequacy.

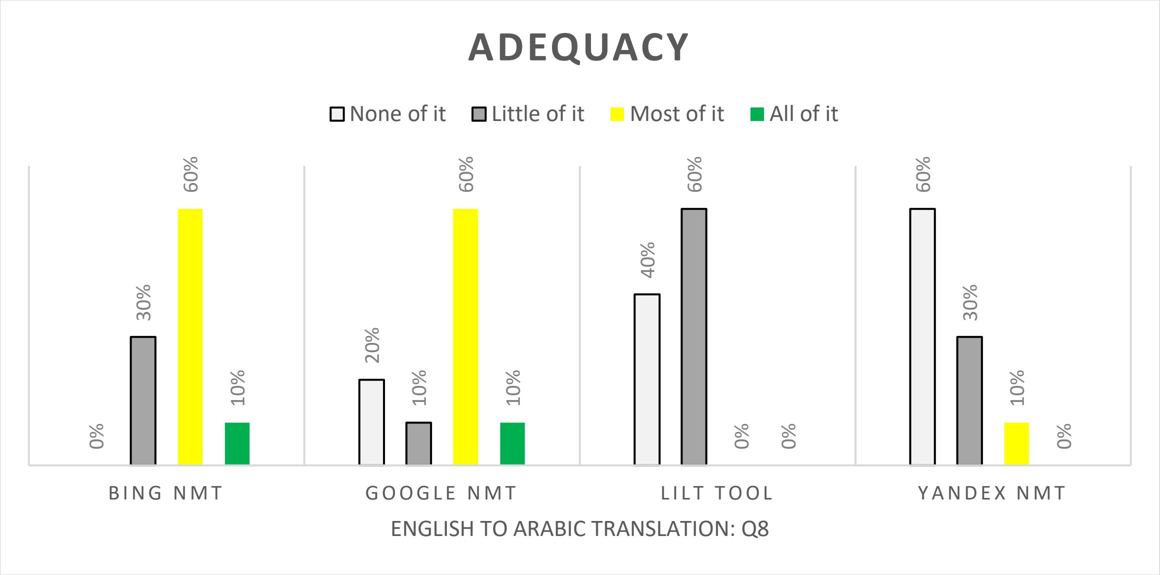


Figure 7.15: Ranking levels of Adequacy in the four MT systems: English to Arabic (Q8)

In the above chart, the combined scores of the coloured bars suggest that the evaluators ranked Google and Bing M. highly (70% each). In contrast, they judged Lilt's output as poor (100%).

In terms of those systems that conveyed the sense of the English source with a high level of Adequacy, the results show that none of the evaluators thought that Lilt and Yandex expressed all of the meaning, whereas one respondent thought that Bing M. and Google did. The same percentage thought that Yandex expressed most of the meaning, while slightly less than two-thirds thought that Bing M. and Google expressed most of it. Regarding a poor level of Adequacy, over half of the respondents thought that Lilt's output and a quarter that Bing M.'s and Yandex's expressed little of the meaning. Further, although no one thought that Bing M, and only two (out of ten) respondents that Google expressed none of the sense of the source sentence, over half the respondents thought that Yandex and over a third thought that Lilt expressed none of its sense.

It can be seen that there was disagreement regarding Google's output: although 10% of the respondents found that the output expressed all of the meaning, 20% judged that it expressed none of it. The difference between the evaluators' responses may be explained by the non-translated word and its impact on the Arabic translation.

b) Fluency

Figure 7.16 (below) presents how the evaluators ranked the Fluency of the Arabic translations.

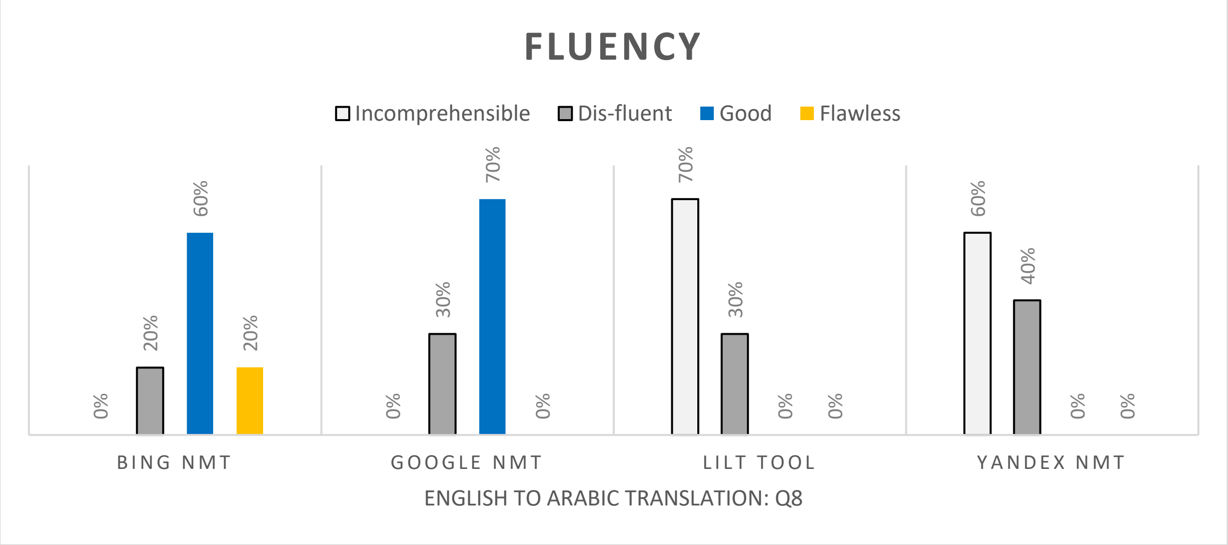


Figure 7.16: Ranking levels of Fluency in the four MT systems: English to Arabic (Q8)

In the above chart, the combined scores of the coloured bars indicate that Bing M. and Google were ranked highly – 80% and 70%, respectively. In contrast, the evaluators judged that the translations produced by Lilt and Yandex were completely poor (100%).

In terms of a high level of Fluency, the results show that none of the evaluators judged that Google, Lilt or Yandex provided stylistically flawless translations, while two (out of ten) respondents judged that neither did Bing M., although the Fluency of its output was judged as good by over half the respondents, while over two-thirds considered Google's translation good. In relation to a poor level of Fluency, over a third thought that Yandex provided a disfluent translation, whereas over a quarter judged Google's and Lilt's, and slightly less judged Bing M.'s translations as disfluent. Although no one thought that Bing M. and Google provided incomprehensible translations, slightly less than two-thirds judged the output of Yandex and over two-thirds judged that of Lilt as incomprehensible.

Table 7.9 presents the average of the mean scores for the Adequacy and Fluency of the translations of Q8 by the four MT systems, in addition to the BLEU and hLEPOR scores.

Parameter	**Bing M. NMT**	**Google NMT**	**Lilt Tool**	**Yandex NMT**
Adequacy	2.8	2.6	1.6	1.5
Fluency	3.0	2.7	1.3	1.4
BLEU scores	0.1936	0.2355	0.3033	0.1936
hLEPOR scores	0.3401	0.3521	0.2112	0.3401

Table 7.9: Average of the mean scores for the four MT systems in terms of Adequacy and Fluency, and the BLEU and hLEPOR scores (Q8)

As can be seen in Table 7.9, the mean Adequacy scores ranged between 1.5 and 2.8, while the mean Fluency scores ranged between 1.3 and 3.0. Bing M. led the other systems in both criteria.

i) In terms of Adequacy, Bing M.'s mean score was the highest at 2.8, Google was in second place with 2.6, Lilt was third with 1.6, and Yandex last with 1.5.

ii) Bing M. was also assigned the highest mean Fluency score (3.0), followed by Google (2.7), Yandex (1.4) and Lilt (1.3). Overall, it can be seen that Lilt and Yandex had slightly higher scores for Adequacy than for Fluency, while Bing M. and Google were assigned a relatively higher score for Fluency than for Adequacy.

Obviously, the Arabic translations of Q8 received a very low rating – except for Bing M., which performed slightly better. It is suggested that the low rating was mainly caused by the non-translation of the proper nouns ‘Kenteris’ and ‘Thanou’.

The inter-rater reliability was slight agreement with 0.11 for Adequacy and with 0.16 for Fluency.

In terms of the metric scores, BLEU’s ranged between 0.1936 and 0.3033, while hLEPOR scores ranged between 0.2112 and 0.3521. Although the average scores of the two metrics were very low – less than 35% – the hLEPOR scores were relatively higher than BLEU’s, and Lilt’s output scored higher in BLEU than in hLEPOR.

i) hLEPOR scored the outputs of Google relatively high with 0.3521, while Bing, M. and Yandex shared second place with 0.3401 each, and Lilt came last with 0.2112.

ii) BLEU, however, scored Lilt’s output the highest with 0.3033 (in hLEPOR, the highest score was 0.3521), Google was second with 0.2355, while Bing, M. and Yandex shared last place with 0.1936 each.

This suggests that the low scoring was mainly caused by the non-translation in Arabic. Nevertheless, Bing M. was rated the best for both Adequacy and Fluency, Bing M. scored the highest in hLEPOR, and Google was placed first in BLEU.

Summary

The following table gives an overview of the mean scores for all four sentences (Q5, Q6, Q7, Q8) and the four MT systems’ English-to-Arabic translations. In the table, A = Adequacy and F = Fluency, as rated by the human evaluators. The automated scores are placed alongside.

NMT system	**Q5**	**Q6**	**Q7**	**Q8**	**Overall score**	**BLEU score**	**hLEPOR score**
Bing M.	A:3.0 F:2.3	A:1.8 F:1.5	A:2.7 F:2.2	A:2.8 F:3.0	A:2.57 F:2.25	0.2431	**0.4412**
Google	A:3.3 F:3.4	A:1.8 F:1.6	A:3.0 F:2.8	A:2.6 F:2.7	**A:2.67** **F:2.62**	0.2334	0.3891
Lilt Tool	A:1.6 F:1.4	A:2.0 F:2.1	A:2.3 F:2.1	A:1.6 F: 1.3	A:1.87 F: 1.72	**0.2685**	0.2703
Yandex	A:2.1 F:1.5	A:1.8 F:1.3	A:1.4 F:1.2	A:1.5 F:1.4	A:1.70 F:1.35	0.2666	0.3727

Table 7.10: Overall scores for Adequacy and Fluency, and the BLEU and hLEPOR scores, for the four NMT systems – English to Arabic

An examination of the ranking of the systems' Arabic translations of the four English sentences shows that the participants' preferences varied. However, the overall scores reveal that the most adequate and most fluent MT system was Google, followed by Bing M., with the other two systems trailing behind them on both these counts. Looking in more detail at the data in Table 7.10 (the overall mean scores), it can be seen that in terms of Adequacy, Google also scored best, with an average of 2.67, Bing M. scored an average of 2.57, third place was occupied by Lilt with an average of 1.87, and in bottom place was Yandex with an average of 1.70. There is no change in the order of the systems when it comes to Fluency: Google was also assigned the top score for Fluency with an average of 2.62, followed by Bing M. with an average of 2.25, then Lilt with an average of 1.72, and finally Yandex with an average of 1.35. Overall, the scores suggest that although the systems' output conveyed a high level of meaning but not a high level of Fluency, the average score was low in comparison to the average scores given for the Arabic-to-English translation. Thus, it could be argued that MT systems perform better when translating from Arabic to English than from English to Arabic.

In terms of the automatic evaluation, hLEPOR calculated that Bing M. produced the best output (0.4412), followed by Google (0.3891), Yandex was placed third (0.3727), while Lilt was in last place with 0.2703. In contrast, BLEU calculated that Lilt produced the best output (0.2685), followed by Yandex (0.2666), although the gap was not wide, and then Bing M. (0.2431). It computed that Google gave the poorest performance (0.2334). The highest value scored by hLEPOR was 0.4412, while the highest BLEU score was 0.2685. This suggests that LEPOR performed better than BLEU.

There was some variation in results in the English-to-Arabic translations. Table 7.11 shows that the evaluators judged Google's translation to be the best for Adequacy (2.67 out of 4) and Fluency (2.62 out of 4). Bing M. achieved the best hLEPOR score (0.4412 out of 1.0), while Lilt produced the highest BLEU score (0.2685 out of 1.0). The overall performance scoring was low over all the evaluation methods, particularly in BLEU.

7.1.4 Translation direction comparison

The outcomes of both the human and automatic evaluation methods show some similarities in that the assessment of the quality of the translation appeared to depend on the translation direction.

7.1.4.1 Evaluators' ratings

The evaluators' rating of the systems' output was higher for the Arabic-to-English translations than for the English-to-Arabic ones. The similarity between the systems in terms of the high ratings achieved by their English output and the low ratings of their Arabic output is presented in Figure 7.17, below.

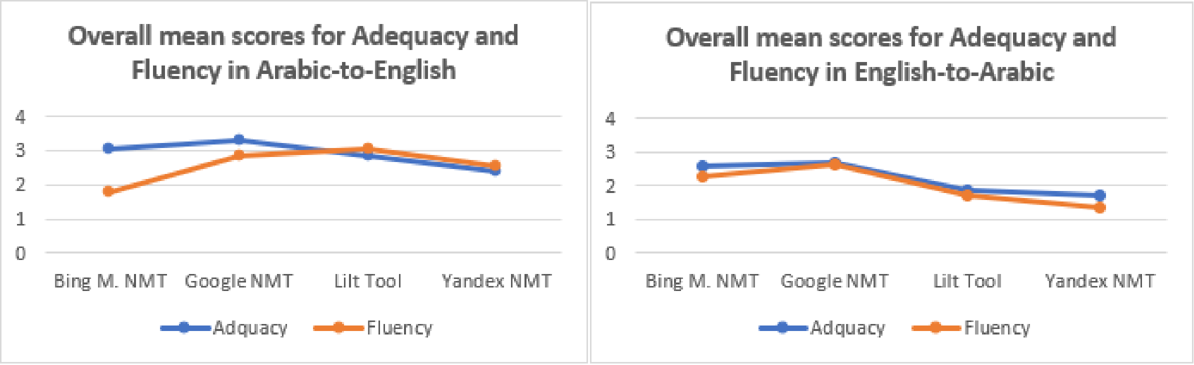


Figure 7.17: Overall mean scores for Adequacy and Fluency in Arabic-to-English translation and vice versa

The figure above clearly shows that, in terms of the Arabic output (on the right), the evaluators consistently rated the systems and their output as relatively similar in terms of Adequacy and Fluency. Regarding the MT systems' English output (on the left), the evaluators rated the Bing M. and Google translations' Adequacy (blue line) as much worse than their Fluency (orange line), whereas Lilt and Yandex were judged as relatively similar for both criteria. The inconsistent rating of the English target may be due to the use of two types of informants: native English speakers were used to rank fluency, while evaluators whose first language was Arabic, with English as a second language, were used to rank adequacy.

7.1.4.2 Automatic ratings

a) hLEPOR scores

The hLEPOR model scored the NMT systems' translations in English higher than their Arabic ones. The similarity in the high rates awarded to the English output and the low rates given to the Arabic output across the systems is displayed in Figure 7.18 (below).

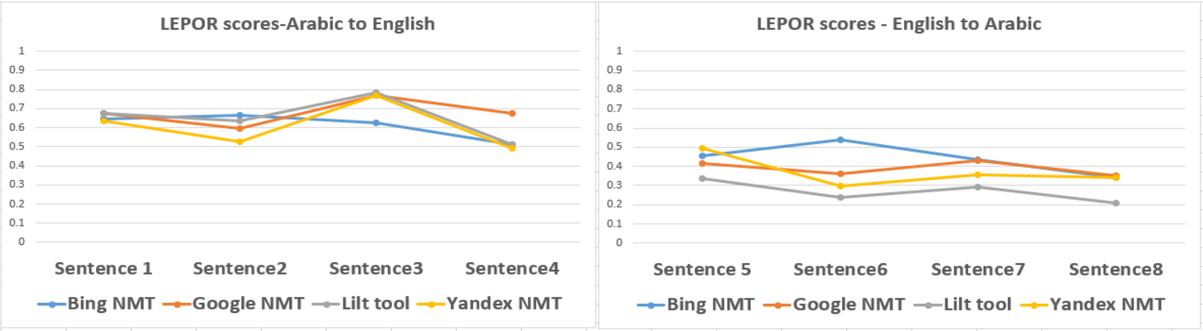


Figure 7.18: The hLEPOR scores for Arabic-to English and English-to-Arabic translations

The figure above clearly shows that hLEPOR's scores of the systems' English output (left side) are higher than its scores for their Arabic output (right side). Furthermore, the metric awarded the systems' English output relatively close scores, while it gave the different systems quite distant scores when their output was Arabic.

b) BLEU scores

Similarly, BLEU gave higher scores to the systems' output in English than in Arabic. The similarity between the different NMT systems in terms of the high rating of their English output and low rating of their Arabic output is displayed in Figure 7.19 (below).

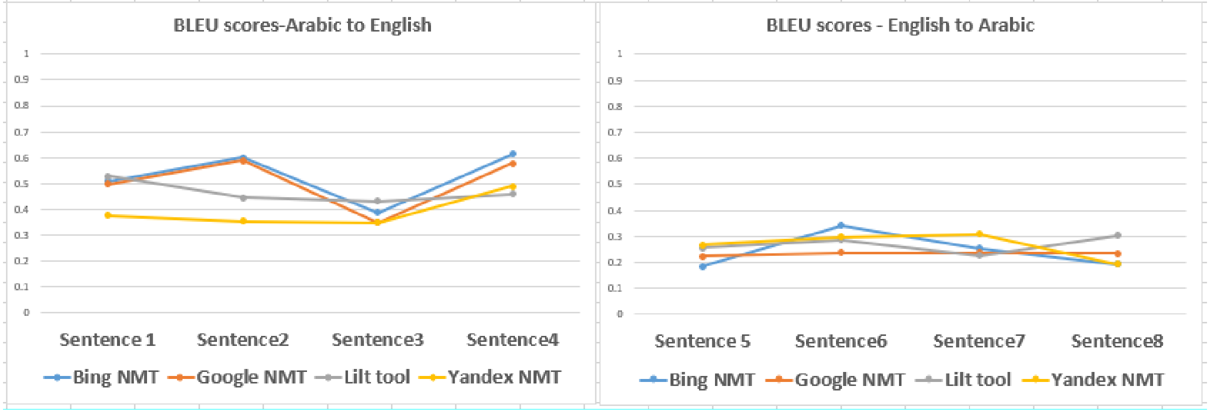


Figure 7.19: The BLEU scores for Arabic-to-English and English-to-Arabic translations

The figure above clearly shows that the BLEU scores for the systems' English output (left side) are relatively higher than for the Arabic output (right side); however, BLEU produced close scores for the systems' output in Arabic.

It can be concluded that, as a whole, the human and automatic evaluations concurred that the quality of the English output of NMT systems was relatively higher than their Arabic output. Further, the results of BLEU – the metric based on lexical similarity – fall into the same band of Fluency ratings, suggesting that the metric performs well in capturing translation Fluency (Lo et al. 2012).

c) BLEU vs hLEPOR scores

The results provided by the automatic metrics show that the hLEPOR scores were closer to 1.0 (a perfect match) than BLEU's scores, whichever the translation direction. In terms of Arabic-to-English translation, Figure 7.20 (below) shows that hLEPOR's scores were relatively higher than BLEU's.

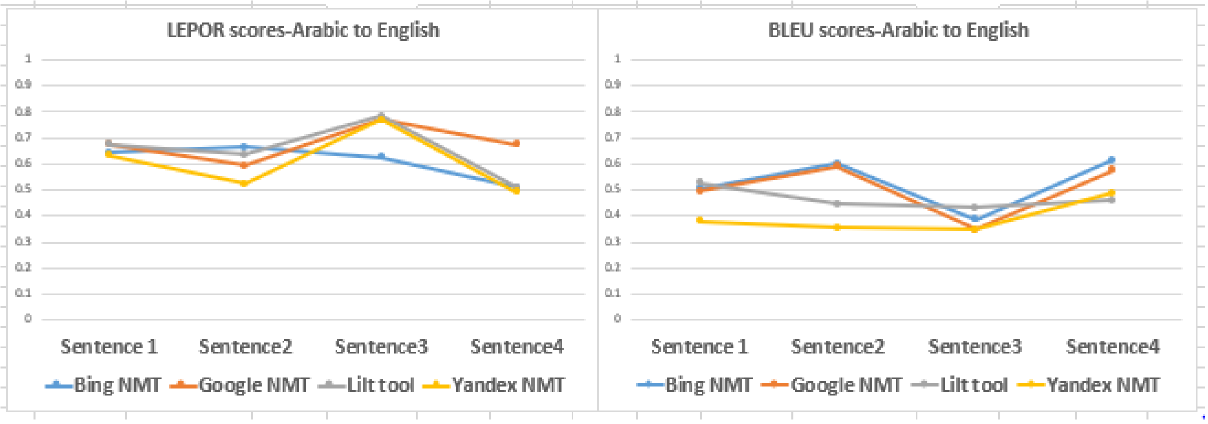


Figure 7.20: hLEPOR scores vs. BLEU scores for Arabic-to-English translation

The model of hLEPOR's high scores occurred despite the fact that only one English translation was used as a reference, while four English translations were used as references for the BLEU evaluation, although it could be assumed that multiple reference translations would increase the match.

Similarly, in terms of English-to-Arabic translation, figure 7.21 (below) shows that the hLEPOR scores were again higher than BLEU's scores.

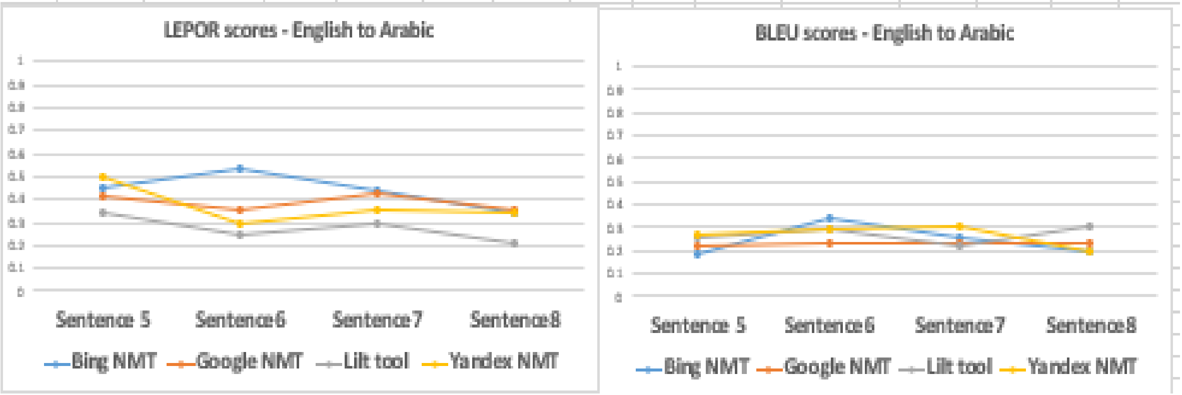


Figure 7.21: hLEPOR scores vs. BLEU scores for English-to-Arabic translations

In both metrics only one Arabic translation was used as a reference translation; however, the hLEPOR scores were higher than those of BLEU. It can be concluded that hLEPOR's scores outperformed BLEU's, whichever the translation direction.

7.2 Discussion

7.2.1 Discussion of findings

The experiment's findings show that the NMT systems produced a good quality of translation in both directions. However, the results suggest that post-editing operations would be needed to adjust the Fluency of their output, especially that of the Bing M. and Google systems.

The level of the Adequacy rating was related to the translation direction – the Arabic-to-English translations were given slightly higher mean scores. The overall score for Google, which received the highest mean score in both directions, was 3.30 out of 4 for its Arabic-to-English translation (see Table 7.5) but 2.67 out of 4 for its English-to-Arabic translation (see Table 7.10). Rankings of 3-4 indicate that the translations were judged to convey 'most of the meaning' and 'all of the meaning', while rankings of 1-2 indicate they conveyed 'none of the meaning' and 'little of the meaning'. The decreasing Adequacy in Google's English-to-Arabic translations may have been partly due to the non-translation of some proper nouns (e.g. 'Orejuela' in Q6), which had major consequences for the Arabic sentence structure. As a result, the participants rated the translated sentence poorly for Adequacy. One reason behind the good-quality translations may have been the style of the test segments which used document-level texts and normal sentence lengths (ranging from 19 to 39 words), and the fact that the neural network architecture benefits from context (Miculicich et al. 2018; Tu et al. 2018).

However, the average scores for Fluency were lower overall than those for Adequacy, and much worse for Bing M.'s and Google's translations from Arabic to English. In terms of the highest Fluency mean score, Google led the other systems in English-to-Arabic translations, while Lilt led in Arabic-to-English translations. Google was ranked 2.62 for English to Arabic (see Table 7.10), while Lilt was ranked 3.05 out of 4 for Arabic to English (see Table 7.5). Rankings of 3-4 indicate that the translations were judged as 'flawless' and 'good', while rankings of 1-2 indicate that they were judged as 'disfluent' and 'incomprehensible' The main advantage of the Fluency judgment was that it was delivered by native speakers of each language pair: the fluency of the systems' Arabic output was judged by native Arabic speakers, while the fluency of the systems' English output was judged by native English speakers.

The judgments of the Adequacy and Fluency of the English-to-Arabic translations and of the Adequacy of the Arabic-to-English translations were supplied by native Arabic speakers with a very advanced level of English, while the judgments of the Fluency of the Arabic-to-English translations were supplied by native English speakers. In the Arabic-to-English translations, despite the two different sets of evaluators, the results for both criteria were quite consistent for Google, Lilt and Yandex, although not for Bing M. Bing M.'s output was assessed inconsistently: the mean score for Adequacy, which was supplied by Arabic informants with a very advanced level of English, was significantly higher than the means score for Fluency, which was supplied by native English speakers. Regarding the automatic evaluation, in terms of Arabic-to-English translation, hLEPOR computed Google as the best, followed by the Lilt tool, while Bing NMT was placed third and Yandex was in last place. BLEU computed Bing NMT as the best, followed by Google, Lilt was placed third, while Yandex gave the poorest result. The highest value scored by hLEPOR was 0.6780, while the highest BLEU score was 0.5274. Regarding English-to-Arabic translation, hLEPOR calculated that Bing M. produced the best output, followed by Google, Yandex was placed third, while Lilt was in last place. In contrast, BLEU calculated that Lilt produced the best output, followed by Yandex, and then Bing M., while Google gave the poorest performance. The highest value scored by hLEPOR was 0.4412, while the highest BLEU score was 0.2685. This suggests that hLEPOR performed better than BLEU. However, the fact that only one reference was used in the rating of the English-to-Arabic translations may have had an impact on the calculations. It was clear that the average values of the English-to-Arabic translations were very low in comparison to the average scores of the Arabic-to-English translations. The scores for Arabic to English may have been higher as the corpus used four English references in contrast to the one reference used for English to Arabic. According to Papineni et al. (2002), additional reference translations increase the BLEU score.

The inter-rater agreement, which is calculated according to Fleiss's kappa measurement (i.e. how consistent the evaluators were in terms of their assessments), was used to assess the level of agreement between the evaluators. The findings of the experiment revealed a degree of inconsistency among the evaluations. Therefore. impossible to draw clear conclusions, but there is a tendency for lower agreement among a larger number of raters to be seen as best. We compare overall agreement in percentage in Table 7.11 below.

	Arabic-to-English translation			English-to-Arabic translations	
	Agreement (Fleiss's kappa) in Adequacy	Agreement (Fleiss's kappa) in Fluency		Agreement (Fleiss's kappa) in Adequacy	Agreement (Fleiss's kappa) in Fluency
Q1	-0.15183	0.596774	Q5	0.170158	0.097342
Q2	-0.03949	0.554795	Q6	-0.02036	-0.02064
Q3	0.04832	0.21875	Q7	0.055055	0.082569
Q4	0.066083	0.651163	Q8	0.118841	0.162178

Table 7.11: Inter-rater Agreements

In terms of Adequacy ratings, whether Arabic to English or vice versa, and Fluency ratings in English to Arabic translation, the inter-rater reliability showed poor and slight agreement. Regarding Fluency ratings in Arabic to English translation, the inter-rater reliability showed fair and moderate agreement. It appears that the inter-rater reliability using a larger number of raters was much lower than for using a small number of raters. A further possible explanation for the lower inter-rater agreement might be that the Arabic-English translation participants had some trouble assessing Adequacy and Fluency with MT systems' output. Although this could lead to questions concerning the reliability of human judgement, it is perhaps more easily explained by the evaluators' lack of familiarity with the Adequacy and Fluency rankings. As the demographic data shows (section 7.1.1), only 40% had previous experience of ranking translations using these criteria, while 60% had none at all.

A further noticeable outcome was the fact that the evaluators rated Google the best in Arabic-to-English translation in terms of Adequacy and in English-to-Arabic translation whichever the quality criteria, possibly reflecting the power of its NMT system (all four systems use the NMT approach). This result appears to confirm the findings of Al-Mahasees' study (2020). However, the interesting result is that Lilt led the other systems in terms of Fluency in Arabic-to-English translation. This may reflect the effect of the interactive and adaptive mechanism adopted by the Lilt tool.

7.2.2 Summary of the qualitative and quantitative data analysis

Drawing on the strengths of triangulation design, the findings of the quantitative analyses informed the development of the qualitative interviews. The participants' answers demonstrated a statistically reliable decrease or increase in Adequacy and Fluency on at least one of the measures, and the qualitative data gave an insight into the quality of the output of NMT systems in translating between Arabic and English. Data were analysed separately and then integrated to provide an overview of a graded scale of Adequacy and Fluency (quantitative results) and the translation quality of the NMT method (qualitative results). By choosing to employ a triangulation design, the researcher was able to substantiate the hypothesis that MT technology has difficulty in handling the challenge of translating Arabic content and to also describe the effects of this difficulty and the areas affected by it.

7.3 Conclusion

This chapter has presented the results of a comparison of the quality of four NMT systems' Arabic<>English translations using an Adequacy and Fluency rating, The fluency of the translations was judged by native speakers of each language pair, as well as by the automatic metrics BLEU and hLEPOR. The study used data extracted from the LDC corpus, which was then translated by Bing M., Google, Lilt and Yandex. The results showed that the NMT systems produced a good quality translation, although the adequacy of the translations from Arabic to English was rated slightly higher than those from English to Arabic. Producing fluent translations from English into Arabic, however, proved more difficult for these systems than vice versa, which suggests that translating from a morphologically rich language is automatically easier than translating into such a language. In the comparison of the four NMT systems, the human evaluators rated Google NMT as the most adequate in both translation directions and the most fluent in English-to-Arabic translation, and Lilt as the most fluent in Arabic-to-English translation. In the comparison performed by the automatic evaluation metrics, Google achieved the best hLEPOR score and Bing M. the best BLEU score in translating from Arabic into English, while Bing M. achieved the best hLEPOR score and Lilt achieved the best BLEU score in translating from English to Arabic.

CHAPTER EIGHT

CONCLUSION

8.0 Introduction

This chapter presents the overall conclusions drawn from this experimental study, which has sought to establish which is the best translation technology tool for Arabic<>English translation. Section 8.1 directly addresses each of the research questions outlined in Chapter One; section 8.2 offers recommendations for TM developers and designers of MT systems; section 8.3 highlights the contributions of this research to the ongoing evaluation of translation tools; and section 8.4 discusses its limitations and suggests the ways in which future research could extend the study.

8.1 Addressing the research questions

The answers to the research questions (RQ1-6) guiding this study test the validity of the research hypotheses detailed in Chapter One (section 1.3).

RQ1: To what useful extent can the matching metrics of TM systems retrieve segments that are semantically identical but different in structure?

The findings show that the matching metrics of current TM systems appear to prevent two highly similar segments, differing only in their word order, from being ranked as useful matches. As the experiment in Chapter Four, section A revealed, the systems' algorithms were unable to recognise a move operation: if segments included a move operation, such as the reordering phenomena in Arabic, their retrieval from the TM database provided very low matches, especially in the case of shorter segments. Further, move strings of different sizes were treated either as multiple-word units or as blocks. The systems' criterion of measurement seems to be computed by using only the surface forms of the segments – no linguistic knowledge is involved.

This inability to identify move operations was confirmed by the results of the experiment in Chapter Four, section B: the retrieval of segments which included a move operation always produced the lowest match score compared with the retrieval of a three-operation edit (addition, deletion or substitution). This is evidence that – with the exception of OmegaT – the TM matching metrics neither recognise the move operation nor treat it as a three-operation edit.

OmegaT, on the other hand, dealt with the move operation as if it were a deletion operation. This suggests that the different ways in which each TM system dealt with the reordering operation had different consequences. The outcome was that if segments contained a reordering operation (move operation), such as the reversal of the subject and the verb in the Arabic segments, their retrieval from the TM resulted in very low matches, especially for the shorter segments.

RQ1 was based on the assumption that the TM user generally establishes 70% as a match threshold. This means that if the match drops under this threshold, the user would have to translate the segment from scratch despite the presence in the TM database of a translation of a semantically identical segment. This confirms the hypothesis that, in such a scenario, the translator would be denied the maximum leverage of matches for lexically and semantically highly similar segments.

RQ2: Do the TM matching algorithms treat a combination of inflectional affixes (which may also be combined with a diacritic mark) linguistically or statistically?

Based on the findings of the experiment in Chapter Five, section A, it can be said that the TM systems' similarity algorithms treat a morphological combination as an intervention on the whole word, as a single-character change, or according to the position of the intervention. For example, DVX dealt with the inflectional affix as an intervention on the whole word; OmegaT and Trados Studio employed a specific mechanism for an individual edit operation (i.e. a single-character intervention) to the inflectional affix; and memoQ's measurement, which was derived from two different ranges, provided a low match range for segments that were based on the number of words and a high match range for segments that were based on the total number of characters. Meanwhile, Memsource's algorithm, which was based on the total number of segment characters, produced high percentages when inserting or removing a one-character prefix; however, when a one-character prefix was changed into a suffix or vice versa, the segment was assigned a lower match. Thus, the TM systems' algorithms, apart from some of Memsource's and memoQ's scores, penalised a combination of inflectional affixes heavily. This is further evidence that TM matching measurements use a purely statistical algorithm in their matching operations, and no linguistic information is involved.

RQ2 was based on the assumption that the high-scoring match would be presented near the top of the list of proposals.

RQ3: Is the absence of a Hamza marker weighted as a minor or major difference by the TM matching metrics?

The findings showed that Trados Studio treated the omission of a Hamza marker as a minor error, suggesting that the system has an effective mechanism for addressing the omission in Arabic. In contrast, the other four systems treated such an omission as a word or character intervention, and not as an orthographic error.

RQ3 was based on the assumption that the range of nearly exact matches would be the best as the two segments were identical in meaning and surface forms except for an omission of the Hamza marker. This nearly exact match would be granted a higher discount.

To conclude, the results of the TM retrieval in terms of the reordering operation and the morphological inflection reveal that the matching scores of TM systems appeared to be based purely on the string of surface forms; the matches fell steadily as the segment length became shorter. In terms of orthography, excepting Trados Studio, the other four systems employed a similar method. Trados Studio, however, dealt with the missing Hamza markers in a different way: it seemed to recognise the variant Hamza and treated the omission accordingly as a minor difference. Further, the TM matching metrics scored the segment retrieval that included a reordering operation in lower fuzzy bands than those provided by the three-operation edit. Furthermore, the matching metrics dealt with the segment retrieval containing a one-character inflection and the omission of the Hamza marker as either an intervention on a whole word or a single character change.

In terms of the comparison of the five TM systems, it can be said that the matching metrics of the five TM systems – apart from Trados Studio in relation to the character-marker omission – do not appear to have any linguistic basis. Trados Studio treated the omission of the marker as a minor error. In contrast, the matching metrics of the DVX, memoQ, OmegaT and also Trados Studio systems, in which the recall was based on the number of words in a segment, show that a one-character inflection was equated with a whole-word substitution. This was the same for the segment with a missing marker in DVX, memoQ and OmegaT, but not Trados Studio. This may be due to the fact that, in a European language, changing a single character often completely changes the meaning of the word. Regarding Memsource’s scores, the matching scores, which were inconsistent regardless of the type of intervention, appeared to rely on the number of characters. The system provided very high matches for the retrieval of segments that included a one-character prefix edit and segments with the omission of the Hamza marker (but only in the word-initial position). However, it seems Memsource assigned high scoring according to the position of the intervention, not according to a linguistic perspective. This may be due to the fact that, a prefix combination may cause less damage to the word form than a suffix combination.

In terms of usability, these results imply that highly repetitive texts are not enough on their own if the TM is to be put to good use; the segments also have to be in the same word order. The current TM matching metrics tend to be based on surface string and edit distance only, and failure to develop the matching measurement could result in a significantly lower usability of TM matches and higher costs. Bearing in mind the limited range of segment lengths tested (three-to-ten words), useful TM retrieval appears to be conditioned by the length of the segments – longer segments produce higher matches. However, longer segments are less likely to be found in a TM database.

RQ4: What degree of adequacy and fluency can different NMT systems achieve when translating between Arabic and English?

The results showed that NMT systems produced a good quality translation, which was more adequate than fluent in both directions: the systems' output largely expressed the meaning of the source text but committed some morphological errors. This may be due to the lack of training data resources in Arabic and its rich morphology. Also, the systems seemed to find producing Arabic translations of English texts more difficult than translating in the other direction. One reason for the decrease in the adequacy of English-to-Arabic translations was the non-translation of English proper nouns.

RQ5: Is there a difference between the scores of BLEU and hLEPOR metrics when translating between Arabic and English?

The findings show that the hLEPOR scores were closer to 1.0 (a perfect match) than BLEU's scores, whichever the translation direction. Further, hLEPOR scored the NMT systems' translations in English higher than their Arabic ones. Similarly, BLEU gave higher scores to the systems' output in English than in Arabic. Overall, hLEPOR performed better than BLEU.

RQ6: Based on the scores obtained from the evaluation methods, which one of the NMT systems provides better translation when translating between Arabic and English?

The findings show that Google NMT was rated by the human evaluators in this study as producing the most adequate and fluent translations in both directions regardless of the quality criteria. In contrast, the automatic metrics differently scored the MT systems, Bing NMT achieved the best BLEU score while Google achieved the best hLEPOR score in terms of Arabic-to-English translation. Bing NMT also achieved the best hLEPOR score regarding English-to-Arabic translation while Lilt achieved the best BLEU score.

Overall, both TM and MT systems provided useful results, according to the function of each system; however, both types of system appear to find dealing with Arabic word order a significant challenge (i.e. handling the reordering operation in the TM retrieval process and accomplishing syntactic re-arrangement, thus ensuring a natural syntactic structure). This seems to be especially the case when dealing with shorter segments of text.

8.2 Dissemination of key findings

To disseminate the findings of the study, the researcher intends to:

- circulate the results via conferences, print and online media, and, most importantly, journal publications;
- use https://www.proz.com/ (an online community and workplace used by professional translators and translation companies) to inform translators of the conclusions of this research;
- join and participate in specialised translation organisations, such as an International Forum for Arabic Translators.

8.3 Recommendations

This section focuses on the ways in which the outcomes of this research could make a difference to the performance of TM and MT systems when translating between Arabic and English. It suggests how the empirical findings could be applied to the future development of translation technology tools, and offers a set of recommendations based on these findings.

8.3.1 Recommendations for TM developers

OmegaT and Trados Studio showed some measure of success when retrieving a move operation by using a block string strategy, while Memsource and some memoQ scores performed well in the retrieval of inflected words. DVX performed the worst out of all the systems every time.

- Treat a string move as a block unit, the findings suggest that if the TM system, when retrieving segments that include a move operation, treats a multiple-word unit as a one-chunk move, it elicits better results than if the segment is treated as an assemblage of discrete words.
- Implement morphological processing: the findings substantiate the proposal of Gupta et al. 2016 a and b, Timonera and Mitkov 2015, and Gupta and Orasan 2014 that the implementation of morphological processing should be incorporated into the matching metrics of TM systems. Another interesting recent work comes from a research group associated with Ruslan Mitkov and looks at using paraphrase to enhance TM recall.[67]
- As a constant in the development of CAT tools, developers should pay attention to the diversity of their teams, and include researchers who speak non-European languages. Users of different languages pairs, like Arabic, could easily reuse their own stored translations for both commercial and academic research purposes.

8.3.2 Recommendations for TM translators

This research shows that, until the TM matching metrics improve, TM translators would do best to:

- set a match threshold lower than 70%: using a 65% match value, for example, would produce more recall in terms of a move operation, in particular in OmegaT;
- take into consideration the finding that MT systems seem to work better with short sentences than do TM systems;
- take care when the segment retrieval includes a move operation, as is often the case in languages with a flexible word order.

[67] http://rgcl.wlv.ac.uk/publications/ruslan-mitkov/

8.3.3 Recommendations for MT designers

The empirical results suggest that both types of translation technology systems still find the different syntactic rules of the two languages problematic (see Chapter Six). Hence, designers of MT systems are recommended to develop a mechanism that can handle the phenomenon of flexible word-order in Arabic with greater success.

Further, MT designers should seek to introduce a means of transliteration that can support the translation of proper nouns from English into Arabic. The researcher believes that this is needed to increase the quality of the MT output of Arabic and strengthen the translation quality of MT in general.

8.4 Contributions to the field

This comparative study of two translation technology systems contributes to the wide range of discussions concerning the evaluation of translation technology tools by revealing that Arabic language features still pose difficulties for these tools.

This study is the first to undertake an empirical investigation into a set of MT systems with the aim of determining which is the best translation technology tool for Arabic<>English translation, conducting experiments to test the retrieval performance of TM systems when the source is the morphologically rich language of Arabic. Moreover, the methodology, which is based on a black-box method and a test suite instrument, can be used as a guide in any future research that wishes to extend the investigation by experimenting with different language pairs and different linguistic features.

Secondly, it demonstrates in detail how lexically and semantically highly similar segments that differ only in their word order may be ranked as low matches, especially the short segments. The results also suggest that translation of short sentences is likely to be better using MT systems than using TM systems, since TM retrieval is processed in a very strict similarity measurement which results in low matching.

The study’s third contribution is the discovery that Trados Studio deals effectively with the retrieval of segments that only differ from the input in terms of the Hamza marker, resulting in the provision of matches in a nearly exact match band.

The fifth contribution of the study is that it is the first to evaluate the use of the Lilt tool as a TM system and hLEPOR as an automatic evaluation metric in translation between Arabic and

English. The findings show that both techniques would have a promising future in evaluating and translating from Arabic to English.

8.5 Limitations and future research

Throughout this research, the very best efforts have been made to maximise the validity of the study. However, it does possess some limitations, which point the way to further research.

8.5.1 The testing of TM systems

Although this comparative study evaluated pure TM systems, there are several components provided in a computer-assisted translation (CAT) setting. that bring TM and MT together. One strategy for integrating the two systems is to use MT Quality Estimation to rank MT results in the TM proposals list. Such a strategy, variants of which are currently used by web server systems such as Memsource and MateCat.[68] may represent a slightly different model based on the convergence of their technologies. It may provide a way of leveraging the advantages of one technology to improve the results of the other – either by using MT suggestion for segments with no or a low TM match or by using the quality human input from a TM system, for example when TM matches are found above a specific threshold, to retrain MT systems. More research is needed to investigate how such a hybrid tool might treat a move operation. On the other hand, there is a further web-based CAT tool called CATaLog,[69] which can be used as a normal CAT tool enhanced with a technique for post-editing TM segments or MT output. A recent version of CATalog has an information retrieval system which includes segment and word alignment. It may deliver a technique of handling the morphological combinations.

Another limitation is related to the language pair this research chose to study. The tests were limited to one language combination and a single direction: Arabic to English. Hence, it may be useful to replicate the methodology with different language pairs that have no strict word

[68] https://www.matecat.com/
[69] http://santanu.appling.uni-saarland.de/CATaLog/

order, such as Greek, Russian or Finnish or belong to morphologically rich languages such as German or Welsh, to see whether the results are replicated.

8.5.2 The testing of NMT systems

With regard to the data set, it should be noted that this study used sentences extracted from a corpus comprising news articles; a corpus from other, more specific domains may produce different results when training MT.

Owing to the limitations of time and availability of resources, especially human resources, the survey was distributed to a specific category of evaluators – graduate or advanced students of translation whose first language is Arabic. In future, the research could be opened out to include experimental studies targeting different categories of translators and including a greater variety of first languages. Another possible shortcoming (linked to the limitations of time and resources mentioned above) is the fact that the research analysed a small amount of data. As it was not possible to analyse large volumes of data, it focused on the quality of the output.

Further, the study was carried out using human evaluation and automatic evaluation; it may be useful to run correlation metrics such as Pearson correlation coefficient between human judgment and systems automated evaluations. The use of correlation criteria can measure the closeness between the manual judgments and the automatic metrics.

A final point is that although this comparative study used a variety of MT systems, there are now other aspiring MT systems, freely available to translators: DeepL,[70]. Thus far, this free translation system does not support Arabic, but once it does it would be interesting to test it to see whether it replicates the results obtained by this research.

[70] https://www.deepl.com/en/translator

Bibliographical references

Abdelaal, N. M., & Alazzawie, A. (2020). ‘Machine Translation: The Case of Arabic- English Translation of News Texts’. In: *Proceedings of Theory and Practice in Language Studies 10*(4), 408–418.

Abdelali, A (2004). ‘Localization in Modern Standard Arabic’. In: *Proceedings of Journal of the American Society for Information Science and Technology 55*(1), 23–28.

Al Mahasees, Z. (2020). *‘Diachronic Evaluation of Google Translate, Microsoft Translator and Sakhr in English-Arabic Translation’.* (Doctoral thesis), The University of Western Australia, Australia. Available online: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Diachronic+Evaluation+of+Google+Translate%2C+Microsoft+Translator+and+Sakhr+in+English-Arabic+Translation%E2%80%99&btnG=

Alanazi, M. S. (2019). *‘The Use of Computer-Assisted Translation Tools for Arabic Translation: User Evaluation, Issues, and Improvements’*. (Doctoral thesis), Kent State University, Kent, USA. Retrieved from ProQuest Dissertations Publishing, (27692255).

Ali, M. S., Alatawi, A., Alsahafi, B., & Noorwali, N. (2020) ‘QUES: A Quality Estimation System of Arabic to English Translation.’ In: *Preceding of (IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. *11*, No. 7, 245-251

Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H., & Haidar, M. (2013). ‘An Opinion Analysis Tool for Colloquial and Standard Arabic’. In: *Proceedings of the Fourth International Conference on Information and Communication Systems (ICICS)*, 23-25.

Allison, B., Guthrie, D., & Guthrie, L. (2006) ‘Another Look at the Data Sparsity Problem’. In: *Proceedings of the 9th International Conference on Text, Speech and Dialogue,* Springer, 327–334.

Almahairi, A., Cho, K., Habash, N., & Courville, A. (2016). ‘First Result on Arabic Neural Machine Translation’. https://arxiv.org/abs/1606.02680

Almahasees, Z. M. (2018). ‘Assessment of Google and Microsoft Bing Translation of Journalistic Texts’. In: *Proceedings of the International Journal of Languages, Literature and Linguistics 4* (3) :231–235.

Almahasees, Z., & Mustafa, Z. (2017). ‘Machine Translation Quality of Khalil Gibran’s The Prophet’. In: *Proceedings of the Arab World English Journal for Translation and Literary Studies 1*(4):151–159.

Almansor, E. H., & Al-Ani, A. (2018). ‘A Hybrid Neural Machine Translation Technique for Translating Low Resource Languages’. In: *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*, 347–356.

Alqudsi, A., Omar, N., & Shaker, K. (2014). ‘Arabic machine translation: a survey.’. In: *Proceedings of the Artificial Intelligence Review 42*(4),549–72.

Alqudsi, A., Omar, N., & Shaker, K. (2019). ‘A Hybrid Rules and Statistical Method for Arabic to English Machine Translation’. In: *Proceedings of the International Conference on Computer Applications & Information Security (ICCAIS)*,1-7.

Alrajeh, A. (2018). ‘*A recipe for Arabic-English neural machine translation*’. ArXiv Preprint ArXiv: https://arxiv.org/abs/1808.06116. Retrieved 27 November from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A+Recipe+for+Arabic-English+Neural+Machine+Translation&btnG=

Ameur, M. S. H., Meziane, F., & Guessoum, A. (2020). ‘Arabic Machine Translation: A survey of the latest trends and challenges. In: *Computer Science Review* 38:100305. https://doi.org/10.1016/j.cosrev.2020.100305

Ameur, M. S. H., Moulahoum, Y., & Guessoum, A. (2015). ‘Restoration of Arabic Diacritics Using a Multilevel Statistical Model’. In: *Proceedings of Computer Science and Its Applications, IFIP Advances in Information and Communication Technology*, 181–92. doi: https://doi.org/10.1007/978-3-319-19578-0_15.

Arthern, P. J. (1978). ‘Machine translation and computerized terminology systems: a translator’s viewpoint.’ In: *Preceding* of *Translating and the Computer, London* Vol. *14*, 77-108.

Ataman, D., & Federico, M. (2018). ‘An Evaluation of Two Vocabulary Reduction Methods for Neural Machine Translation’. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, Boston, 97-110.

Ataman, D., Aziz, W., & Birch, A. (2019). ‘A latent morphology model for open-vocabulary neural machine translation’. In: *Proceedings of International Conference on Learning Representations (ICLR2020).* Available online: https://openreview.net/forum?id=BJxSI1SKDH.

Azzano, D. (2011). *‘Placeable and localizable elements in translation memory systems’*. (Doctoral thesis), Ludwig Maximilian University of Munich, Germany. doi: 10.5282/edoc.1384. Available online: https://edoc.ub.uni-muenchen.de/13841/2/Azzano_Dino .

Babych, B. (2014). ‘Automated MT evaluation metrics and their limitations.’ In: *Tradumàtica,* (12), 464-470.

Badr, I., Zbib, R., & Glass, J. (2009). ‘Syntactic Phrase Reordering for English-to-Arabic Statistical Machine Translation’. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* Athens, Greece 86–93.

Badr, I., Zbib, R., & Glass, J. (2008). ‘Segmentation for English-to-Arabic Statistical Machine Translation’. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, Columbus, Ohio, 153-156.

Baldwin, T. (2009). ‘The Hare and the Tortoise: Speed and Accuracy in Translation Retrieval’. In: *Machine Translation 23*(4), 195–240. doi: 10.1007/s10590-009-9064-7.

Baldwin, T., & Tanaka, H. (2000). ‘The Effects of Word Order and Segmentation on Translation Retrieval Performance’. In: *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, Saarbrücken, Germany, 35–41.

Banerjee, S., & Lavie, A. (2005). 'METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.' In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization,* Ann Arbor, Michigan 65-72.

Baquero, A. S., & Mitkov, R. (2017). 'Translation Memory Systems Have a Long Way to Go'. In: *Proceedings of the Workshop Human Informed Translation and Interpreting Technology,* 44–51.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). 'Neural versus Phrase-Based Machine Translation Quality: A Case Study'. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Austin, Texas, 257–267.

Bisazza, A., & Federico, M. (2010). 'Chunk-Based Verb Reordering in VSO Sentences for Arabic-English Statistical Machine Translation'. In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. USA, 235–243.

Bloodgood, M., & Strauss, B. (2015). 'Translation memory retrieval methods. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* Gothenburg, Sweden, 202–210.

Bojar, O., Chatterjee, R., Federmann, C., et al. (2016). 'Findings of the 2016 Conference on Machine Translation'. In: *Proceedings of the First Conference on Machine Translation':* Association for Computational Linguistics. Berlin, Germany, 131–198.

Bowker, L. (2002). *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). 'The Mathematics of Statistical Machine Translation: Parameter Estimation'. In: *Computational Linguistics 19*(2):263–311.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, et al. (1990). 'A Statistical Approach to Machine Translation'. In: *Computational Linguistics 16*(2):79–85.

Bulté, B., & Tezcan, A. (2019). ‘Neural fuzzy repair: Integrating fuzzy matches into neural machine translation.’ In: *Preceding of the 57th Annual Meeting of the Association-for-Computational-Linguistics (ACL),* Florence, Italy,1800-1809.

Bulté, B., Vanallemeersch, T., & Vandeghinste, V (2018). ‘M3TRA: Integrating TM and MT for Professional Translators’. In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 69—78.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., & Schroeder, J. (2007). ‘(Meta-) Evaluation of Machine Translation.’ In: *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 136-158).

Callison-Burch, C., Koehn, P., Monz, C., & Zaidan, O. (2011). ‘Findings of the 2011 workshop on statistical machine translation.’ In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, 22-64.

Carroll, J. B. (1966). ‘An experiment in evaluating the quality of translations.’ In: *Mechanical Translation and Computational Linguistics, vol.9,* (3-4), 55-66.

Castilho, S., Moorkens, J., Gaspari, F., et al. (2017). ‘A Comparative Quality Evaluation of PBSMT and NMT Using Professional Translators’. In: *Proceedings of Machine Translation Summit XVI,* Nagoya, Japan.

Cer, D., Yang, Y., Kong, S. Y., Hua, et al. (2018). ‘Universal Sentence Encoder for English’. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations),* Brussels, Belgium,169–174.

Chan, S. W. (Ed.). (2015). *Routledge Encyclopedia of Translation Technology*. London and New York: Routledge.

Chatzitheodorou, K. (2015). ‘Improving Translation Memory Fuzzy Matching by Paraphrasing’. In: *Proceedings of the Workshop Natural Language Processing for Translation Memories*, Hissar, Bulgaria, 24–30.

Cho, K., Van Merriënboer, B., Gulcehre, C., et al. (2014). ‘Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation’. In: *Proceedings of the Empirical Methods Natural Lang. Process. (EMNLP),* Doha, Qatar, 1724–1734.

Cormode, G., & Muthukrishnan, S (2007). ‘The String Edit Distance Matching Problem with Moves’. In: *ACM Transactions on Algorithms 3*(1),1–19. doi: 10.1145/1186810.1186812.+.

Creswell, J. W. and Plano Clark, V., (2007). *‘Designing and Conducting Mixed Methods Research.*’ Thousand Oaks, CA: Sage Publications

Daelemans, W., & Hoste, V. (Eds.) (2009). *‘Evaluation of Translation Technology.’* Linguistica Antverpiensia 8. (NL-LeOCL) 840637314.

Daems, J., & Macken, L. (2019). ‘Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. In: *Machine Translation 33,* 117–134. https://doi.org/10.1007/s10590-019-09230-z.

Denscombe, M. (2014). *The Good Research Guide: For Small-Scale Social Research Projects.* Buckingham: Open University Press.

Ding, S., Renduchintala, A., & Duh, K. (2019). ‘A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation’. In: *Proceedings of Machine Translation Summit XVII Volume 1,* Dublin, Ireland, 204– 213.

Doddington, G. (2002). ‘Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.’ In: *Proceedings of the second international conference on Human Language Technology Research,* SanDiego*, CA* (pp. 138-145).

Dorr, B. J., Hovy, E. H., & Levin, L. S. (2004). ‘Machine Translation: Interlingual Methods’. *Encyclopedia of Language and Linguistics.* 2nd edition, Brown, Keith (ed.), Elsevier, Oxford, UK.

El Kholy, A., & Habash, N. (2012). ‘Orthographic and Morphological Processing for English--Arabic Statistical Machine Translation’. In: *Machine Translation 26*(1–2):25–45. doi: 10.1007/s10590-011-9110-0.

El Kholy, A., & Habash, N. (2010). 'Techniques for Arabic Morphological Detokenization and Orthographic Denormalization'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC),* Valletta, Malta.

El Marouani, M., Boudaa, T., & Enneya, N. (2017). 'AL-TERp: extended metric for machine translation evaluation of Arabic.' In: *Proceedings of the International Conference on Applications of Natural Language to Information System,* Springer, Cham, 156-161.

Elming, J. (2008). *Syntactic Reordering in Statistical Machine Translation*. (Doctoral thesis), Copenhagen Business School (CBS), Frederiksberg, Denmark. Available online: http://hdl.handle.net/10419/208702

Ezzeldin, A. M., & Shaheen, M. (2012). 'A survey of Arabic question answering: Challenges, tasks, approaches, tools, and future trends.' In: *Proceedings of the 13th International Arab Conference on Information Technology (ACIT 2012),* 1-8.

Farghaly, A. (2010). 'Arabic Machine Translation: A Developmental Perspective'. In: *Proceedings of the International Journal on Information and Communication Technologies*, 3(3), 3-10.

Farghaly, A., & Senellart, J. (2003). 'Intuitive Coding of the Arabic Lexicon'. In: *Proceedings of the Machine Translation Summit IX*, " Louisiana, USA.

Federico, M., Cattelan, A., & Trombetti, M. (2012). 'Measuring user productivity in machine translation enhanced computer assisted translation.' In: *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas* (AMTA). *https://scholar.google.com/scholar?hl=en&as_sdt=2005&sciodt=0%2C5&cites=6541617979700408077&scipsc=&q=Measuring+user+productivity+in+machine+translation+enhanced+computer+assisted+translation&btnG=*

Flanagan, K. (2015). 'Subsegment Recall in Translation Memory-Perceptions, Expectations and Reality'. In: *Journal of Specialised Translation (23)*, 64–88.

Görög, A. (2014). 'Quality Evaluation Today: The Dynamic Quality Framework'. In: *Proceedings of Translating and the Computer 36*, London: United Kingdom, 155–164.

Gow, F. (2003). *'Metrics for Evaluating* Translation Memory *Software'*. (Doctoral thesis), University of Ottawa., Canada. Available online: http://www.chandos.ca/Metrics_for_Evaluating_Translation_Memory_Software.pdf.

Grönroos, M., & Becks, A (2005). 'Bringing Intelligence to Translation Memory Technology.' In: *Proceedings of the International Conference Translating and the Computer 27.* London: ASLIB.

Gupta, R., & Orăsan, C. (2014). 'Incorporating Paraphrasing in Translation Memory Matching and Retrieval'. In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMachine Translation-2014*. Dubrovnik, Croatia: European Association for Machine Translation, 3–10.

Gupta, R., Orăsan, C., Liu, Q., & Mitkov, R. (2016.b). 'A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval Using Paraphrases'. In: *Proceedings of the 19th International Conference on Text, Speech, and Dialogue,* Springer International Publishing, 259-269.

Gupta, R., Orăsan, C., Zampieri, M., Vela, M., van Genabith, J., & Mitkov, R. (2016.a). 'Improving Translation Memory Matching and Retrieval Using Paraphrases'. In: *Machine Translation 30*(1):19–40. doi: 10.1007/s10590-016-9180-0.

Habash, N. (2007). 'Arabic Morphological Representations for Machine Translation'. In: *Proceedings of Arabic Computational Morphology: Knowledge-based and Empirical Methods*, 263–85.

Habash, N. (2010). '*Introduction to Arabic Natural Language Processing'*. Morgan & Claypool Publishers.

Habash, N., & Rambow, O. (2007). 'Morphophonemic and Orthographic Rules in a Multi-Dialectal Morphological Analyzer and Generator for Arabic Verbs'. In: *Proceedings of International Symposium on Computer and Arabic Language (ISCAL)*, Riyadh, Saudi Arabia.

Habash, N., & Sadat, F. (2006). 'Arabic Preprocessing Schemes for Statistical Machine Translation.' In: *Proceedings of the Human Language Technology Conference of the NAACL,* New York City, USA, 49–52.

Hadla, L. S., Hailat, T. M., & Al-Kabi, M. N. (2015). 'Comparative Study Between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study.' In: *Proceedings of the International Journal of Advanced Computer Science and Applications, 6* (11), 215-223.

Han, A. L. F., Wong, D. F., Chao, L. S., He, L., & Lu, Y. (2014). 'Unsupervised Quality Estimation Model for English to German Translation and Its Application in Extensive Supervised Evaluation'. In: *Proceedings of The Scientific World Journal. Issue: Recent Advances in Information Technology,* 1–12.

Han, A. L., Wong, D. F., & Chao, L. S. (2012). 'LEPOR: A robust evaluation metric for machine translation with augmented factors.' In: *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012),* Mumbai, India, 441–450.

Han, A. L., Wong, D. F., & Chao, L. S., He, L., Lu, Y., Xing J., & and Zeng, X. (2013). 'Language-independent Model for Machine Translation Evaluation with Reinforced Factors.' In: *Proceedings of the 14th International Conference of Machine Translation Summit, Nice, France*, 215–222.

Han, A. L.F. (2017). *'LEPOR: An Augmented Machine Translation Evaluation Metric'*. (Doctoral thesis), University of Macau, Macao. Available online: https://arxiv.org/abs/1703.08748

Hatem, A., & Omar, N. (2010). 'Syntactic Reordering for Arabic- English Phrase-Based Machine Translation'. In: *Proceedings of Database Theory and Application, Bio-Science and Bio-Technology, Communications in Computer and Information Science, Vol. 118,* Verlag Berlin Heidelberg, 198–206.

Hodász, G., & Pohl, G. (2005). 'MetaMorpho Translation Memory: A Linguistically Enriched TM'. In: *International Workshop: Modern Approaches in Translation Technologies*. Borovets, Bulgaria, 26–30.

Hu, X., Li, G., Xia, X., Lo, D., Lu, S., & Jin, Z. (2018). ‘Summarizing source code with transferred API knowledge.’ In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence. AAAI,* 2269–2275.

Hutchins, J. (1998). ‘The Origins of the Translator’s Workstation’. In: *Machine Translation 13*(4),287–307. doi: 10.1023/A:1008123410206.

Hutchins, J. (1999). The Development and Use of Machine Translation Systems and Computer-Based Translation Tools. In: *Proceedings of the International Conference on Machine Translation and Computer Language Information Processing.* Beijing, China, 1–16.

Hutchins, J. (2005). ‘Example-Based Machine Translation: A Review and Commentary’. In: *Machine Translation* 19(3):197–211. doi: 10.1007/s10590-006-9003-9.IBM Archives: 701 Translator (1954). Retrieved 27 November 2020 from: http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html.

Hutchins, W. J., & Somers, H. L. (1992). *An Introduction to Machine Translation*. volume 362. London: Academic Press London.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J (2017). ‘Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation’. In: *Proceedings of Transactions of the Association for Computational Linguistics, vol. 5,* 339–351.

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). ‘Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions’. *ArXiv Preprint ArXiv:1610.01108.*

Kalchbrenner, N., & Blunsom, P. (2013). ‘Recurrent Continuous Translation Models’. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP),*1700–1709.

Kay, M. (1980/1997). ‘The Proper Place of Men and Machines in Language Translation’. In: *Machine Translation 12*(1–2):3–23.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Koehn, P., & Monz, C. (2006). 'Manual and automatic evaluation of machine translation between European languages.' In: *Proceedings on the Workshop on Statistical Machine Translation, New York, New York, 102-121.*

Koehn, P., Och, F. J., & Marcu, D. (2003). 'Statistical Phrase-Based Translation'. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1,* NAACL '03, 48–54.

Lavie, A. (2010). 'Evaluating the output of machine translation systems.' MT Summit Tutorial, page 86.

Lee, Y. S.2004. 'Morphological Analysis for Statistical Machine Translation'. In: *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL),* Boston, Massachusetts, USA.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., ... & Arnold, D. (1996). 'TSNLP – Test Suites for Natural Language Processing'. In: *Proceedings of the 16th International Conference on Computational Linguistics,* Copenhagen, Denmark, 711–717.

Levenshtein, V. I. (1966). 'Binary Codes Capable of Correcting Deletions, Insertions, and Reversals'. In: *Proceedings of Soviet Physics-Doklady, vol. 10*(8),707–710.

Linguistic Data Consortium. (2005). ACE (automatic content extraction) English annotation guidelines for events. Retrieved 27 November 2020 from https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

Liu, X., Wong, D. F., Liu, Y., Chao, L. S., Xiao, T., & Zhu, J. (2019). 'Shared-Private Bilingual Word Embeddings for Neural Machine Translation'. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. 3613–3622.

Lo, C. K., Tumuluru, A. K., & Wu, D. (2012). ‘Fully automatic semantic MT evaluation’. In: *Proceedings of the Seventh Workshop on Statistical Machine Translation,* Montreal, Canada, 243–252.

Lo, C. K., Tumuluru, A. K., & Wu, D. (2012). ‘Fully automatic semantic MT evaluation.’ In: *Proceedings of the Seventh Workshop on Statistical Machine Translation,* Montreal, ´ Canada, *243-252.*

Lommel, A. (2018). ‘Metrics for translation quality assessment: a case for standardising error typologies.’ In: *Translation Quality Assessment*, Springer, Cham. 109-127.

Lommel, A., Görög, A., Melby, A., Uszkoreit, H., Burchardt, A., & Popović, M. (2015). ‘Harmonised metric. Project Report, QT21 project.’ Retrieved from https://www.taus.net/qt21-project

Ma, Y., He, Y., Way, A., & van Genabith, J. (2011). ‘Consistent translation using discriminative learning-a translation memory-inspired approach.’ In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA*. 1239-1248.

Macklovitch, E., & Russell, G. (2000). ‘What’s been forgotten in translation memory’. Envisioning Machine Translation in the Information Future: In: *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas.* Cuernavaca, Mexico, 137–146

Marzouk, S., & Hansen-Schirra, S. (2019). ‘Evaluation of the Impact of Controlled Language on Neural Machine Translation Compared to Other Machine Translation Architectures’. In: *Machine Translation 33*(1):179–203. doi: 10.1007/s10590-019-09233-w.

Massardo, I., van der Meer, J., & Khalilov, M. (2016). TAUS translation technology landscape report. De Rijp: TAUS. Retrieved 27 November 2020 from (https://www.taus.net/think-tank/reports/translate-reports/taus-translation-technology-landscape-report-2016#download-purchase).

Maučec, M. S., & Donaj, G. (2019) 'Machine Translation and the Evaluation of Its Quality'. In: *Trends in Computational Intelligence*. doi: 10.5772/intechopen.89063. Retrieved 20 November 2020 from https://www.intechopen.com/books/recent-trends-in-computational-intelligence/machine-translation-and-the-evaluation-of-its-quality

Melby, A. K. (1981). 'A bilingual concordance system and its use in linguistic studies.' In

Mesa-Lao, B. (2020). Explicitation and Translation Editing Environments. Retrieved 20 November from https://scholar.google.com/scholar?hl=en&as_sdt=2005&sciodt=0%2C5&cites=12146948972852493232&scipsc=&q=Explicitation+and+Translation+Editing+Environments.&btnG=

Miculicich, L., Ram, D., Pappas, N., & Henderson, J. (2018). 'Document-Level Neural Machine Translation with Hierarchical Attention Networks'. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Brussels, Belgium, 2947–2954.

Mitkov, R. (2005). 'New Generation Translation Memory Systems'. In: *Panel Discussion at the 27th International ASLIB Conference 'Translating and the Computer,* London. United Kingdom.

Moorkens, J. (2012). *'Measuring Consistency in Translation Memories: A Mixed-Methods Case Study'.* (Doctoral thesis), Dublin City University, Ireland. Available online: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=%E2%80%98Measuring+Consistency+in+Translation+Memories%3A+A+Mixed-Methods+Case+Study%E2%80%99&btnG=

Moorkens, J., & O'BRIEN, S. (2017). 'Assessing User Interface Needs of Post-editors of Machine Translation'. In: *Proceedings of Human Issues in Translation Technology*, Routledge, 127–148.

Muravev, Y. (2020). 'Machine translation and legal tech in legal translation training.' In: *Proceedings of the International Scientific Conference-Digital Transformation on Manufacturing, Infrastructure and Service, 1-7.*

Muthukrishnan, S. M., & Șahinalp, S. C. (2002). ‘Simple and Practical Sequence Nearest Neighbors with Block Operations’. In: *Proceedings of the Symposium on Combinatorial Pattern Matching (2002), Lecture Notes in Computer Science vol. 2373,* Spring Verlag, 262–278.

Nagao, M. (1984). ‘A Framework of a Mechanical Translation between Japanese and English by Analogy Principle’. In: *Proceedings of the International NATO Symposium on Artificial and Human Intelligence,* North-Holland, Amsterdam, 173-180.

Neme, A (2011). ‘A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers’. In: *Proceedings of the International Workshop on Lexical Resources (WoLeR*), Slovenia, 78-85.

Nirenburg, S., & Wilks, Y. (2000). ‘Machine Translation’. In: *Advances in Computers, 52,* 159-188.

Nishimura, Y., Sudoh, K., Neubig, G., & Nakamura, S. (2019). ‘Multi-Source Neural Machine Translation with Missing Data’. In: *Proceedings of the Second Workshop on Neural Machine Translation and Generation (WNMachine Translation),* Melbourne, Australia, 92–99.

O’Brien, S. (2007). ‘Eye-Tracking and Translation Memory Matches’. In: *Perspectives: Studies in Translatology 14*(3):185–205.

Olohan, M. (2021). ‘Post-editing: A Genealogical Perspective on Translation Practice.’ In: *M. Bisiada (Ed.), Empirical Studies in Translation and Discourse*, Berlin.1-25. Language Science Press. https://doi.org/10.5281/zenodo.4450077

Ortega, J. E., Forcada, M. L., & Sanchez-Martinez, F. (2020). ‘Fuzzy-match repair guided by quality estimation.’ In: *IEEE Transactions on Pattern Analysis and Machine Intelligence. https://scholar.google.com/scholar?hl=en&as_sdt=2005&sciodt=0%2C5&cites=9501521549384792421&scipsc=&q=Fuzzy-match+repair+guided+by+quality+estimation&btnG=*

Ortega, J. E., Sánchez-Martınez, F., & Forcada, M. L. (2016). ‘Fuzzy-match repair using black-box machine translation systems: what can be expected.’ In: *Proceedings of AMTA*, Vol. 1, 27-39.

Ortega, J., Sánchez-Martínez, F., Turchi, M., & Negri, M. (2019). 'Improving translations by combining fuzzy-match repair with automatic post-editing.' In: *Proceedings of the Machine Translation Summit XVII, Dublin, Ireland,* 256–266.

Oudah, M., Almahairi, A., & Habash, N. (2019). 'The Impact of Preprocessing on Arabic-English Statistical and Neural Machine Translation'. In: *Proceedings of Machine Translation Summit XVII Volume*, Dublin, Ireland, 214–221.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). 'BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 311–318.

Paul, M., Federico, M., & Stüker, S. (2010). 'Overview of the IWSLT 2010 evaluation campaign.' In: International Workshop on Spoken Language Translation (IWSLT) 2010.

Pekar, V., & Mitkov, R. (2007). 'New generation translation memory: content-sensivite matching. In: Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters.

Phillips, A. B., Cavalli-Sforza, V., & Brown, R. D. (2007). 'Improving Example Based Machine Translation Through Morphological Generalization and Adaptation'. In: *Proceedings of the 9th Machine Translation Summit (Machine Translation Summit IX),* Copenhagen, Denmark, 369-375.

Planas, E. (2005). 'SIMILIS Second-generation Translation Memory software.' In, *Proceedings of the 27th International Conference on Translating and the Computer (TC27),* London, United Kingdom.

Planas, E., & Furuse, O. (1999). 'Formalizing Translation Memories'. In: *Proceedings of Machine Translation Summit VII,* Singapore,331-339.

Quah, C.K. (2006). '*Translation and Technology'*, Palgrave Textbooks in Translation and Interpretation, Palgrave MacMillan.

Quaranta, B. (2011). ‘Arabic and Computer-Aided Translation: An Integrated Approach’. *Translating and the Computer*, 33, 17-18. Retrieved November 2020 from http://www.machinetranslation-archive.info/Aslib-2011-Quaranta.pdf

Ranasinghe, T., Orasan, C., & Mitkov, R. (2020). ‘Intelligent Translation Memory Matching and Retrieval with Sentence Encoders’. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation,* Lisboa, Portugal, 175–184.

Reinke, U. (2013). ‘State of the Art in Translation Memory Technology’. In: *Translation: Computation, Corpora*, Cognition, 3(1), 27–48.

Sado Al-Jarf, R. (2007). ‘SVO Word Order Errors in English-Arabic Translation’. In: *Translators' Journal 52*(2), 299–308.

Sajjad, H., Dalvi, F., Durrani, N., Abdelali, A., Belinkov, Y., & Vogel, S. (2017). ‘Challenging Language-Dependent Segmentation for Arabic: An Application to Machine Translation and Part-of-Speech Tagging’. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* Vancouver, Canada, 601– 607.

Saldanha, G., & O'Brien, S. (2014). *Research Methodologies in Translation Studies*. London and New-York: Routledge.

Salem, Y., Hensman, A., & Nolan, B. (2008). ‘Implementing Arabic to English Machine Translation using the role and reference grammar linguistic model’. In: *Proceedings of the Eighth Annual International Conference on Information Technology and Telecommunication (ITT 2008),* Galway, Ireland, 103-110.

Santy, S., Dandapat, S., Choudhury, M., & Bali, K. (2019). ‘INMT: Interactive Neural Machine Translation Prediction’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* Hong Kong, China, 103–108.

Sennrich, R., Haddow, B., & Birch, A. (2015). ‘Improving Neural Machine Translation Models with Monolingual Data’. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016),* Berlin, Germany, 86–96.

Shamsan, M. A. H. A., & Attayib, A. M. (2015). ‘Inflectional Morphology in Arabic and English: A Contrastive Study’. In: *Proceedings of International Journal of English Linguistics, 5*(2), 139-150.

Shapira, D., & Storer, J. A. (2002). ‘Edit Distance with Move Operations’. In: *Proceedings of the 13th Annual Symposium on Combinatorial Pattern Matching,* Fukuoka, Japan, 85–98.

Shapiro, P., & Duh, K. (2018). ‘Morphological Word Embeddings for Arabic Neural Machine Translation in Low-Resource Settings’. In: *Proceedings of the Second Workshop on Subword/Character Level Models,* New Orleans, 1–11.

Simard, M., & Fujita, A. (2012). ‘A Poor Man’s Translation Memory Using Machine Translation Evaluation Metrics’. In: *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (A Machine Translation A),* San Diego, California, USA.

Simard, M., & Isabelle, P. (2009). ‘Phrase-based machine translation in a computer-assisted translation environment.’ In: *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII),* Ottawa, Ontario, Canada. 120-127.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). ‘A study of translation edit rate with targeted human annotation.’ In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* Cambridge, Massachusetts. 223-231.

Somers, H. (2003). *Computers and Translation: A Translator’s Guide*. Amsterdam: John Benjamins Publishing.

Somers, H., & Fernández Díaz, M. G. (2004). ‘Translation Memory vs. Example-based MT–What’s the difference?’ In: *Proceedings of International Journal of Translation, 16*(2):5–33.

Šoštarić, M. (2018). ‘Advanced Fuzzy Matching in the Translation of EU Texts’. In: *Journal of Translation Studies and Terminology (5),* 26-71.

Soudi, A., Farghaly, A., Neumann, G., & Zbib, R. (Eds.) (2012). *Challenges for Arabic Machine Translation*. John Benjamins Publishing.

Stent, A., Marge, M., & Singhai, M. (2005). 'Evaluating evaluation methods for generation in the presence of variation.' In: *International Conference on Intelligent Text Processing and Computational Linguistics,* Springer, Berlin, Heidelberg, 341-351.

Su, K. Y., Wu, M. W., & Chang, J. S. (1992). 'A new quantitative quality measure for machine translation systems.' In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics,* Nantes, France.

Sutskever, I., Vinyals, O., & Le, Q. V (2014). 'Sequence to Sequence Learning with Neural Networks'. In: *Advances in Neural Information Processing Systems,* 2096–2104.

Tan, X., Chen, J., He, D., Xia, Y., Qin, T., & Liu, T. Y. (2019). 'Multilingual Neural Machine Translation with Language Clustering'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19),* Hong Kong, China, 962–972.

Tezcan, A., Bulté, B., & Vanroy, B. (2021). 'Towards a Better Integration of Fuzzy Matches in Neural Machine Translation through Data Augmentation.' In: *Informatics* Multidisciplinary Digital Publishing Institute. (Vol. 8, No. 1, p. 7).

Thawabteh, M. A. (2013). 'The Intricacies of Translation Memory Tools: With Particular Reference to Arabic-English Translation'. In: *Localisation Focus: The International Journal of Localisation 12*, 79–90.

Thierry P. (2017). *Machine Translation*. The MIT Press, Cambridge, Massachusetts, London, England.

Toral, A., & Sánchez-Cartagena, V. M. (2017). 'A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1,* Valencia, Spain, 1063–1073.

Tu, Z., Liu, Y., Shi, S., & Zhang, T. (2018). 'Learning to Remember Translation History with a Continuous Cache'. In: *Transactions of the Association for Computational Linguistics 6,* 407–420.

Turian, J. P., Shea, L., & Melamed, I. D. (2006). ‘Evaluation of machine translation and its evaluation.’ In: *Proceedings of the MT Summit IX, New Orleans, pp 386–393*

Utiyama, M., Neubig, G., Onishi, T., & Sumita, E. (2011). ‘Searching Translation Memories for Paraphrases’. In: *Machine Translation Summit, volume 13*, Xiamen, China, 325–331.

Vanallemeersch, T., & Vandeghinste, V. (2015). ‘Assessing Linguistically Aware Fuzzy Matching in Translation Memories’. In, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey,153–160.

Wagner, R. A., & Fischer, M. J. (1974). ‘The String-to-String Correction Problem’. In: *Association for Computing Machinery (ACM) 21*, 168-173.

Wang, O. (2014). *‘The Application of Explicit Semantic Analysis in Translation Memory Systems’.* (Doctoral thesis), Imperial College London, United Kingdom. doi: https://doi.org/10.25560/45427

Way, A. (2020). ‘Machine translation: where are we at today?’ In: *Erik Angelone, Maureen Ehrensberger Dow, and Gary Massey, editors,* The Bloomsbury Companion to Language Industry Studies, 311-332.

White, J. S. (1995). ‘Approaches to black box MT evaluation.’ In: *Proceedings of Machine Translation Summit V* (page. 10).

White, J. S. (2003). ‘How to evaluate machine translation.’ In: *H. Somers. (Ed.) Computers and Translation: a translator’s guide. Ed. J.* Benjamins B.V., Amsterdam, Philadelphia, 211-244.

White, J. S., O’Connell, T. A., & O’Mara, F. E. (1994). ‘The ARPA MT evaluation methodologies: evolution, lessons, and future approaches.’ In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas,* Columbia, Maryland,193–205

Whyman, E. K., & Somers, H. L. (1999). ‘Evaluation Metrics for a Translation Memory System. In: *Software-Practice and Experience, 29*(14):1265–84.

Williams, J., & Chesterman, A. (2014). *The Map: A Beginner's Guide to Doing Research in Translation Studies*. Manchester: Routledge.

Wolff, F., Pretorius, L., & Buitelaar, P. (2014). 'Missed Opportunities in Translation Memory Matching'. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC),* Reykjavik, Iceland, 4401-4406.

Wolff, F., Pretorius, L., Dugast, L., & Buitelaar, P. (2016) Methodological pitfalls in automated Translation Memory evaluation. In: *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016),* Portorož, LREC, 21-28.

Wong, T. M., & Kit, C. (2008). Word Choice and Word Position for Automatic Machine Translation Evaluation. In: *A Machine TranslationA 2008 Workshop: Metrics of the Association for Machine Translation in the Americas,* Waikiki, Hawai'i.

Wu, Q. (2017). 'A Brief Overview of Attention Mechanism. Retrieved 27 November 2020 from (https://medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129).

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'. In: *ArXiv:1609.08144 [Cs]*.

Xu, J., Crego, J. M., & Senellart, J. (2020). 'Boosting neural machine translation with similar translations.' In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1580-1590.

Xuan, H. W., Li, W., & Tang, G. Y. (2012). 'An Advanced Review of Hybrid Machine Translation (HMT)'. In: *Procedia Engineering 29*, 3017–3022. doi: 10.1016/j.proeng.

Zollmann, A., Venugopal, A., & Vogel, S. (2006). 'Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation'. In: *Proceedings of the Human Language Technology Conference of the NAACL,* New York City, USA, 201–204.

Appendices

Appendix One (1): Researcher’s TM (RTM)

Arabic source	English target
شجب الرئيس الإسلاموفبيا.	President denounced Alaslamovbaa.
ترك المهرج المنصة.	The clown left the platform.
قرأ الولد كتابه.	The boy read his book.
يرتبَ الفتى المكان.	The boy arranges the bed.
يلعبَ الفتى بالكرة.	The boy plays with the ball.
يخرج الأب مسرورا.	Father comes out happy.
يضغط الزر الأخضر.	Press the green light.
تبدأ امرأة بالتساؤل.	A woman begins to wonder.
خرج السيدة مسرعة.	The lady went out speeding.
شرق الشمس مبكرا.	The sun rose early.
كتبت الطفلة القصة.	The girl wrote the story
تفرح الأسرة بالمولود.	The family is happy with the baby.
تهبط الطائرة بسلام.	The plane lands safely.
دعم الفريق المدرب الحالي.	The team gave support to the current coach.
حمل ساعي البريد الطرود.	The postman carried the post package.
جمع الفلاح الثمار الناضجة.	The farmer collected the ripe fruits.
يبحثَ الناس عن أيديولوجيته.	People look for his ideology.

يفقدَ الرئيس شعبية كبيرة.	The president loses great popularity.
يحدثَ باحث تويتر بأفكاره.	A researcher updates using his own thoughts.
يسبح البحار في البحر.	The seaman swims in the sea.
يمشي جحا وراء الحمار.	Joha walks behind the donkey.
يقرأ الطالب في كتابه.	The boy reads his book.
تظمت الجمعية زواج جماعيا.	Assembly organised marriage collectively.
رسمت الفنانة لوحة جميلة.	The artist painted a beautiful painting.
درست شركة فرنسية الخطة.	A French company was preparing a plan.
تطبخ الأم وجبة الغذاء.	The mother cooks the lunch.
تزين الفتاة الطبق بالسلاطة.	The girl decorates this dish with some salad.
تأخذ الموظفة إجازة طويلة.	The employee takes a long vacation.
ترسل الزوجة لزوجها رسالة.	Wife sends a letter to his husband.
شرب الطفل الحليب الطازج صباحا.	The child drank fresh milk in the morning.
نظر الوالد إلى ولده مفتخرا.	The father looked at his son proudly.
فحص طبيب العيون عين أمي.	The option examined my mother’s eye.
يحسبَ سائق التاكسي الأجرة بدقة.	The taxi driver calculates the fare accurately
يذهبَ الرجل إلى عمله بالحافلة.	The man goes to his work by bus.
عزفَ المسيقار ألحان وطنية رائعة.	The musician played wonderful patriotic tunes.
يسكن الطالب مع عائلة إنجليزية.	The student lives with an English family.
جلست الجدة على الكرسي الجميل.	The grandmother has sit on a beautiful chair.
حضرت المديرة مع نائبها الاجتماع.	The manger attended the meeting with her deputy.

سمحت المحكمة الابتدائية بدخول المتضررين.	The court allowed the effect people to entre.
شملت الجولة المشروع المتوقف بالقرية.	The tour included the stalled project in the village.
تدفع مستأجرة البيت الأجرة كاملة.	The lady pays the rental in full.
فتحت النافدة بقوة فوقع الستار.	The window is opened forcefully; then the curtain fell.
اكل الرجل مع زوجته في المطعم.	The husband ate with his wife at the restaurant.
صرف المدير مكافئة لكل الموظفين بسخاء .	The manger granted all employees annually.
خلع طبيب الأسنان سن المريض ببراعة.	The dentist took off the patient’s tooth brilliantly.
يسحبَ الرئيس الوفد المشارك من المؤتمر.	The president takes out the delegation from the conference.
يرفضَ زعيم المتمردين مقابلة مبعوث الوساطة.	The rebellion leader refuses to meet the mediation envoy.
تهدد المطر الغزير والسيول سكان القرية.	Heavy rain and floods threaten the villagers.
يقرأ الشيخ المسن الصحف يوميا للمتعة.	The president takes out the delegation from the conference.
حدثت الحرب بين الدولتين لأسباب متعددة.	The war happened between the two states for multiple reasons.
ظهرت روح التطوع بين الثقافات المختلفة.	The spirit of volunteerism appeared among cultures
تعمل الوزارة على تأمين الدعم المالي.	The Ministry is working to secure financial support.
تقسم الأم الكعكة إلى ست قطع.	Mother slices the cake into six pieces.
تشرح المعلمة الفقرة شرحا تفصيليا دقيقا.	The teacher explains the section clearly and accurately.
تقدم الصحافة تقريرا عن العصابة الإجرامية.	The press releases reports on the criminal gang.

حضر الوفد الوطني المشارك أعمال القمة المصغرة.	The participating national delegation attended the mini-summit.
فحص دكتور متخصص نبضات القلب والتنفس والدماغ.	A specialist doctor examined the heartbeat, breathing and brain.
هدد انقطاع التيار الكهربائي حياة المرضى بالقرية.	Power outages threatened patients in the village.
يخصمَ الرئيس التنفيذي للشركة المزاية عن المتغيبين.	Chief Executive Officer of company deducts awards for absentees.
يسألَ الأستاذ سؤال ويطلب أجابة مختصرة ودقيقة.	The professor asks a question and needs for a brief and accurate answer.
يخلطَ الدهان اللون الأحمر مع الاصفر بحرفية.	The paint man mixes the colours of red with yellow professionally.
خصم الرئيس التنفيذي للشركة المزاية عن المتغيبين.	Chief executive Officer of company deducts awards for absentees.
سأل الأستاذ سؤال ويطلب أجابة مختصرة ودقيقة.	The professor asks a question and needs for a brief and accurate answer.
خلط الدهان اللون الأحمر مع الاصفر بحرفية.	The paint man mixes the colours of red with yellow professionally.
حجز ت المسافرة حجرة فردية بالفندق بوسط المدينة.	The traveller booked a single room in the city centre.
حذرت الحكومة الاهالي بعدم الاقتراب من المنحدر.	The government warned for the people not to approach the slope.
غيرت وسائل التواصل الحديثة من حياة الناس.	Modern means of communication have changed people’s lives
تصدر المحكمة حكما نهأئيا بالسجن مدا الحياة.	The court issues a final ruling a life sentence.

تجدد الاسرة تاكيدها للمحامي بعدم سحب القضية.	The family reiterates to the lawyer not to withdraw the case.
تسبب العملة المزورة انهيار في الاقتصاد الوطني.	The counterfeit currency leads to a collapse in the national economy
أسعار النفط مشتعلة.	Oil prices are on fire.
إطلاقها للصواريخ تهديد.	Her firing of rockets is a threat.
أليس هناك مصعد؟	Isn't there an elevator?
بدأت غالبية بالمغادرة.	The majority started leaving.
الآراء و التعليقات.	Views and comments.
الأغنام تنتج الصوف.	Sheep produce wool.
المبتدأ و الخبر.	Subject and predicate
الكلأ تعني العشب.	Herbal means grass.
التقيؤ يعني الاستفراغ.	The vomiting means emptiness.
شواطئ بحيرة طبريا.	The shores of Lake Tiberia.
أكبر من أخيه بسنة .	A year older than his brother.
أغلبية الثلثين مطلوبة للفوز.	A two-thirds majority is required to win.
إلا أن الأمر مستبعد.	Although, it is unlikely.
تأمر الشريعة الإسلامية بالذبح.	Shariah is ordered to slaughter.
الأدب الليبي يناقش التراث.	Libyan literature is discussing heritage.
الإنفاق الحكومي شحيح للغاية.	Government spending is very scarce.
الخطأ يُعالج بتكرار المحاولة.	The wrong action is treated by trying again.

سيصدأ الحديد لتعرضه للهواء.	The iron is rusted when exposed to air.
أتجرؤ أن تتهمني بالغدر.	you dare to accuse me of treachery.
الموانئ البحرية ثروة طبيعية.	Ports Maritime are natural resources.
أصدقاء المسجد يؤتون لزيارة الشيخ.	His friends of the mosque are coming to visit the Sheikh.
أسامح نفسي على الماضي ولكن.	I forgive myself for my past but.
أصداء الملاعب برنامج رياضي رائع.	Echoes stadiums wonderful sports program.
تآكلت أسنانه بسب أكل الحلوة.	His teeth eroded because of eating sweets.
تأخرت الأستاذ عن الفصل لدقائق.	The tutor was late for minutes.
يترأس الملك اجتماعا وزاريا مهما.	The King chairs a significant ministerial meeting.
الظمأ يذهب عند إفطار الصائم.	Thirst goes when breaking the fast.
يتوضأ المسلم لأداء الصلوات الخمس.	A Muslim perform ablution for performing the five daily prayers.
يختبئ خلف الصخور و الشقوق.	It is hiding behind rocks and crevices.
التجشؤ طرد الهواء من المعدة.	Belching expels the air from the stomach.
أطلقت المحطة الفضائية قمرا جديداً للفضاء.	The space station launched a new space satellite.
أعترف المتهم بدخوله البيت ليلاً للسرقة.	The accused man admitted that he entered the house at night to steal.
أعراض أعصاب المعدة الإرهاق والإمساك.	Symptoms of stomach nerves are fatigue and constipation.
متألق دائما يأبني سأمنحك جائزة رائعة.	Always, brilliant my son, I will give you a great prize.

تأثير الأكياس البلاستكية سيئ على البيئة.	The effect of plastic bags is bad on the environment.
الفأر يعيش بالقرب من مصادر الطعام.	The mouse lives near the food sources.
يكافأ المتميزين لحصولهم على المراتب الأولى.	The clever students are rewarded for having first ranks.
يتبرأ الأب من أولاده التاركين للصلاة.	Father acquitted one of his children who do not pray.
يفاجئ الإمام المصلين بصوت مؤثر جدا.	The Imam surprises the worshippers with a very impressive voice.
تباطؤ ضربات القلب يؤدي إلى التعب.	The slowed heart strikes causes to fatigue.
أنفقت الجمعية الخيرية أموالا ضخمة على الإيتام.	The charity has spent huge money on orphans.
أبداء المشورة قبل اتخاذ القرار شيء جيد.	Getting advice before making a decision it is a nice thing.
أغاني الأطفال لا ينصح بها قبل الخامسة.	The kids' songs are not recommended before the fifth.
سأفعل ذلك لاحقا بعد نهاية الجزء المتبقي.	I will do it later after the end of the remaining part.
بدأتم قرأت القرآن منذ العمر ست سنوات.	You started reading the Koran six years ago.
امرأة مكافحة تريد تربية أبنائها الأيتام بالحلال.	A striving woman wants to raise her orphan children.
يبتدأ بخطوة ثم تتبعها خطوات ليحقق مبتغاه.	It starts a step, and then it is followed by steps to make its objective.
يمتلأ المكان بالزوار على أخره.	The visitors filled the place completely.
يلتجئ إلى افتعال وصناعة الأزمات الاقتصادية.	Resort to fabricating and making economic crises.

Appendix Two (2): (Survey)

Title: Evaluating of Output Quality of Four MT Systems between Arabic and English

Introductory information

About the survey

This page provides you with important information about this research. Before beginning the survey, please read the information on this page and then tick the box at the end of the page to confirm your participation and begin the survey.

About this research

Khaled Ben Milad, a PhD researcher at the University of Swansea, UK, is conducting this research. This survey is a part of the PhD work which aims to rank the output quality of FOUR MT (MT) systems [GOOGLE TRANSLATE, BING MICROSOFT TRANSLATOR, YANDEX TRANSLATE, and LILT TOOL] in terms of Arabic<>English translation according to TAUS quality evaluation criteria: Adequacy and Fluency. For more information regarding the TAUS quality evaluation criteria, please click on ' i ' at the bottom-right below. Instructions for applying the criteria are explained below at the top of pages 3 & 4. There will be 4 translation samples for the Arabic to English direction and the same number for English to Arabic. The demographic information on page 2 is being obtained to better understand the users of the MT systems only.

Supervision team

Supervisors. Andrew Rothwell (a.j.rothwell@swan.ac.uk) & Maria Fernandez Parra (m.a.fernandezparra@swansea.ac.uk)

School Research Ethics Committee: COAHresearchethics@swansea.ac.uk

If you have any questions about this research that have not been answered on this page, please contact the researcher on the email: 882922@swansea.ac.uk

Who can take part?

The survey aims to collect data from professional translators - anybody who may be either a freelance translator, teacher of translation or advanced student (MA or PhD level of translation).

Participants' information.

Participation in this survey is completely voluntary. Participants are free to opt out of the research at any time without prejudice by contacting the researcher and asking to withdraw.

All data supplied will be stored securely, anonymously and confidentially; the data are expected to be kept for a maximum of one year, after which they will be deleted.

By ticking the choice (YES) below, you confirm that:

(a) You have read this information above and understood the purpose of this study

(b) You have read this information above and you are one of the categories of participant that the survey aims to collect data from.

(c) You understand that all data are anonymous and that there will not be any connection between the demographic information provided and the main data.

(d) You understand that there are no known risks or hazards associated with participating in this study.

(e) By submitting this questionnaire, you agree that your answers, which you have given voluntarily, can be used anonymously for research purposes only.

*I confirm that I consent to participate in this survey:

Yes ……….   No ……..

Demographic data

Demographic data will not be used in any way that can be linked to the participants; the data will be used to provide background about a participant. Please read the sub-questions below, then choose 'Yes or No' answer.

*Demographic Questions

Which category of participant are you? [a freelance translator, teacher of translation or student]

Do you use MT systems to translate (in either direction) between Arabic and English?

Answer

Have you previously been involved in evaluating the output quality of MT in terms of adequacy and fluency?

Answer

Do you usually edit the MT systems' output?

Answer

Is the participant a professional translator?

Answer

Arabic to English translation

Please read the source text (ST) and the output of FOUR MT systems, then rank Adequacy and Fluency on a scale of 1 to 4 for each, according to the following definitions and criteria:

ADEQUACY: 'How much of the meaning expressed in the source fragment appears in the translation fragment?'

1= None of it.

2= Little of it.

3= Most of it.

4= All of it.

FLUENCY: Is the target text well-formed grammatically "so that it contains correct spellings, adheres to common

use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker"?

1= Incomprehensible.

2= Dis-fluent.

3= Good.

4= Flawless.

ST

وأضاف اعتقد ان العلاقة وصلت الى مستوى لم نشهده من حيث الصدق والصراحة مع مسئولين

.كوريين ديمقراطيين اجتمعنا معهم خلال الايام الاربعة الماضية

Output MT 1. [Added I think the relationship has reached the level of what we see in terms of honesty and frankness with officials of Korea the Democrats we met with them during the last four days.]

ADEQUACY

FLUENCY

Output MT 2. ["I think the relationship has reached a level that we have not seen in terms of honesty and honesty with Democratic Korean officials we have met with over the past four days," he said.]

ADEQUACY

FLUENCY

Output MT 3.["I think the relationship has reached a level that we have not seen in terms of honesty and frankness with the DPRK officials we have met with over the past four days," he added.]

ADEQUACY

FLUENCY

Output MT 4.[I think the relationship has reached a level that we have not witnessed in terms of sincerity and sincerity with democratic Korean officials we have met in the past four days.]

ADEQUACY

FLUENCY

ST

واوضح هذا الاكتشاف ان بعض الحيوانات الثديية كبيرة الحجم التى عاشت فى هذا العصر قد
تكون آكلة للحوم ولديها الشجاعة الكافية لمنافسة الديناصورات على الغذاء ومكان المعيشة.

Output MT 1. [He explained this discovery to some of the mammals of great size who lived in this era may be carnivorous and have the courage to compete with dinosaurs for food and knowledge.]

ADEQUACY

FLUENCY

Output MT 2. [Some large mammals that lived in this age may be carnivorous and have the courage to compete with dinosaurs for food and living space, the discovery said.]

ADEQUACY

FLUENCY

Output MT 3. [This discovery indicated that some large-sized mammals that lived in this age may be carnivores and have the courage to compete with dinosaurs for food and the place of living.]

ADEQUACY

FLUENCY

Output MT 4. [The discovery said some large mammal animals that lived in this era may be a meat eater and have the courage to compete with dinosaurs for food and living space.]

ADEQUACY

FLUENCY

ST

وقال ان اوغندا لا يمكن ان تتفاوض. اننا اصغر من ان نتفاوض انك لا تستطيع ان تكون ضعيفا
وتتفاوض.

Output MT 1. [He said that Uganda could not negotiate. I'm too young to negotiate you can't be weak and negotiating.]

ADEQUACY

FLUENCY

Output MT 2. [Uganda cannot negotiate. We are too small to negotiate that you cannot be weak and negotiate.]

ADEQUACY

FLUENCY

Output MT 3. [He said that Uganda could not negotiate.We are too young to negotiate that you cannot be weak and negotiate.]

ADEQUACY

FLUENCY

Output MT 4. [He said Uganda could not negotiate. We are too small to negotiate that you cannot be weak and negotiated.]

ADEQUACY

FLUENCY

ST

يذكر ان اخر مأساة اسفرت عن خسائر جسيمة فى صفوف السويديين وقعت عام 1994 عندما
غرق مركب فى بحر البلطيق, الامر الذى ادى الى غرق 892 شخصا بينهم 551 سويديا.

Output MT 1. [Recall that another tragedy resulted in heavy losses in the ranks of the Swedes occurred in 1994 when the boat sank in the Baltic Sea, which led to the sinking of 892 people, including 551 Swedes]

ADEQUACY

FLUENCY

Output MT 2. [The last tragedy, which caused heavy damage to Swedes, occurred in 1994 when a boat sank in the Baltic Sea, killing 892 people, including 551 Swedes.]

ADEQUACY

FLUENCY

Output MT 3. [It is noteworthy that the last tragedy resulted in massive losses among the Swedes, which occurred in 1994 when a boat sank in the Baltic Sea, which resulted in the drowning of 892 people, including 551 Swedes.]

ADEQUACY

FLUENCY

Output MT 4. [The latest tragedy caused heavy losses among the Swedes in 1994 when a boat drowned in the Baltic Sea, which sank 892 people, including 551 Swedes]

ADEQUACY

FLUENCY

English to Arabic translation

Please read the source text (ST) and the output of FOUR MT systems, then rank Adequacy and Fluency on a scale of 1 to 4 for each, according to the following definitions and criteria:

ADEQUACY: ‘How much of the meaning expressed in the source fragment appears in the translation fragment?’

1= None of it.

2= Little of it.

3= Most of it.

4= All of it.

FLUENCY: Is the target text well-formed grammatically "so that it contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker”?

1= Incomprehensible.

2= Dis-fluent.

3= Good.

4= Flawless.

ST

De Villepin also said that the number of practicing Muslims in France is in about the same range as for other religions, in other words less than 10% out of five million people.

من 10 %من أصل خمسة ملايين شخص دو فيلبان قال أيضا أن عدد من ممارسة المسلمين في فرنسا هو في نفس .
1 MT Output.النطاق مثل الأديان الأخرى ، وبعبارة أخرى أقل

ADEQUACY

FLUENCY

كما قال دو فيليبان ان عدد المسلمين الممارسين فى فرنسا فى نفس النطاق تقريبا بالنسبة للديانات الاخرى ، وبعبارة اخرى
2 MT Output.اقل من 10 فى المائة من بين خمسة ملايين شخص

ADEQUACY

FLUENCY

أخرى أقل من 10 ٪ من أصل خمسة ملايين شخص كما قال دي فيلبان إن عدد المسلمين الممارسين في فرنسا هو في .
3 MT Output.نفس النطاق تقريبًا بالنسبة للأديان الأخرى ، وبعبارة

ADEQUACY

FLUENCY

أوقال دي فيلبين أيضاً أن عدد المسلمين الذين يمارسون مهنة في فرنسا في نفس النطاق الذي يتدربون عليه في الديانات
4 MT Output.الأخرى، وبعبارة أخرى أقل من 10 في المائة من خمسة ملايين شخص

ADEQUACY

FLUENCY

ST

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

أوريخويلا بدت هادئة تماما كما كان يجري أدى إلى الطائرة الأمريكية التي من شأنها أن تأخذه إلى ميامي في ولاية .
فلوريدا.1 MT Output

ADEQUACY

FLUENCY

Orejuela هادئة جدا كما كان يجري نقله إلى الطائرة الأمريكية التي من شأنها أن تأخذه إلى ميامي في ولاية فلوريدا.
بدا.2 MT Output

ADEQUACY

FLUENCY

بدا.3 MT Output Orejuela ًا تماما لأنه كان يقود إلى الطائرة الأمريكية التي ستأخذه إلى ميامي في فلوريدا هادئ ً

ADEQUACY

FLUENCY

Orejuela.4 MT Output بدا هادئاً تماماً بينما كان يقود إلى الطائرة الأمريكية التي ستأخذه إلى (ميامي) في (فلوريدا)

ADEQUACY

FLUENCY

ST

The statement said that the Arab taking a position, especially its intellectuals and thinkers, is a historical responsibility whose surrender is a surrender of the right to expression and a betrayal of every message and value, spiritually and culturally.

وقال البيان ان العربية اتخاذ موقف ، وخاصة المثقفين والمفكرين هي المسؤولية التاريخية التي الاستسلام هو
الاستسلام.للحق في التعبير و الخيانة من كل رسالة قيمة روحيا وثقافيا 1 MT Output

ADEQUACY

FLUENCY

وقال البيان ان اتخاذ العرب لموقف وخاصة مثقفه ومفكربه مسؤولية تاريخية ان استسلامها هو تنازل عن حق التعبير
وخيانة لكل رسالة وقيمة روحيا وثقافيا. 2 MT Output

ADEQUACY

FLUENCY

وقال البيان إن العرب الذين يتخذون موقفا ، وخاصة المثقفين والمفكرين ، مسؤولية تاريخية ، واستسلامها هو تنازل عن
الحق في التعبير وخيانة لكل رسالة وقيمة روحيا وثقافيا. 3 MT Output

ADEQUACY

FLUENCY

وقال البيان إن العرب الذين يتخذون موقفا، ولا سيما المفكرين والمفكرين، مسؤولية تاريخية، يستسلمون الحق في التعبير
وخيانة لكل رسالة وقيمتها وروحيا وثقافيا. 4 MT Output

ADEQUACY

FLUENCY

ST

Kenteris (31 years old) and Thanou (29 years old) missed a drug test on August 12 on the eve of the Olympic Games in Athens, saying they had been in a motorcycle traffic accident.

عاما) غاب عن اختبار المخدرات في 12 آب عشية الألعاب الأولمبية في أثينا ، قائلا أنها كانت في دراجة نارية في حادث
1 MT Output .Kenteris) 31 عاما (Thanou) 29 مروري

ADEQUACY

FLUENCY

وكان كانينيليس (31 عاما) وثانو (29 عاما) تغيبا عن اختبار المخدرات في 12 اب/اغسطس عشية دورة الالعاب
MT Output 2.الاولمبية.في اثينا، مؤكدين انهما تعرضا لحادث سير على الدراجات النارية

ADEQUACY

FLUENCY

Kenteris) 31 عاما) وً Thanou) 29 ، عاما) عن اختبار مخدرات في 12 أغسطس عشية الألعاب الأولمبية في أثيناً
MT Output 3.غاب قائلين إنهما تعرضا لحادث مروري على دراجة نارية

ADEQUACY

FLUENCY

عاما) اختصار للمخدرات في 12 أغسطس عشية الألعاب الأولمبية في أثينا، قائلاً 4 .أنهما كانا في حادث حركة الدراجات
MT Output . فقد كان Kenteris) 31 سنة (and) 29النارية

ADEQUACY

FLUENCY

Thanks for participating in the questionnaire

Appendix Three (3): Experiment A (segments including a move unit)

TM Input
3WSL (word-sentence-length)
يشجب باول الإسلاموفبيا
يواجه ساركوزي رمضان
امرأة بدأت بالتساؤل
ًأبي سيكون فخورا
ًيبدو هذا جيدا
4WSL
.الأمر سيكون مثيرا للاهتمام
،كهذه رسالة تكون مضرة
تنظم السعودية زواج الفتيات
يبدو اليورو بديلا جيدا
ًيزور غول سوريا غدا
يواجه الرئيس أزمة كبيرة
يحدث باحث تويتر بأفكاره

5WSL
الاستنساخ يظل مسألة مثيرة للجدل.
توقف إيران العمرة خلال رمضان
يتحرك العالم لاحتواء انتشار الأنفلونزا
يتراجع التدخين عالميا ويتزايد عربيا
يستهدف تفجير سفارة فرنسا بنواكشوط
تشيد إيران بالمكاسب التكنولوجية النووية
تنشيء الكويت لجنة للطاقة النووية
يحبط خطاب البابا قادة المسلمين.
يصنع المنهج الصحيح مجتمع متقدم
ينتهك جنون الإرهاب حقوق الانسان
يكافح زيت الزيتون أمراض القلب
6WSL
يستغل مراهق ثغرة في موقع تويتر
يقلب أوباما القاهرة رأسا على عقب
تترقب مصر تشكيل الحكومة الإسرائيلية الجديدة

تستعد غزة لمهمة القصف الإسرائلي الجديد
القذافي أجبر كل شخص على الحضور
يتصارع تويتر مع هجوم الدودة الرابع
تقبل حزب الله الهزيمة في انتخابات
تختتم قمة الدوحة على إيقاع التذمر
تلحق الأزمة العالمية الأذى بحقوق الإنسان
تظهر روح التطوع بين الثقافات المختلفة
7WSL
تنشغل باريس مع الشعب الفلسطيني في السلام.
تعترض إسرائيل سفينة المساعدات الليبية لقطاع غزة
تحتفل القدس بمهرجان فلسطين للادب الاسبوع القادم
يهدد أولمرت برد قوي على صواريخ غزة
تحظر إسرائيل الاحتفال بالقدس عاصمة للثقافة العربية
يكتشف سورى علاجاً جديداً لمرض سرطان الثدى
تواجه المرأة السعودية حظراً على صالات الجيم
يستعد سعد الحريري لتولي رئاسة وزارة لبنان

تهاجم القوات الهندية الفنادق لتطلق سراح الرهائن
يجري مستشفى فرنسي عمليات زرع للوجه واليدين
تحذر الحكومة المصرية من إختفاء مقابر الفراعنة
تواجه العملة الخليجية الموحدة المزيد من العقبات.
يهدد انقطاع التيار الكهربائي حياة المرضى بغزة
يقطع الرئيس الإسرائيلي رابين العلاقات مع الفاتيكان
8WSL
تعمق فنزويلا الجراح بدعوة مسؤولي السلطة الوطنية الفلسطينية
يقول موسوي إنه الفائز المؤكد في الانتخابات الإيرانية
تعدم الصين شخصين فى فضيحة حليب الأطفال المسمم
تشيد حماس بهجوم الجرار الذي وقع في القدس
تحث اليابان مواطنيها على الهدوء حيال صواريخ كوريا
يدعو الرئيس الصومالي المسلحين لعقد هدنة في رمضان
تقاضي امراة مسلمة قاضي في حادث غطاء الرأس
يتقدم أحمدي نجاد على موسوي في النتائج الأولية
يختتم مؤتمر حوار الأديان أعماله بدعم قيم التسامح

تفتح السلطة الوطنية الفلسطينية بعثة دبلوماسية في فنزويلا
تشعر رابطة الكتاب الأردنيين بالصدمة بسبب حكم السجن
تحرر قوات النيتو بأفغانستان صحفي نيويورك تايمز البريطاني
تعود اللغة الآرامية المهددة بالخطر مرةً أخرى بسوريا
تمنح خطة الانسحاب من العراق بصيصا من الأمل
يكرّمون رجال الدين في حلب سعد الله ونّوس
يسلم أول متهم بجرائم دارفور نفسه للجنائية الدولية
9WSL
تعلن السعودية مشاركتها بمؤتمر إعادة إعمار غزة في مصر
يبحث البرلمان أوضاع المصريين بالسعودية وتطورات قضية الطبيبين المجلودين
يزعم العراق القبض على خلية مزعومة تابعة لتنظيم القاعدة
يتعهد موسوي في إيران بمراجعة قوانين المرأة 'غير العادلة'
تنفي الحكومة المقالة علاقة المقاومة بصواريخ أطلقت من غزة
يفسر الرئيس السوري سبب فشل المحادثات السابقة مع إسرئيل.
تهدد كوريا الشمالية بشن حرب لحماية صوارخها بعيد المدى
يقترح حزب ليبرمان فرض حظر على إحياء النكبة العربية

تستعيد المملكة البريطانية حواراتها مع حزب الله مرةً أخرى
تدشن نتائج الانتخابات الإسرائيلية سنوات من الفتور مع مصر
تكشف وزارة الدفاع الأمريكية عن صور التعذيب داخل السجون
تتهم المحكمة الجنائية الدولية رئيس السودان بارتكاب جرائم حرب
تحظر تي موبايل بألمانيا استخدام سكايب على أي فون
يتوقف الرئيس الأمريكى باراك أوباما فى زيارة مفاجئة بالعراق
يقرأ فريدرك كانوتية لاعب اشبيلية الفاتحة قبل نزول الملعب
يلتقي المبعوث الأمريكي جورج ميتشل وفدا من حركة فتح
تلامس القضايا التي يطرحها المنتدى اهتمامات اساسية للمواطن العربي
10WSL
يصدر الجيش أوامر بوقف تحضيرات زيارة البابا في بيت لحم
يرمي المشروع إلى حصر وتوثيق جميع الآثار الموجودة على مصر
يدرس أوباما احتجاز الإرهابيين المشتبه فيهم إلى أجل غير مسمى
يشيد دبلوماسي في انفراج العلاقات بين الولايات المتحدة و سوريا
يعطي فايس بوك المستخدمين الفرصة للتعبير عن رأيهم بشأن السياسات
يعني انقطاع الانترنت التوقف عن الانتاج في العديد من المؤسسات

يؤكد استطلاع رأي أن عيد الحب يهدف العالم الغربي فقط
يشكك نظام غير آمن بمواطنيه يعد بداية النهاية للنظام القائم
يسحب مكتب السياحة الإسرائيلي إعلاناته بعد تقديم شكاوى من الخرائط
يعلن رئيس الوزراء التايلاندي حالة الطوارئ، وينجو من هجوم جماهيري
تقدم صحيفة الأخبار المصرية معطيات حول القمة العربية في الدوحة.
تخفق الوكالة الدولية للطاقة الذرية مجددا في اختيار خلفا للبرادعي
يحذر تقرير جديد للأمم المتحدة من تزايد الضغوط على المياه
تسخر تي شيرتات الجيش الإسرائيلي من عمليات القتل في غزة
شملت الدعوات الى حفل الزفاف العديد من مشاهير العالم العربي

Appendix Four (4): Experiment B (Segments including a three-operation edit)

TM Input
تحقيق يمكن هذه الغاية بطريقتين:
الاستنساخ يظل مسألة مثيرة للجدل.
البابا خطاب يحبط قادة المسلمين.
تحقيق هذه الغاية بطريقتين:
الاستنساخ مسألة مثيرة للجدل.
البابا يحبط قادة المسلمين.
يدوي يمكن تحقيق هذه الغاية بطريقتين:
يدوي يظل الاستنساخ مسألة مثيرة للجدل.
يدوي خطاب البابا يحبط قادة المسلمين.
فريق تحقيق هذه الغاية بطريقتين:
صحف الاستنساخ مسألة مثيرة للجدل.
فريق البابا يحبط قادة المسلمين.
إسرائيل قيام بتوجيه ضربة عسكرية لإيران،

؟البرقع حظر أم منع الحظر نفسه
.أنهم الآن سيسحقون أي مظاهرات أخرى
،إسرائيل بتوجيه ضربة عسكرية لإيران
.أنهم سيسحقون أي مظاهرات أخرى
؟البرقع أم منع الحظر نفسه
،محلية قيام إسرائيل بتوجيه ضربة عسكرية لإيران
.يمكن الآن أنهم سيسحقون أي مظاهرات أخرى
.عسكرية على الولايات المتحدة مواصلة الرد الحذر
،فريق إسرائيل بتوجيه ضربة عسكرية لإيران
.يظل أنهم سيسحقون أي مظاهرات أخرى
.يظل الولايات المتحدة مواصلة الرد الحذر
.تنشغل باريس مع الشعب الفلسطيني في السلام
.ثمة سيكون مركز للإعلام الدولي بجامعة القاهرة
.الخليجية العملة الموحدة تواجه المزيد من العقبات
.تنشغل مع الشعب الفلسطيني في السلام
.ثمة مركز للإعلام الدولي بجامعة القاهرة

الخليجية الموحدة تواجه المزيد من العقبات.
صحف باريس تنشغل مع الشعب الفلسطيني في السلام.
فريق سيكون ثمة مركز للإعلام الدولي بجامعة القاهرة.
سيكون العملة الخليجية الموحدة تواجه المزيد من العقبات.
عملية تنشغل مع الشعب الفلسطيني في السلام.
عملية ثمة مركز للإعلام الدولي بجامعة القاهرة.
عسكرية الخليجية الموحدة تواجه المزيد من العقبات.
نظام يعتمد البطاقات مسبقة الدفع على ثلاثة عوامل:
هذه استخدمت التجهيزات أيضاً لتوجيه الهوائي بشكل صحيح.
الزيارة ستتم بين الساعة العاشرة صباحا والسادسة مساء.
كل يعمل قطاع وكأنه شبكة إيثرنت محلية مستقلة.
نظام البطاقات مسبقة الدفع على ثلاثة عوامل:
كل قطاع وكأنه شبكة إيثرنت محلية مستقلة.
هذه التجهيزات أيضاً لتوجيه الهوائي بشكل صحيح.
صحف يعتمد نظام البطاقات مسبقة الدفع على ثلاثة عوامل:
فريق يعمل كل قطاع وكأنه شبكة إيثرنت محلية مستقلة.

سيكون استخدمت هذه التجهيزات أيضاً لتوجيه الهوائي بشكل صحيح.
للجدل نظام البطاقات مسبقة الدفع على ثلاثة عوامل:
ضربة كل قطاع وكأنه شبكة إيثرنت محلية مستقلة.
كمبيوتر هذه التجهيزات أيضاً لتوجيه الهوائي بشكل صحيح.
الخطاب سيمتد لخمسين دقيقة، هذه مدة طويلة نوعا ما.
الكونجرس يعمل على سرعة تنفيذ مشروع قانون تحفيز الاقتصاد.
مركز يقع الولوج البعيد ضمن محطة محلية للبث الإذاعي.
هيئات تستخدم الإغاثة والسلطات المحلية أيضاً الشبكة بشكل مكثف.
الحزب أعضاء الوطني الديمقراطي يظلون كما هم أينما حلوا!
مركز الولوج البعيد ضمن محطة محلية للبث الإذاعي.
هيئات الإغاثة والسلطات المحلية أيضاً الشبكة بشكل مكثف.
الحزب الوطني الديمقراطي يظلون كما هم أينما حلوا!
صحف يقع مركز الولوج البعيد ضمن محطة محلية للبث الإذاعي.
فريق تستخدم هيئات الإغاثة والسلطات المحلية أيضاً الشبكة بشكل مكثف.
سيكون أعضاء الحزب الوطني الديمقراطي يظلون كما هم أينما حلوا!
صحف مركز الولوج البعيد ضمن محطة محلية للبث الإذاعي.

عسكرية هيئات الإغاثة والسلطات المحلية أيضاً الشبكة بشكل مكثف.
سيكون الحزب الوطني الديمقراطي يظلون كما هم أينما حلوا!

Appendix five (5): Experiment C (Segments including an inflection verb affix)

TM Input
3WSL (word-sentence-length)
3a [يــ]
.يشجب الرئيس الإسلاموفبيا
.يترك المهرج المنصة
.يقرأ الولد كتابه
3b with َ
.رتبَ الفتى المكان
.لعبَ الفتى بالكرة
3c no َ
.خرج الأب مسرورا
.ضغط الزر الأخضر
3d [تــــ]
.تبدأ امرأة بالتساؤل

.تخرج السيدة مسرعة
.تشرق الشمس مبكرا
3e[ـــت]
.كتبت الطفلة القصة
.فرحت الأسرة بالمولود
.هبطت الطائرة بسلام
4WSL
4a [يـ]
.يدعم الفريق المدرب الحالي
.يحمل ساعي البريد الطرود
.يجمع الفلاح الثمار الناضجة
4b with
.بحثَ الناس عن أيديولوجيته
.فقدَ الرئيس شعبية كبيرة
.حدثَ باحث تويتر بأفكاره
4c no

سبح البحار في البحر.
مشي جحا وراء الحمار.
قرأ الطالب في كتابه.
4d [تـــ]
تنظم الجمعية زواج جماعيا.
ترسم الفنانة لوحة جميلة.
تدرس شركة فرنسية الخطة.
4e[ـــت]
طبخت الأم وجبة الغذاء.
زينت الفتاة الطبق بالسلاطة.
أخذت الموظفة إجازة طويلة.
رسلت الزوجة لزوجها رسالة.
5WSL
5a [يـــ]
يشرب الطفل الحليب الطازج صباحا.
ينظر الوالد إلى ولده مفتخرا.

يفحص طبيب العيون عين أمي.
5b with ´
حسبَ سائق التاكسي الأجرة بدقة.
ذهبَ الرجل إلى عمله بالحافلة.
عزفَ المسيقار ألحان وطنية رائعة.
5c no ´
سكن الطالب مع عائلة إنجليزية.
عزف المسيقار ألحان وطنية رائعة.
5d [تـــ]
تجلس الجدة على الكرسي الجميل.
تحضر المديرة مع نائبها الاجتماع.
تسمح المحكمة الابتدائية بدخول المتضررين.
5e[ـــت]
شملت الجولة المشروع المتوقف بالقرية.
دفعت مستأجرة البيت الأجرة كاملة.
فتحت النافدة بقوة فوقع الستار.

6WSL
6a [يــ]
.ياكل الرجل مع زوجته في المطعم
. يصرف المدير مكافئة لكل الموظفين بسخاء
.يخلع طبيب الأسنان سن المريض ببراعة
6b with
.سحبَ الرئيس الوفد المشارك من المؤتمر
.رفضَ زعيم المتمردين مقابلة مبعوث الوساطة
6c no
.هدد المطر الغزير والسيول سكان القرية
.قرأ الشيخ المسن الصحف يوميا للمتعة
6d [تـــ]
.تحدث الحرب بين الدولتين لأسباب متعددة
.تظهر روح التطوع بين الثقافات المختلفة
.تعمل الوزارة على تأمين الدعم المالي
6e[ـــت]

قسمت الأم الكعكة إلى ست قطع.
شرحت المعلمة الفقرة شرحا تفصيليا دقيقا.
قدمت الصحافة تقريرا من العصابة الإجرامية.
7WSL
7a [يـ]
يحضر الوفد الوطني المشارك أعمال القمة المصغرة.
يفحص دكتور متخصص نبضات القلب والتنفس والدماغ.
يهدد انقطاع التيار الكهربائي حياة المرضى بالقرية.
7b with َ
خصمَ الرئيس التنفيدي للشركة المزاية عن المتغيبين.
سألَ الأستاذ سؤال ويطلب أجابة مختصرة ودقيقة.
خلطَ الدهان اللون الأحمر مع الاصفر بحرفية.
7c no َ
خصم الرئيس التنفيدي للشركة المزاية عن المتغيبين.
سأل الأستاذ سؤال ويطلب أجابة مختصرة ودقيقة.
خلط الدهان اللون الأحمر مع الاصفر بحرفية.

7d [تـــ]
تحجز المسافرة حجرة فردية بالفندق بوسط المدينة.
تحذر الحكومة الاهالي بعدم الاقتراب من المنحدر.
تغير وسائل التواصل الحديثة من حياة الناس.
7e[ـــت]
صدرت المحكمة حكما نهأئيا بالسجن مدا الحياة.
جددت الاسرة تاكيدها للمحامي بعدم سحب القضية.
سببت العملة المزورة انهيار في الاقتصاد الوطني.

Appendix Six (6): Experiment D (Segments with Hamza marker omission)

TM Input
3WSL (word-sentence-length)
Removing a Hamza of Alif in an initial-word position
اسعار النفط مشتعلة.
اطلاقها للصواريخ تهديد.
اليس هناك مصعد؟
Removing a Hamza of Alif in a mid-word position
بدات غالبية بالمغادرة.
الاراء و التعليقات.
الاغنام تنتج الصوف.
Removing a Hamza of Alif in a word-final position
المبتدا و الخبر.
الكلا تعني العشب.
التقيو يعني الاستفراغ.

شواطى بحيرة طبريا.
4WSL
Removing a Hamza of Alif in an initial-word position
اكبر ثالث مصدر للنحاس.
اغلبية الثلثين مطلوبة للفوز.
الا أن الأمر مستبعد.
Removing a Hamza of Alif in a mid-word position
تامر الشريعة الإسلامية بالذبح.
الادب الليبي يناقش التراث.
الانفاق الحكومي شحيح للغاية.
Removing a Hamza of Alif in a word-final position
الخطا يعالج بتكرار المحاولة.
سيصدا الحديد لتعرضه للهواء.
أتجرو أن تتهمني بالغدر.
الموانى البحرية هبة طبيعية.
5 word-SL

Removing a Hamza of Alif in an initial-word position
اجرة البيت كانت مكلفة جدا.
اسامح نفسي على الماضي ولكن.
اصداء الملاعب برنامج رياضي رائع.
Removing a Hamza of Alif in a mid-word position
تاكلت أسنانه بسب أكل الحلوة.
تاخرت الأستاذ عن الفصل لدقائق.
يتراس الملك اجتماعا وزاريا مهما.
Removing a Hamza of Alif in a word-final position
الظما يذهب عند إفطار الصائم.
يتوضا المسلم لأداء الصلوات الخمس.
يختبى خلف الصخور و الشقوق.
التجشو طرد الهواء من المعدة.
6WSL
Removing a Hamza of Alif in an initial-word position
اطلقت المحطة الفضائية قمرا جديداً للفضاء.

اعترف المتهم بدخوله البيت ليلاً للسرقة.
اعراض أعصاب المعدة الإرهاق و الإمساك.
Removing a Hamza of Alif in a mid-word position
متالق دائما يأبني سأمنحك جائزة رائعة.
تاثير الأكياس البلاستكية سيئ على البيئة.
الفار يعيش بالقرب من مصادر الطعام.
Removing a Hamza of Alif in a word-final position
يكافا المتميزين لحصولهم على المراتب الأولى.
يتبرا الأب من أولاده التاركين للصلاة.
يفاجى الإمام المصلين بصوت مؤثر جدا.
تباطو ضربات القلب يؤدي إلى التعب.
7WSL
Removing a Hamza of Alif in an initial-word position
انفقت الجمعية الخيرية أمولا ضخمة على الإيتام.
ابداء المشورة قبل اتخاذ القرار شيء جمي.
اغاني الأطفال لا ينصح بها قبل الخامسة.

Removing a Hamza of Alif in a mid-word position
سافعل ذلك لاحقا بعد نهاية الجزء المتبقي.
بداتم قرأت القرآن منذ العمر ست سنوات.
امراة مكافحة تريد تربية أبنائها الأيتام بالحلال.
Removing a Hamza of Alif in a word-final position
يبتدا بخطوة ثم تتبعها خطوات ليحقق مبتغاه.
يمتلا الطابق الأول و كذلك نصف الثاني.
يلتجى إلى افتعال و صناعة الأزمات الاقتصادية.
الوجه أحد علامات الحياة الزوجية السعيدة تلألو.

Appendix Seven (7): Experiment E (Segments used in a TM test translated using a MT system)

MT Input
3WSL (word-sentence-length)
باول يشجب الإسلاموفبيا
ساركوزي يواجه رمضان
بدأت امرأة بالتساؤل
سيكون أبي فخوراً
هذا يبدو جيداً
4WSL
سيكون الأمر مثيرا للاهتمام.
رسالة كهذه تكون مضرة،
السعودية تنظم زواج الفتيات
اليورو يبدو بديلا جيدا
غول يزور سوريا غداً
الرئيس يواجه أزمة كبيرة

باحث يحدث تويتر بأفكاره
5WSL
يظل الاستنساخ مسألة مثيرة للجدل.
إيران توقف العمرة خلال رمضان
العالم يتحرك لاحتواء انتشار الأنفلونزا
التدخين يتراجع عالميا ويتزايد عربيا
تفجير يستهدف سفارة فرنسا بنواكشوط
إيران تشيد بالمكاسب التكنولوجية النووية
الكويت تنشيء لجنة للطاقة النووية
خطاب البابا يحبط قادة المسلمين.
المنهج الصحيح يصنع مجتمع متقدم
جنون الإرهاب ينتهك حقوق الانسان
زيت الزيتون يكافح أمراض القلب
6WSL
مراهق يستغل ثغرة في موقع تويتر
أوباما يقلب القاهرة رأسا على عقب

مصر تترقب تشكيل الحكومة الإسرائيلية الجديدة
غزة تستعد لمهمة القصف الإسرائلي الجديد
أجبر القذافي كل شخص على الحضور
تويتر يتصارع مع هجوم الدودة الرابع
حزب الله تقبل الهزيمة في انتخابات
قمة الدوحة تختتم على إيقاع التذمر
الأزمة العالمية تلحق الأذى بحقوق الإنسان
روح التطوع تظهر بين الثقافات المختلفة
7WSL
باريس تنشغل مع الشعب الفلسطيني في السلام.
إسرائيل تعترض سفينة المساعدات الليبية لقطاع غزة
القدس تحتفل بمهرجان فلسطين للادب الاسبوع القادم
أولمرت يهدد برد قوي على صواريخ غزة
إسرائيل تحظر الاحتفال بالقدس عاصمة للثقافة العربية
سورى يكتشف علاجاً جديداً لمرض سرطان الثدى
المرأة السعودية تواجه حظراً على صالات الجيم

سعد الحريري يستعد لتولي رئاسة وزارة لبنان
القوات الهندية تهاجم الفنادق لتطلق سراح الرهائن
مستشفى فرنسي يجري عمليات زرع للوجه واليدين
الحكومة المصرية تحذر من إختفاء مقابر الفراعنة
العملة الخليجية الموحدة تواجه المزيد من العقبات.
انقطاع التيار الكهربائي يهدد حياة المرضى بغزة
الرئيس الإسرائيلي رابين يقطع العلاقات مع الفاتيكان
8WSL
فنزويلا تعمق الجراح بدعوة مسؤولي السلطة الوطنية الفلسطينية
موسوي يقول إنه الفائز المؤكد في الانتخابات الإيرانية
الصين تعدم شخصين فى فضيحة حليب الأطفال المسمم
حماس تشيد بهجوم الجرار الذي وقع في القدس
اليابان تحث مواطنيها على الهدوء حيال صواريخ كوريا
الرئيس الصومالي يدعو المسلحين لعقد هدنة في رمضان
امراة مسلمة تقاضي قاضي في حادث غطاء الرأس
أحمدي نجاد يتقدم على موسوي في النتائج الأولية

مؤتمر حوار الأديان يختتم أعماله بدعم قيم التسامح
السلطة الوطنية الفلسطينية تفتح بعثة دبلوماسية في فنزويلا
رابطة الكتاب الأردنيين تشعر بالصدمة بسبب حكم السجن
قوات النيتو بأفغانستان تحرر صحفي نيويورك تايمز البريطاني
اللغة الآرامية المهددة بالخطر تعود مرةً أخرى بسوريا
خطة الانسحاب من العراق تمنح بصيصا من الأمل
رجال الدين في حلب يكرّمون سعد الله ونّوس
أول متهم بجرائم دارفور يسلم نفسه للجنائية الدولية
9WSL
السعودية تعلن مشاركتها بمؤتمر إعادة إعمار غزة في مصر
البرلمان يبحث أوضاع المصريين بالسعودية وتطورات قضية الطبيبين المجلودين
العراق يزعم القبض على خلية مزعومة تابعة لتنظيم القاعدة
موسوي يتعهد في إيران بمراجعة قوانين المرأة 'غير العادلة'
الحكومة المقالة تنفي علاقة المقاومة بصواريخ أطلقت من غزة
الرئيس السوري يفسر سبب فشل المحادثات السابقة مع إسرئيل.
كوريا الشمالية تهدد بشن حرب لحماية صوارخها بعيد المدى

حزب ليبرمان يقترح فرض حظر على إحياء النكبة العربية
المملكة البريطانية تستعيد حواراتها مع حزب الله مرةً أخرى
نتائج الانتخابات الإسرائيلية تدشن سنوات من الفتور مع مصر
وزارة الدفاع الأمريكية تكشف عن صور التعذيب داخل السجون
المحكمة الجنائية الدولية تتهم رئيس السودان بارتكاب جرائم حرب
تي موبايل بألمانيا تحظر استخدام سكايب على أي فون
الرئيس الأمريكى باراك أوباما يتوقف فى زيارة مفاجئة بالعراق
فريدرك كانوتية لاعب اشبيلية يقرأ الفاتحة قبل نزول الملعب
المبعوث الأمريكي جورج ميتشل يلتقي وفدا من حركة فتح
القضايا التي يطرحها المنتدى تلامس اهتمامات اساسية للمواطن العربي
10WSL
الجيش يصدر أوامر بوقف تحضيرات زيارة البابا في بيت لحم
المشروع يرمي إلى حصر وتوثيق جميع الآثار الموجودة على مصر
أوباما يدرس احتجاز الإرهابيين المشتبه فيهم إلى أجل غير مسمى
يشيد دبلوماسي في انفراج العلاقات بين الولايات المتحدة و سوريا
فايس بوك يعطي المستخدمين الفرصة للتعبير عن رأيهم بشأن السياسات

انقطاع الانترنت يعني التوقف عن الانتاج في العديد من المؤسسات
استطلاع رأي يؤكد أن عيد الحب يهدف العالم الغربي فقط
نظام غير آمن يشكك بمواطنيه يعد بداية النهاية للنظام القائم
مكتب السياحة الإسرائيلي يسحب إعلاناته بعد تقديم شكاوى من الخرائط
رئيس الوزراء التايلاندي يعلن حالة الطوارئ، وينجو من هجوم جماهيري
صحيفة الأخبار المصرية تقدم معطيات حول القمة العربية في الدوحة.
الوكالة الدولية للطاقة الذرية تخفق مجددا في اختيار خلفا للبرادعي
تقرير جديد للأمم المتحدة يحذر من تزايد الضغوط على المياه
تي شيرتات الجيش الإسرائيلي تسخر من عمليات القتل في غزة
الدعوات الى حفل الزفاف شملت العديد من مشاهير العالم العربي

Appendix Eight (8): Reference translation extracted from LDC (Linguistic Data Consortium) corpus “Arabic-English corpus LDC2004T18”

	Source Arabi		**Reference English translations (LDC corpus)**
Q1	وأضاف اعتقد ان العلاقة وصلت الى مستوى لم نشهده من حيث الصدق والصراحة مع مسئولين كوريين ديمقراطيين اجتمعنا معهم خلال الايام الاربعة الماضية	R1	He added ‘I believe relations have reached an unprecedented level of truth and openness with the Democratic Korean officials we have met with over the past four days.’
		R2	He added ‘I believe that the relationship reached an unprecedented level of trust and frankness with the Democratic Korean officials we met during the past four days.’
		R3	He added ‘I believe that the relationship has reached an unprecedented level of honesty and candor with the Democratic Korean officials we met with during the past four days.’
		R4	He went on to say, ‘the relationship, I think, has reached the level that we’ve not seen in terms of candor and openness with the Democratic Korean officials we met over the past four days.’
Q2	واوضح هذا الاكتشاف ان بعض الحيوانات الثديية كبيرة الحجم التى عاشت فى هذا العصر قد تكون آكلة للحوم ولديها الشجاعة الكافية لمنافسة الديناصورات على الغذاء ومكان المعيشة	R1	This discovery reveals that some large-sized mammals which lived in that era may have been carnivorous and brave enough to compete with dinosaurs for food and living space.
		R2	The discovery revealed that some large-size mammals that lived in that era may have been carnivorous, and brave enough to compete with dinosaurs for food and living space.
		R3	This discovery revealed that some large mammals living during that era could have been carnivorous and may have had enough courage to compete with the dinosaurs over food and shelter.

		R4	This discovery revealed that some large size mammals living in that era could have been carnivorous and brave enough to compete with dinosaurs for food and living space.
Q3	وقال ان اوغندا لا يمكن ان تتفاوض. اننا اصغر من ان نتفاوض انك لا تستطيع ان تكون ضعيفا وتتفاوض.	R1	He said Uganda cannot negotiate. We are too small to negotiate ... you cannot be weak and negotiate.
		R2	He said Uganda cannot negotiate. We are too small to negotiate... you cannot be weak and negotiate.
		R3	He said Uganda cannot negotiate. We are too small to negotiate. You cannot be weak and negotiate.
		R4	He said Uganda cannot negotiate. We are too small to negotiate... you cannot be weak and negotiate.
	يذكر ان اخر مأساة اسفرت عن خسائر جسيمة فى صفوف السويديين وقعت عام 1994 عندما غرق مركب فى بحر البلطيق, الامر الذى ادى الى غرق 892 شخصا بينهم 551 سويديا	R1	It is noteworthy the last tragedy which resulted in a huge Swedish death toll was in 1994 when a boat sank in the Baltic Sea, killing 892 people among them 551 Swedes.
		R2	The last tragedy that caused large scale casualties among the Swedes took place in 1994 when a ship sank in the Baltic Sea, which led to the drowning of 892 people, of whom were 551 Swedes.
		R3	The last tragedy causing mass casualties of Swedes happened in 1994, when a ferry sank in the Baltic Sea and 892 people drowned, including 551 Swedes.
		R4	The last tragedy which resulted in enormous losses in Swedish ranks happened in 1994, when a ship sank in the Baltic Sea,

			leading to the drowning of 892 people, among them 551 Swedes.

	Source- English		Reference Arabic translations (LDC corpus)
Q5	De Villepin also said that the number of practising Muslims in France is in about the same range as for other religions, in other words less than 10% out of five million people.	R1	وقال دو فيلبان ايضا ان عدد المسلمين الممارسين في فرنسا هم تقريبا بنفس الشريحة لدى الاديان الاخرى اقل من 10 %من اصل خمسة ملايين نسمة
Q6	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.	R1	وبدا اوريخويلا هادئا عندما اقتيد الى الطائرة الاميركية التي ستنقله الى ميامي في فلوريدا.
Q7	The statement said that the Arab taking a position,[71] especially its intellectuals and thinkers, is a historical responsibility whose surrender is a surrender of the right to expression	R1	واعتبر البيان ان اتخاذ موقف من قبل العرب وخاصة مثقفيهم ومفكريهم مسؤولية تاريخية كل تنازل عنها تنازل عن الحق في التعبير وخذلان لكل رسالة وقيمة، روحا وحضارة

[71] Although this is not grammatically correct English, we cannot intervene in the corpus sentence.

	and a betrayal of every message and value, spiritually and culturally.		
Q8	Kenteris (31 years old) and Thanou (29 years old) missed a drug test on August 12 on the eve of the Olympic Games in Athens, saying they had been in a motorcycle traffic accident.	R1	وقد تخلف كنتيريس)31 عاما (وثانو)29 عاما (عن الخضوع لفحص للكشف عن المنشطات في 12 اب/اغسطس عشية انطلاق دورة الالعاب الاولمبية التي أقيمت في أثينا، وادعيا بانهما تعرضا لحادث سير على دراجة نارية