DOCTOR OF PHILOSOPHY

Building and verifying parallel corpora between Arabic and English

Alkahtani, Saad

Award date:
2015

Awarding institution:
Bangor University

Bangor University

Doctoral Thesis

Building and verifying parallel corpora between Arabic and English

Author:
Saad Alkahtani

Supervisor:
Dr. William J. Teahan

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Computer Systems

in the

School of Computer Science

November 23, 2015

Abstract

Arabic and English are acknowledged as two major natural languages used by many countries and regions. Reviews of previous literature conclude that machine translation (MT) between these languages is disappointing and unsatisfactory due to its poor quality.

This research aims to improve the translation quality of MT between Arabic and English by developing higher quality parallel corpora. The thesis developed a higher quality parallel test corpus, based on corpora from *Al Hayat* articles and the OPUS open-source online corpora database.

A new Prediction by Partial Matching (PPM)-based metric for sentence alignment has been applied to verify quality in translation between the sentence pairs in the test corpus. This metric combines two techniques; the traditional approach is based on sentence length and the other is based on compression code length. A higher quality parallel corpus has been constructed from the existing resources. Obtaining sentences and words from two online sources, *Al Hayat* and OPUS, the new corpus offers 27,775,663 words in Arabic and 30,808,480 in English. Experimental results on sample data indicate that the PPM-based and sentence length technique for sentence alignment on this corpus improves accuracy of alignment compared to sentence length alone.

Acknowledgements

This thesis would not have been possible without the help of so many people during its writing. First of all, I am very grateful to the eminent and enthusiastic professor, William J. Teahan. As my principal mentor and my dedicated supervisor, he made various important recommendations which improved the thesis in both content and structure. Had I not read and had experiences in undertaking research on machine translation with Dr. Teahan at Bangor University, I would not have decided to pursue my study in the direction of machine translation and parallel corpora. It was he who first suggested I focus my Masters dissertation on this research direction, and his research achievements so far have inspired me to explore my own contribution to improving MT in Arabic and English in both academic and practical applications.

Next, I wish to thank my mother and also my wife. They are the most important people in the world to me and I would not have finished my doctoral project as well as my thesis without their care and support. I also want to thank my children, who have been one of the main reasons driving me to continue to conduct this research because I believe my project on PPMD-based MT will make their lives better and more confident in the future. My children are my constant source of happiness, because they have helped me to resist the pressures of this project, always telling me that I am doing a great thing to make the world better.

Last but not least, I must acknowledge my appreciation to my colleague and working partner, Mr. Wei Liu. I first met Wei on the Java programming course at Masters level at the University of Bangor, and his unique understanding, knowledge and keenness to help other people left a great impression on me. We have been friends and research partners ever since. Wei also continues to conduct parallel corpora research between Chinese and English on his PhD. Consequently, we have always discussed problems which occurred in my research and his advice on has always been very helpful.

Without the help of all those mentioned above, I could not have finished this research smoothly and on schedule. They deserve to share in my joy and sense of achievement.

Contents

List of Figures

List of Tables

To my father who dreamed to seeing me finish my PhD before he passed away.

Chapter 1

Introduction

1.1 General Background to the Research

In 2013, Mohamed Abdelmageed Mansour, Assistant Professor of English Linguistics, at Assiut University in Egypt, reflected on the difficulties that he faced as an Arabic-speaking researcher specializing in language:

In a world of a revolutionary computer technology in the field of linguistics, it seems that the common practice among the Arab linguists in the Arab world is very much frustrating. The only thing that an Arab linguist who is conducting linguistic research can do is painstakingly sit in his own office either contriving his linguistic data or extracting his own corpus – a tedious process that involves reading through printed texts and manually recording his data. The linguistic results of this huge effort are not highly accurate because these data are far removed from real language use, not empirical or representative (Mansour 2013: 81).

The sense of frustration that he voiced is all the more understandable when he goes on to list some of the many uses that a good quality corpus can serve, based on his knowledge of those which already exist for the English language. He argues that a corpus can be used to trace the origins and development of a language, revealing the factors that have brought about the emergence of different dialectal variants. It can also be employed to produce geolinguistic maps that show where these variants occur and the territory they occupy. It can be used by lexicographers to facilitate the process of creating dictionaries, improving the information which they contain, and addressing some of the deficiencies currently found in Arabic dictionaries.

Moreover, he explains, corpora benefit areas of textual analysis such stylistics and pragmatics, helping to reveal previously unsuspected linguistic features, and they allow for more complete descriptive grammars to be constructed.

He then extols the virtues of parallel or bilingual corpora which can help facilitate research in the field of contrastive linguistics. Stubbs (1996) has previously likened the importance of the development of corpora as a tool in the study of language to that of the invention of the telescope in astronomy in that it allows linguists to see phenomena and to discover patterns that were not previously suspected. Bilingual corpora also allow both teachers and students to study Arabic and English grammatical structures, by focusing on real language use in context which according to Gavioli (1997 cited in Mansour 2013: 85) is an "incredibly effective language-learning activity".

Mansour finally refers briefly to the importance of corpora that contain translated texts from two or more languages since these can facilitate research in translation studies, assist in training translators and contribute to advancing linguistic translation theories. As an Associate Professor, Mansour's focus is perhaps understandably on the more academic applications of parallel corpora in terms of the research opportunities they offer. However, in the commercial translation environment, language corpora have become increasingly important in natural language processing, and in machine translation (MT) in particular.

Essentially, large-scale parallel corpora act as a vital resource. They are often used in the case of language modelling for the purposes of training statistical models which can in turn be used for constructing, developing and improving statistical-based MT systems and software. It is important to conduct research on corpora-based MT because this represents many different possibilities in various everyday scenarios, being applicable to diverse areas such as dialogue recognition, human-machine conversation, language processing and information sorting.

When MT technology is combined with large corpora it can become a vital and invaluable translation tool in the commercial environment or in the context of large multilingual entities such as the United Nations or the European Union, helping to make the process of translation faster, more accurate and more efficient.

(Mansour, 2013) notes in his article that "although corpora are widely available for English, there is very little available for the Arabic language [...] Throughout the Arab world we do not have one single corpus that we [Arabs] created ourselves"

(p.82). Mansour's comments are all the more surprising, when we bear in mind that Arabic is a global language, the official language of some 22 countries, one of the six official languages of the United Nations (UN, 2013) and is used as the liturgical language by Islamic communities (or ummah) spread across the five continents (Ghazzawi, 1986; Kitchen et al., 2009).[1]

An additional problem is that the few existing corpora of Arabic/English parallel texts are either of poor quality or expensive to purchase and, as a result, it is difficult for researchers and students to undertake academic studies in this area. Moreover, many of the existing corpora in the commercial domain are essentially based on pre-existing specialised dictionaries which serve as technical translation aids rather than MT soft packages (Guidère, 2002). The knock-on effect of not having sufficient quantities of this linguistic material available is that currently MT between Arabic and English produces results which are often low in terms of their accuracy and high in price, forcing business and organizations to perform the vast majority of their translation without the assistance of a computer, making this a time-consuming and often error-filled process (Taghipour et al., 2011).

Various reasons explain the scarcity of parallel corpora for Arabic/English, the first being that Arabic has always been considered one of the most difficult languages for both written and spoken language processing "due to its morphological, syntactic, phonetic and phonological properties" (Zughoul and Abu-Alshaar, 2005: 1022). In addition, processing information for inclusion in a corpus manually requires a large team of experts, significant financial backing, and normally takes a long time. Consequently, since the 1990s, the process of corpus development has usually been carried out by national governments or large third-party institutions such as universities (Koehn, 2005). Typically, a consortium of private and public organizations will often work together on such projects. Examples include the British National Corpus (BNC) (`http://www.natcorp.ox.ac.uk/corpus`), a 100 million-word collection of samples of written and spoken English from the late twentieth century. This corpus was the result of a collaboration between major dictionary publishers, academic research centres at the Universities of Oxford and Lancaster, and the British Library, illustrating the scope of the resources required.

Mansour's 2013 observations concerning the importance of language corpora and limited availability of these in the case of Arabic provide an eloquent rationale

[1] Arabic and its characteristic linguistic features are discussed in more detail in Chapter Two of this thesis.

arguing for the urgent need for work in this area. This research aims to respond to this situation. The study of corpora can support the growth of natural language processing methods. As a trend in global social interaction and communication, translating languages, also known as sub-languages, are frequently required. The nature and characteristics of translating language require the support of MT technology. In 1998, at the COLING (Computational Linguistics) Conference in Montreal the general consensus reached by participants was that:

- High quality MT requires two important sub-technologies, namely statistical models and large corpora.
- The studies of corpora-based MT should focus on the accuracy of translating outcomes.
- The developments of MT and corpora should be driven by social requirements and the interaction of natural languages.

Nonetheless, currently most MT systems and software continue to be developed solely on the basis of statistical models, which creates several problems for both MT systems and software. Many recent studies in MT and corpora acknowledge that due to the poor quality of the corpus in terms of accuracy and the limitations of semantic alignment technology, maintaining and updating corpora can present significant difficulties for end users (see, for example, Munteanu and Marcu, 2005; Stubbs, 1996).

Moreover, the metric which is used in existing MT systems and software has been acknowledged as another key problem in this field (Koehn, 2005) which has served as another motivator for this doctoral research. Therefore, the author decided to test whether a new hybrid MT metric which used a compression-based technique called Prediction by Partial Matching (PPM) would help to solve this problem.

1.2 Research Aims and Objectives

The primary aim of this study is to investigate novel compression-based methods which can be used as a means of aligning parallel corpora consisting of Arabic and English translated materials. Its secondary aim is to provide the tools which

can be used to help build a high-quality and low-cost parallel corpus of data from these two languages.

Therefore, the research objectives of this study can be described as follows:

- To produce a state-of-the-art Literature Review on the area of MT with a specific focus on the problem of aligning parallel corpora (Chapter Three).
- To improve and extend the existing parallel corpora of Arabic and English translated texts (Chapter Four).
- To apply a PPM-based compression metric to the problem of sentence alignment in the extended parallel corpus and evaluate the effectiveness of the PPM metric in comparison to other methods (Chapter Five).
- To investigate new hybrid sentence alignment methods that take advantage of both the traditional sentence alignment method and the PPM method, applying this novel hybrid method to the new parallel corpus (Chapter Six).

1.3 Contribution of this research

As established in previous studies, it is currently difficult to conduct research on parallel Arabic/English corpora since many of the existing corpora of this kind in the marketplace are expensive to purchase and highly specialised in nature. The few examples of Arabic/English corpora which are freely accessible contain a high degree of inaccuracy. So the first contribution of this thesis is to deliver techniques that will help with the construction of parallel corpora in Arabic and English since without this essential resource, future research possibilities in this field will remain limited.

In addition to developing a method for distinguishing between accurate and inaccurate translations, a further contribution of this research will be the creation of a new hybrid alignment method which can be used for sentence alignment within the corpus. It is anticipated that this will establish a solid foundation on which future developments can be built.

Although this thesis focuses mainly on alignment at the sentence level, it is hoped that these results will nonetheless serve to expand existing knowledge concerning

the use of compression-based sentence alignment for parallel corpora. The novel hybrid sentence alignment method developed as part of this study could also be applied to other levels of the sentence and could even be used at the level of individual words, meaning that this might be adopted as a mainstream alignment method for the next generation.

In terms of practical output, the Arabic/English parallel corpus developed for use in this study is intended to demonstrate that there is a better and a cheaper method of building an Arabic/English MT corpus. This should result in parallel corpora which are unique in that they combine the characteristics of being of high quality and freely accessible. It is hoped that this will encourage other researchers and students, and even some commercial developers, to contribute to making improvements to this parallel corpus and to expanding it continually. Involvement of this kind and end user feedback would help with the further development of improved parallel corpora for the future.

1.4 Publications

Two academic papers based on this doctoral research have already been published, presenting the most important experimental results of this research project.

The first of these publications, entitled "A new parallel corpus of Arabic/English", is based on Chapter Four of this thesis and it details the development process of the extended parallel corpus. The paper was delivered at the Eighth Saudi Students Conference sponsored by the Saudi Ministry of Higher Education and King Abdullah University of Science and Technology (KAUST). The conference was held at Imperial College, London, United Kingdom. Focusing on a corpus containing some 27.8 million Arabic words and 30.8 million English terms, this co-authored paper examined how the entire dataset was corrected, by reducing noise and eliminating mistranslations with the aim of producing the most accurate Arabic/English parallel corpus currently in existence.

The second paper, entitled "A new hybrid metric for verifying parallel corpora of Arabic/English" is based on Chapters Five and Six. It was presented at the Fifth International Conference on Computer Science, Engineering and Applications (ICCSEA2015) held in Dubai City in the United Arab Emirates and was

published as part of the dblp computer science bibliography, hosted by the University of Trier (http://dblp.uni-trier.de/). This paper discussed a new metric for use with sentence alignment that was employed to verify the quality of the translation in an Arabic/English parallel corpus. The new metric is a hybrid technique combining a sentence length instrument and the compression code length. The paper also presented the experimental findings of the research, showing that this novel combined metric improves levels of accuracy and also increases the likelihood of successful recognition of both satisfactory and unsatisfactory translations.

1.5 Structure of the Thesis

In terms of its general structure, this thesis is organized in two parts. Following the general introduction, Chapters One and Two provide the theoretical context for this study by describing the unique linguistic features of Arabic and reviewing the literature relating to machine translation. The remaining chapters make up the second part of the thesis which details the various stages and processes involved in developing existing parallel Arabic/English corpora resources and transforming these into the test corpora to be used for this study which pilots a novel hybrid method for sentence alignment in parallel Arabic/English corpora.

Chapter Two focuses on the Arabic language itself and describes those specific linguistic features that distinguish it from English and which make it a particularly challenging medium for language processing.

Chapter Three presents the literature review and is designed to summarise existing knowledge and research related to Arabic statistical MT and to assess the usefulness of these previous contributions to the present study. It begins by discussing the origins and growth of MT to contextualise this current study and then reviews previous research which has focused on the issue of techniques for improving sentence alignment in parallel corpora since this represents the research foundation for this current study. The fundamentals for computing language coding for both Arabic and English are then explained and the various types of MT systems are compared in order to determine which of these systems is thought to produce the best results in terms of accuracy. The difference between various translation models including Direct Translation and Interlingua are also contrasted and their advantages and shortcomings highlighted. Finally, two key elements of

this research are discussed in detail, namely the n-gram model, originally formulated by Shannon (1948), and the data compression technique known as Prediction by Partial Matching (PPM).

Chapter Four highlights the limitations of previous parallel corpora developments and discusses some proposals concerning how these could be improved. A detailed step-by-step description is then provided of the methodology used to develop the new parallel Arabic/English corpus from scratch. The chapter records the various milestones in this development process including the filtering of the parallel corpus sources, the categorisation of data, and the examination and revision of the new test corpus.

Chapter Five elaborates how the processes of comparison, analysis and verification were carried out on the new parallel corpus to enhance its accuracy. The traditional and new methods employed to test the quality of the sentence pairs and the overall quality of the corpus are described. The former method involves the use of a distance metric based on length for sentence alignment, while the new method works by compressing code lengths. By using both these techniques, the researcher was able to determine the most appropriate thresholds for both the traditional code-length based corpus and the code-ratio one, using these to improve the accuracy of the sentence alignment. These results provided the researcher with insights into how to create a new hybrid method for sentence alignment, which is discussed in the following chapter.

Chapter Six describes the development of the novel hybrid method for ensuring sentence alignment and then discusses the experimental results. The importance and significance of the role of sentence alignment in development of parallel corpora is evaluated. This new method for sentence matching, which combines sentence length ratio (SLR) and compression code length ratio (CR) was applied to verify the accuracy of the translation between the sentence pairs. In addition, a threshold mechanism, based on whether the SLR or CR values had been exceeded, was used to filter out unsatisfactory translations.

Chapter Seven acknowledges the limitations of this research and discusses various suggestions for future developments in this area. After summarising the main points from the previous chapters, overall conclusions are drawn from the study, and the contributions which it has made are evaluated from both empirical and practical perspectives.

Chapter 2

Arabic: A uniquely challenging medium for language processing

2.1 Introduction

Before proceeding onto the Literature Review proper in Chapter Three, it is important to provide some contextual information about Arabic itself and to describe some of the specific linguistic features that distinguish it from English and which also contribute to making it a particularly challenging medium for language processing and for machine translation.[1] Indeed Soudi et al. (2007) claim it "poses some unique problems for language technology research" (p.1).

This chapter begins then by establishing the sociolinguistic situation of Arabic as a so-called diglossic language and then examines the language's geographical spread from the Arabian Peninsula as a result of the historical expansion of the Islamic Empire and the subsequent development of multiple local variants or dialects of Arabic. The chapter then describes the Arabic alphabet and writing system, and the distinctive features which it possesses. The description of the language concludes by highlighting where useful those features of a morphological and syntactic nature which can pose particular problems for machine translation when Arabic is paired with English.

[1]The first edited collection of articles focusing specifically on the challenges of machine translation from English into Arabic was published in 2012 see Soudi, A., Farghaly, A. Neumann, G. and Zbib, R. 2012. Challenges for Arabic Machine Translation Amsterdam: John Benjamins.

Arabic (العربية) forms part of what is known as the Semitic branch of the Afro-Asiatic family of languages and it functions as the official or co-official language for over 260 million native Arabic speakers, in numerous countries throughout Saharan Africa, the Middle East and the Arabian Peninsula.[2] It is one of the most spoken world languages with total users estimated at the time of writing at more than 330 million people (Soudi et al., 2007). It is also used as one of the six languages which have official recognition within the United Nations and its related organisations and functions as the official language of the Arab League. It also has a special place in the hearts of Muslims around the globe as the language of the Islamic scriptures, the Qur'ān, of salat (their obligatory daily prayers) and of their liturgical rites.

As the *Al-Jazeera* journalist Rasheed (2008) notes Arabic is also considered to be "the tie that binds [...] acting as the main pillar of Arab solidarity, national unification and Arab cultural unity". It has links to pan-Arab nationalism and political activism and is viewed as one of the key markers of Arab identity (Haeri, 2000).

2.2 The sociolinguistic situation of Arabic: Diglossia

The sociolinguistic situation of contemporary Arabic is often described as being a classic example of a linguistic phenomenon which is known as diglossia. As Saiegh-Haddad (2004) noted, this term was first introduced by the sociolinguist Charles A. Ferguson in 1959, and is applied to a linguistic situation that includes four specific features:

(a) A differentiation between the written and the oral modes;

(b) A rigid socio-functional complementarity of two separate sets of functions performed by two different linguistic codes;

(c) A rich and dominant written literary tradition;

[2]These countries are Algeria, Bahrain, Chad, Comoros, Djibouti, Egypt, Eritrea, Iraq, Israel Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen.

(d) Linguistic relatedness between the two linguistic codes: the written and the spoken (Saiegh-Haddad, 2004, p.1).

In other words, this means that two separate varieties of the same language exist but each is used for different situations or purposes, with one of the forms enjoying a higher level of sociocultural prestige than the other (Al-Batal, 1992; Haeri, 2000).

Specifically in the case of Arabic, this usually means that educated Arabs of any nationality will normally be able to use both their local or national variant of the language (the colloquial form of Arabic used for everyday interactions) and Modern Standard Arabic (MSA) which is the medium of communication used throughout all levels of the education system. Colloquial Arabic is normally referred to simply as *Ammiya* and its formal equivalent as *Fus'ha*. Habash (2012) draws a useful distinction by referring to these respectively as "the language of the heart" and the "the language of the mind". *Fus'ha* and *Ammiya* are so distinctive that some researchers argue they should be viewed as two separate languages (Ayari, 1996; Versteegh, 2001).

Fus'ha is form of Arabic which serves throughout the Arabic-speaking world as the normal medium for most written and formal spoken purposes. However it is not normally used in everyday conversation (Versteegh, 2001). *Fus'ha* has a high degree of uniformity and usually functions as the official national language in all Arab states, being used for official communications (including religious sermons and political or administrative addresses) (Al-Toma, 1969). Moreover this prestige form is used in the media, literature and academic discourse (both spoken and written) and is considered to be "the language of the educated and the intelligentsia, often used to display their knowledge" (UNDP, 2003, p.122), acting as a written lingua franca for Arabs of different nationalities.

Fus'ha is composed of a complex mix of elements: (a) Classical or Qur'ānic Arabic; (b) literary Arabic, and (c) MSA. MSA is a more recent variety of Arabic created largely thanks to the mass media forms of the press, television and radio. Some believe that the spread of MSA may be accelerated by satellite television, in particular the popularity of the Qatar-based news and current affairs channel Al-Jazeera, and the Internet (Versteegh, 2001).

In addition, all Muslims in Arabic-speaking countries will also have been exposed since their earliest years to Classical Arabic which is the variety of the language

used in the Qur'ān but there are no native speakers of Classical Arabic just as no one would claim Latin as their mother tongue. Classical Arabic is the language of the Qur'ān and thus of Islam, the second largest religion in the world. Due to this connection with Islamic scripture Arabic has been used very widely as a second language throughout the Muslim world, meaning that it has also greatly influenced other non-Arabic languages spoken in Islamic countries, including Persian and Turkish. For historical reasons, Arabic has also influenced a number of European languages, in particular Spanish which has numerous Arabic loanwords (Versteegh, 2001).

In recent years, the United Nations Development Programme (UNDP) has produced a series of Arab Human Development Reports (AHDR) which are intended to identify challenges to development in the Arab world and to propose means for addressing these. In 2003, the AHDR focused on the need for a "radical revision of education systems in Arab countries" (UNDP, 2003, p.123) in order to create a knowledge society to help overcome these challenges. The report specifically referred to the Arabic language as "the rallying point for the intellectual, spiritual, literary and social activities incarnated in [...] Arab Islamic civilisation" (UNDP, 2003, p.122). At the same time, however, the report observes that "classical Arabic is not the language of cordial, spontaneous expression, emotions, feelings and everyday communication. It is not a vehicle for discovering one's inner self or outer surroundings" (UNDP, 2003, p.121). Instead, the report stresses, this takes place through the medium of *Ammiya*, the colloquial varieties of Arabic which native speakers use to communicate in everyday life.

In theory, then, the use of MSA should ensure that regardless of the geographical origin of the written texts incorporated into an Arabic corpus the content of this should be able to be used for translation purposes in any Arabic-speaking country. Inclusion of spoken Arabic, however, would pose problems of comprehension since many of the dialects are mutually unintelligible and if educated Arabs from different countries converse with each other, they will sometimes switch to MSA for the sake of communication.

2.3 The Geographical Spread of Arabic

As previously noted, since Arabic is used across vast geographical areas "from Iraq in the north to Somalia in the south, Bahrain in the east to the western shores of Mauritania" (Rasheed, 2008), it has split into multiple regional variants or dialects. These may differ widely along geographical, religious and socio-economic lines from one Arab country to another and may even differ between communities within the same country (Holes, 1995). As geographical distance between speakers increases mutual intelligibility decreases. *Ammiya*, the local spoken variant of Arabic, has often been influenced by the characteristics of other local languages and by foreign words, particularly those of former colonial powers. For example, the spoken form of Arabic in the Maghreb is known asas *darija* (الدارجة literally meaning dialect) and has integrated many Berber, French and Spanish words whilst Libyan Arabic still contains many loan words from Italian (Frías Conde, 2000).

Figure 2.1 shows the multiple varieties of Arabic and their development from Classical or Qur'ānic Arabic. The Western Arabic grouping covers all the countries of North Africa (Libya, Morocco, Tunisia, and Algeria) except for Egypt and Sudan. A distinction can be drawn between Maghrebi and Saharan forms of Arabic because the underlying influences are different and the former often prove incomprehensible for the rest of the Arabophone world. Maltese, with some 330,000 speakers, evolved more independently due to its cultural and religious differences with the Arab world and is written in Latin script. Maltese is considered to be a separate Semitic language but is clearly marked by Italian elements. Since the Maltese are Christians, they do not have exposure to the Classical Arabic of the Qur'ān (Frías Conde, 2000).

The Eastern Arabic area includes the rest of the Arabic-speaking world, from Egypt and Sudan to the Asiatic zone of the Middle East and the Arabian Peninsula. The Arabic dialects of this region have remained closer to Classical Arabic because they have undergone less change. Egyptian Arabic and Levantine Arabic (covering the territories of Syria, Palestine, Jordan and Lebanon) are almost identical, with virtually no comprehension problems between speakers. Although Iraqi Arabic is close to Egyptian and Levantine dialects, it also shares some features with Peninsula Arabic, although Yemeni has its own characteristics (Versteegh, 2001).

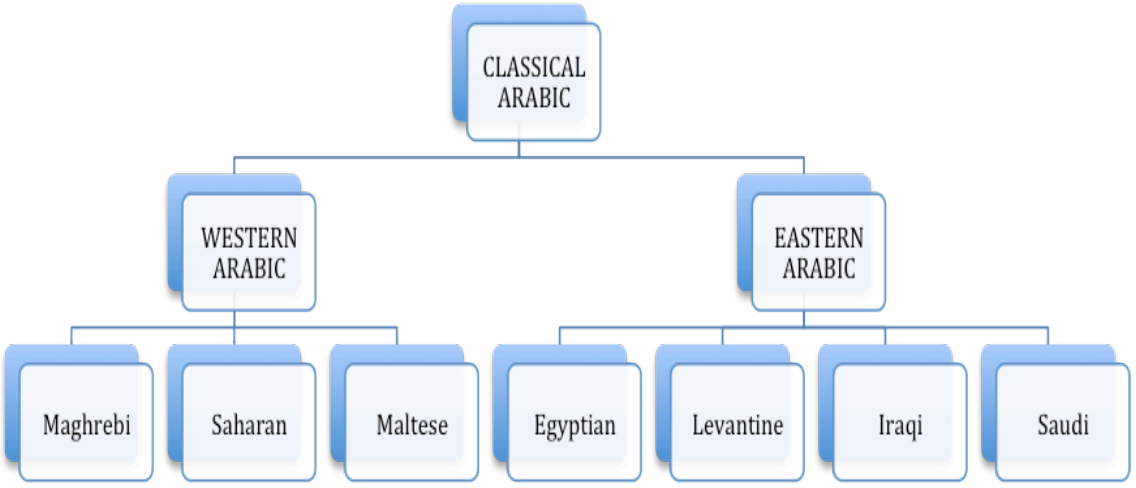

FIGURE 2.1: Varieties of Arabic Source: Adapted from a model devised by Frías Conde (2000)

Egyptian Arabic has the greatest number of speakers (52 million) and is still generally considered to be the most prestigious dialect in the Arabic-speaking world thanks largely to the cultural importance which this country enjoyed historically. The dialect of the capital, Cairene Arabic or *Masri*, is understood by virtually all Arabic speakers.

This illustrates that the utility of a spoken Arabic corpus tends to be limited to a specific geographical area or region since not only do the varieties has their own distinctive phonetic and phonological features but also vary to a great degree in terms of the vocabulary used by speakers, their grammar and syntax.

2.4 The Arabic Alphabet and Writing System[3]

The Arabic alphabet consists of 28-letters which are used to represent the 34 phonemes of consonants and long vowels (/a:/ /i:/ /u:/) necessary for the written codification of the language (see Figure 2.2). As previously noted, Arabic belongs to the group of Semitic languages, which evolved from Nabataean and Aramaic and thus, like them, it is written horizontally from right to left, the opposite way to English (Elbeheri et al., 2006).[4] Also unlike English script, Arabic has no equivalent of upper- or lower-case letters. A total of 17 characters are used to form the Arabic alphabet together with dot-like diacritics (one, two or three in number) which are known as *i'jām* (اعجام). These diacritics, thus, play a vital role in distinguishing between letters, as seen in the letters /n/ noon/, /b/ beh, /t/ teh and /th/ theh as shown in Figure 2.2.

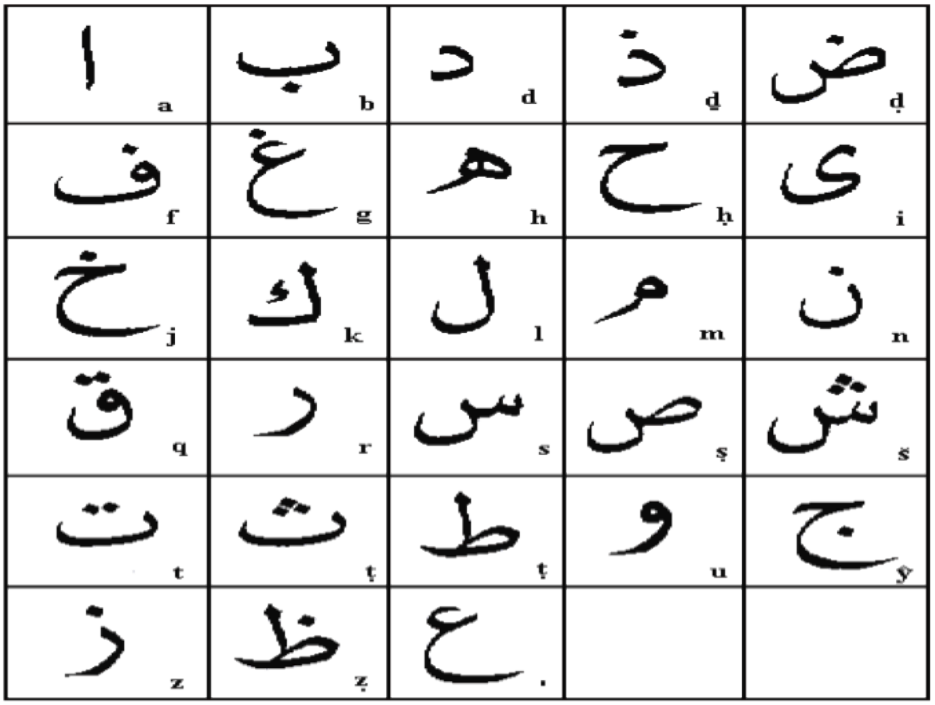

FIGURE 2.2: Arabic Alphabet with Transliteration Source: Frías Conde (2000)

In written Arabic, there is no set of graphemes that are used to represent sounds of the short vowels /a/, /i/ and /u/. Instead a set of diacritics (known as

[3]Arabic script is also used with modifications in Farsi, Urdu and Pashto.

[4]Numbers however are written left to right and enumeration systems may vary from region to region (Habash, 2012).

Harakāt حركات) are used to represent these sounds and they appear above or below the character, as seen in Figure 2.3.[5]

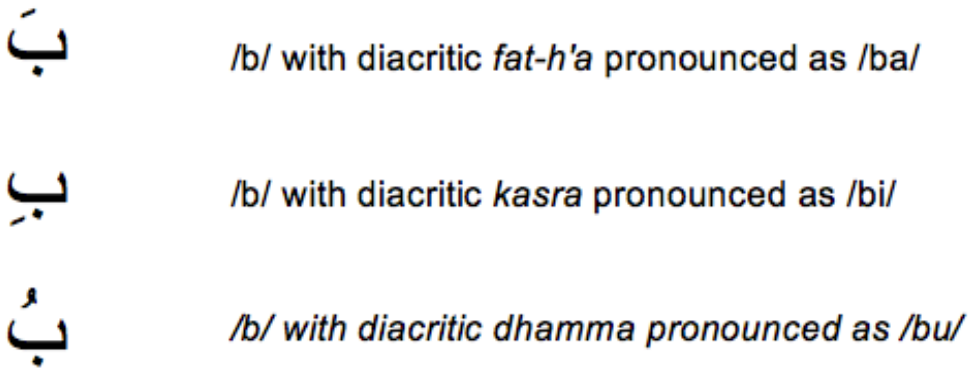

FIGURE 2.3: The Short Vowel Diacritics

As Beland (2001) explains, these *Harakāt* markings are only used in certain circumstances including on religious texts (to help with interpretation) and they are also used on teaching materials and schoolbooks used by children at primary school in the Arabophone world and also textbooks designed for foreign language learners. However as soon as children are past the initial stage of learning to read and write, the marks are gradually taken away. Short vowels are largely unmarked in the majority of texts experienced by readers after the initial years of education. Without these markers, many Arabic words are homographic, meaning that their written forms appear identical, and this makes them semantically and phonologically ambiguous when they are presented out of context, as shown in the examples presented in Table 2.1. Diacritics are thus also used in those cases where ambiguity may distort understanding.

Linguists refer to these two types of orthography in Semitic languages like Arabic as being "shallow" or "deep". This orthography is called "shallow", vocalised or vowelised when diacritical marks or *Harakāt* are used in the text to represent short vowels and "deep", non-vocalised or non-vowelised when these are not added (Versteegh, 2001).

Arabic script is cursive, meaning that many, though not all, of its letters are connected to each other by means of ligatures, like handwritten English (Elbeheri et al., 2006). However unlike English in which each space in a line of written

[5]A number of other diacritics are also used to indicate the pronunciation of consonants, a phenomenon known as germination or nunation (Versteegh, 2001).

Shallow Arabic	Deep Arabic	Transliteration	Meaning
عَلِمَ	علم	a'alima	He knew
عُلِمَ	علم	a'ulima	It has been known
عَلَّمَ	علم	a'allama	He taught
عِلْم	علم	a'ilm	Knowledge

TABLE 2.1: Examples of Shallow and deep Arabic Orthography

script is used to indicate different words because all the letters within one word always link to each other, in Arabic this is not necessarily the case. Only 22 of the 28 letters which make up the Arabic alphabet can join to both the letters which precede them and the letters which follow; the other six can only be joined to the letter which precedes them. Thus when these six Arabic characters appear in the text one or more spaces may be created within a word depending on how many of these letters are used within the word. These two factors, namely the cursive nature of Arabic writing and the word-internal visual spacing, may make it difficult for inexperienced readers to find the boundaries between words in Arabic (Elbeheri et al., 2006).

In addition, Arabic letters vary their graphic form according to their position they occupy within a word and on the surrounding characters, with some having four separate forms. These changes depend on whether the letter comes at the beginning (initial) of the word, in the middle (medial) or at the end (final).

- Initial means that the character can be linked only to the character which follows it.
- Middle/medial means that the character can be linked both to the character that precedes it and the character that follows it
- Final means that the character can be linked only to the character which precedes it.

When shown as single letters of the alphabet, letters are known as isolated. As Figure 2.2 shows, these letters can take on very distinctive forms.

Position in word	Isolated	Initial	Medial	Final
Letter forms	غ	غـ	ـغـ	ـغ

TABLE 2.2: The four written forms of the Arabic Letter *GHAYN*

One problem worth mentioning with regard to parallel corpora is highlighted by Habash (2012), namely, transliteration, which is the conversion of one language script into another, and is typically required with the use of proper names. Habash illustrates the problem using one of the examples which caused most problems in terms of transliteration from Arabic to English, the name of former Libyan leader Moammar Gaddafi, and notes that the Library of Congress lists 72 different spellings for this name. Clearly this type of variation makes it more difficult to determine if two sentences in Arabic and English are suitable translations of each other.

2.5 Arabic morphology

Elbeheri et al. (2006) note that another feature which Arabic shares with other languages in the Semitic group can be found in the way it forms words by using consonantal roots (which mostly consist of three letters) and then inserting different patterns of sets of vowels between these, as can be seen in the example in Figure 2.4 which uses the root /k-t-b/ /ك-ت-ب/

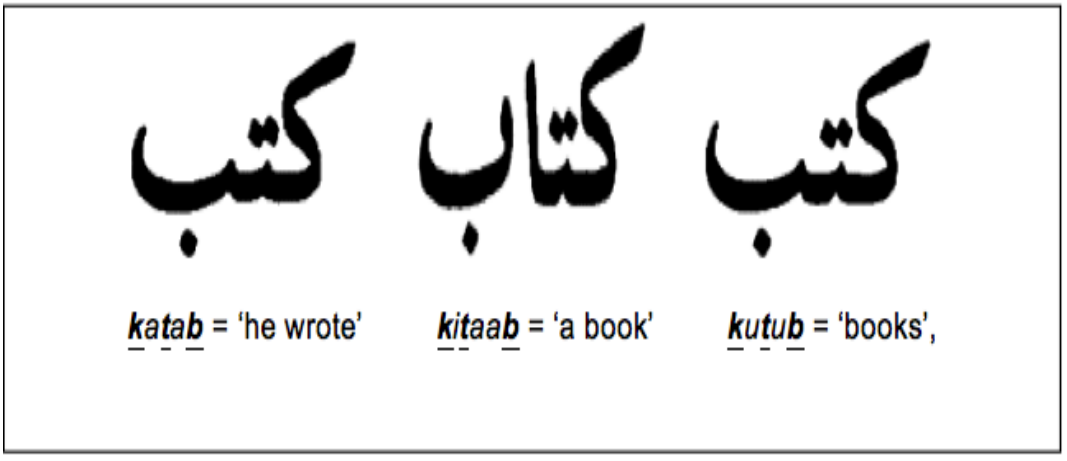

FIGURE 2.4: An illustration of a consonantal root

In this case the root K-T-B is associated with the notion of writing. Elbeheri et al. (2006) explain that "Variations in shade of meaning are obtained first by varying the vocalisation of the simple root, and second, by the use of prefixes, suffix and infixes" (Elbeheri et al., 2006: 144). Each variation in this pattern creates a different meaning. However, the same order is maintained by these letters when they form any word which came from the same root, for example **maktab** means 'office'; **maktaba** is 'library' and its plural 'libraries', **maktabat**. Thus,

as Elbeheri et al. (2006) emphasise in Arabic these consonantal roots indicate the conceptual content of a word (as in the example of K-T-B being linked to book) but the pattern of vowels which with they are used indicate the grammatical function it is intended to fulfil. Every variation in this pattern can create a different meaning referring to tense of a verb, plural or singular, etc. In general, in order to understand Arabic words it is necessary to look at the consonantal root which they contain and the pattern of vowels. This means that although Arabic is a complex language, it follows predictable patterns due to the consonantal root system which underpins most words.

Arabic is referred to by linguists as a highly agglutinative language (Versteegh, 2001). Elbeheri et al. (2006) explain that this means that prefixes and suffixes can be placed before or after words to indicate information about many different grammatical features including the tense of verbs, whether an adjective is masculine or feminine, whether a noun is definite or indefinite. In addition, pronouns of various types can also be placed within words or attached to them. This also applies to certain conjunctions and prepositions. Soudi et al. (2007) observe that this can create a real difficulty for the purposes of machine translation and alignment because it is possible for one single word in Arabic to correspond to a whole sentence in English because of the numerous additions that can be made to the word root.

Habash (2012) usefully summarised some of the key differences in terms of orthography, morphology and syntax which are likely to create problems when dealing with Arabic/English machine translation in the following way:

AREA OF CONCERN	**ARABIC**	**ENGLISH**
Orthographic ambiguity	More	Less
Orthographic inconsistency	More	Less
Morphological inflections	More	Less
Morpho-syntactic complexity	More	Less
Word order freedom	More	Less

Table 2.3: Arabic/English Linguistic Differences

2.6 Conclusion

The fact that Arabic is spoken in geopolitically strategic areas of the world and that there are strong business links between the West and the Middle East in particular have highlighted the need to develop fast and accurate Arabic/English machine translation and these efforts not surprisingly have been led by the United States Department of Defense and leading industrial companies such as Google and IBM (Soudi et al., 2007). However as this chapter has shown, natural language processing can be a particularly complex process, posing multiple challenges for 'Arabic' with its multiple variants or dialectal forms. A further difficulty arises when attempting to produce parallel corpora for this language pair of Arabic and English due to the multiple linguistic differences which exist between them. These challenges and their solutions are examined in the second part of thesis.

Chapter 3

Literature Review

3.1 Introduction

There is no doubt that the quality of a machine translation system depends on both the reliability and the rationality of the algorithms behind it. In order to improve the quality of translation in terms of its accuracy and also the speed at which translating tasks can be carried out, various types of machine translation systems have been developed and these have been divided into different classifications on the basis of the translation algorithms which they employ(Habash and Sadat, 2006).

Before embarking on the more detailed discussion of the technical aspects of the project which forms the focus of this study, this chapter presents a review of some of the concepts that are considered to be fundamental to the process of machine translation, at the same time defining and clarifying the key terms which will be employed throughout this thesis and mapping the contours of this area as an academic field.

This chapter, then, is organised as follows. The chapter continues by briefly tracing the history of Machine Translation between natural languages, charting the changing fortunes of this field since its earliest origins. The focus then shifts in order to provide a general introductory overview of machine translation systems. After discussing some of the best-known translation models such as Direct Translation and Interlingua, statistical machine translation systems are examined in more detail since this it is argued here that this method is best placed to offer

the most feasible solution for multi-respondent translation between two natural languages.

The main emphasis in this context is on the pioneering work of Claude Shannon (1948), often referred to as "the father of information theory",[1] and Warren Weaver (1949), who popularized the American mathematician's ideas. Their Shannon-Weaver model of communication and the associated concepts of the noisy channel and n-gram model were explored here also. The later extension of these models by Brown et al. (1990), researchers at the IBM T. J. Watson Research Centre, is also considered.

The second part of the chapter examines a key element in machine translation, the corpus, considering the different forms that this may take and their role in successful machine translation. This part of the literature review also considers previous research in the area of sentence alignment, a essential means of improving the quality of parallel corpora.

The chapter ends by examining the various ways of encoding Arabic within a computer-based system.

3.2 Early developments in Machine Translation

The history of translation between two natural languages can be traced back to the beginning of human culture. Its major mission is to expand the informativeness of one language, decrease misunderstandings in dialogue, and also contribute to the growth of cultures (Nida, 1975). Language translation has been seen as a valuable Social Science-oriented industry aimed at helping people to develop international relationships since the advent of ocean commerce in the 1780s which was driven by colonial interests.

At the end of the 1930s, the French engineer Artsouni (Bar-Hillel, 1964) first proposed that it might be possible to translate one natural language into another by means of a process which involved the coding of text and then aligning sentences using a dictionary-based computer database. However, it was actually Shannon and Weaver (1949) who coined the term 'machine translation' to describe this

[1]See for example his obituary penned by I. James (2009) which appeared in *Biographical Memoirs of Fellows of the Royal Society*, 55 (257-265).

process in their publication, "translation memo". This subsequently led to a flood of publications on machine translation after World War II.

Though machine translation nowadays is a successful pioneer of the most innovative computing technologies, there was a three-decade 'Dark Ages' period in this area of research from 1949 to 1974, when the Automatic Language Processing Advisory Committee (ALPAC) in the U.S. denied that it would be possible to realising and implementing machine translation. As a negative consequence, the main research institutions including those in the U.S., the U.K., France and even the Soviet Union, all stopped their research on translation. However, some milestones were achieved quickly after that period.

Private research institutions and organisations such as Xerox, Princeton University, and Columbia University later developed prototypes based on the development of the Internet (Klavans and Gonzalo, 2006). The notion of a corpus, more specifically a parallel corpus, became more and more popular in machine translation research (Koehn, 2009). Various products were developed by following the traditional translation rules that compare the length of translation units, whereas empirical studies argued that the translation process could be better compared by how much entropy value was included in a pair of analysis units (Koehn, 2009).

3.3 The need for Machine Translation

Language is not simply a medium of communication or the expression of one's views but is the very wheel of the chariot of culture and civilisation, so to speak. Currently, quick and easy access to accurate and timely information is considered essential to so many areas of life and when much of this is information is available in English, the international language of business, economics, science and technology, the importance of translation has become ever greater. As our ability to process, access and store large quantities of data has advanced exponentially in recent years, the long-awaited dream of machine translation appears to be getting closer to fulfilment.

According to Hutchins (1986, 2001), the necessity for an automated system for translating language has existed for a long time. Machine Translation, also known as computer aided translation, is a computer technique based program that allows one natural language to be transform into one or more different languages,

either on verbal and written linguistic contents (IngilizceTurkce.Gen.Tr., 1998). Machine Translation occupies a particularly important place in the modern world, the foundation of which is information (Brown et al., 1993).

The World Wide Web has enabled users of different languages to access information despite the language barrier. With the expansion of computer-based translation activities, processes that enable the production of text and documents in multilingual and bilingual environments are becoming popular. Furthermore, the possibility of Machine Translation being considered a significant element for facilitating international communication in this information age is emerging (Yamabana, 2006).

As previously noted, the idea of using automated systems for translation is far from new. Yet, a number of factors such as hardware limitations and slow programming languages have restrained the development of Machine Translation, and researchers have relied heavily on the dictionary-based approach and the application of statistical methods. Weaver (1949) proposed the incorporation of computers into the process of translation in March 1947. In 1954, the first successful operational translation system was demonstrated (Homiedan, 1997) and was designed by IBM in joint collaboration with Georgetown University. Since then, a number of theories and models have been explored with the aim of devising a fully automatic system that could generate high-quality results.

3.4 Machine Translation models

Machine Translation systems are classified according to a number of different criteria: input text, applications used, the level of analysis employed and the type of technology used. With regards to the level of analysis criterion, translation models are generally classified into three categories: Direct Translation Models, Semantic Transfer Models and Interlingua Models.

Figure 3.1 illustrates the different levels of analysis in a machine translation system and how these relate to the three types of models mentioned above.

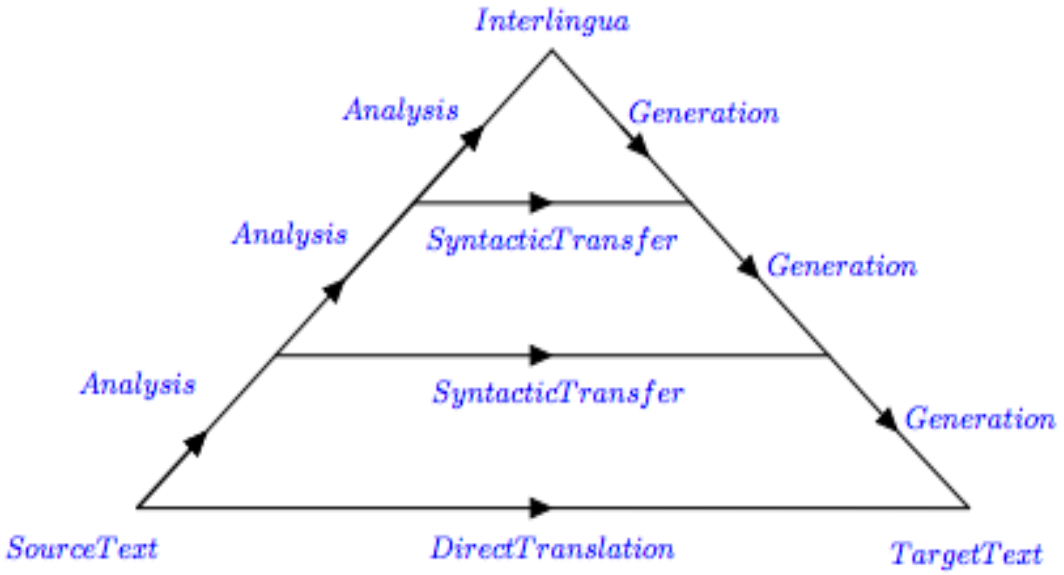

FIGURE 3.1: Different Levels of Analysis in a Machine Translation System (Vauquois, 1968)

3.4.1 Direct Translation Models

At the beginning of Machine Translation development, users had to use direct translation to translate their files. In this, the least sophisticated approach to machine translation, word-by-word translation is carried out using a bilingual dictionary. This approach poses significant challenges for language pairings that are different in nature and also for the obvious reason that a single word may have many translations, depending on the context in which it is being used. At that time, it was slower and resulted in less effective translation (Hutchins and Somers, 1992).

Direct Translation Systems perform the translation directly and dissect the entire translation process into different stages with each stage dealing with a particular task. For example, one stage deals with morphological analysis, another with prepositions, and another with subject and verb arrangements, and so forth. These models rely on ad hoc rules that are hand coded and do not rely heavily on linguistic knowledge. Direct translation is suitable for translating one lexicon of language data and its equivalents in the other target language (Hutchins and Somers, 1992), making them appropriate for texts that are formulaic and not complex in nature. Hutchins and Somers (1992) gave the example of Xerox-controlled Multinational Customised English used for writing technical instruction manuals which make use of highly specific vocabulary and sentence structures.

3.4.2 Syntactic Transfer Models

Scholars next suggested that instead of this direct process, the original data should be analysed and then translated into the other language on the basis of different syntactic transferring algorithms. Syntactic Transfer models are based on the use of transformer architectures to analyse the sentences of the source language syntactically. This analysis is done by assigning part of speech tags to each word in the sentence and then by forming a parse tree of the relationships between these parts. The next stage involves the generation of a parse tree for the target language. By introducing various methods of data analysis, alignment and syntactic transferring, the accuracy and translating speed of MT began to significantly improve (Imamura et al., 2004).

However, this model has frequently been criticised for generating strange literal translations. For example, Figure 3.2 illustrates how a syntactic transfer model is used to translate between English and German (Durrett et al., 2012). Since German has a different sentence structure and some vocabulary can have negative + positive characteristics, the word "demand" in English might be translated as 'verlangen' or although here is not enough information to support this judgement. Thus, a possible result of the Genrman-English MT would make the German sentence in a wrong order for its English audience. The syntactic transfer model focuses on the collocated terms of 'Gewerkschaften' and 'Verzicht', the translation is confirmed as 'demand' = 'verlangen'.

3.4.3 Semantic Transfer Models

Semantic Transfer Models do not use language-based representations but semantic-based representations and are considered to improve the accuracy of translation. Although the main basis of these models is semantics, they are heavily focused on the languages being translated. Relationships arise because of the specific grammatical differences existing between the two languages (Abb et al., 1996).

The main advantages of using a semantic transfer model of machine translation are as follows:

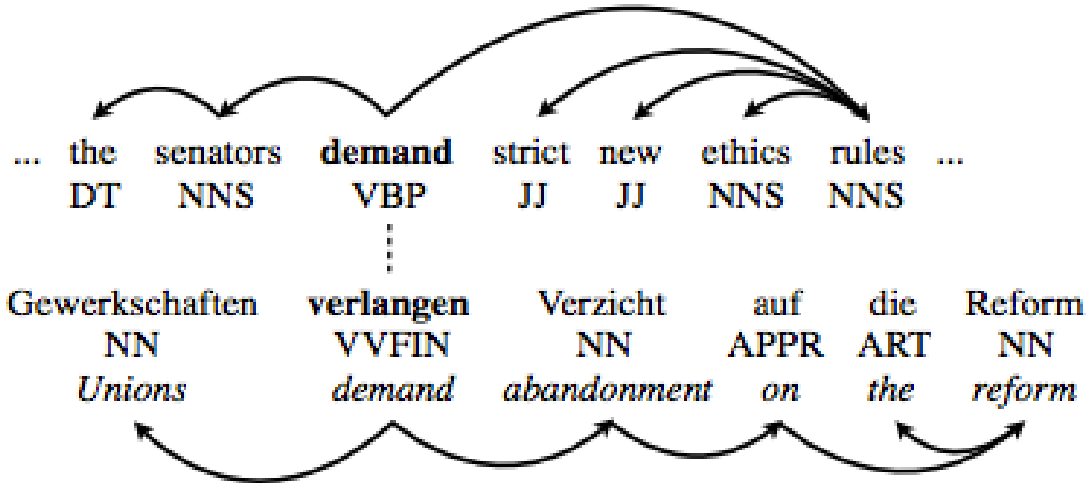

FIGURE 3.2: An example of the Syntactic Transfer Model for Sentences in English and German (Durrett et al., 2012).

- Machine translation is often used to translate text between two languages which, more often than not, have morphological as well as syntactic differences. Using the semantic transfer concept, it is easy to find an abstraction for these.
- Divergences of various kinds are taken care of using semantic transfer methodology.
- These models allow for a better modularity as compared to various other models such as the Interlingua model.

Dorna and Emele (1996) used the semantic transfer concept to develop the Verbmobil system, a major project in the machine translation area which included some 100 research personnel and nearly three dozen institutes. However, the development of a multi-linguistic Machine Translation system proceeded very slowly until the Interlingua model was suggested.

3.4.4 Interlingua Models

In basic terms, when it is difficult to translate one language directly into another one the Interlingua model converts the source text to be translated into an intermediate abstract Interlingua representation common to all languages. This

abstraction is then used for reconversion into the second language. This representation can be used for translating one language into any number of other languages. The advantage offered by these models is that the number of transfer systems required between languages is greatly reduced. However, formulating an abstract language capable of representing all concepts in all languages is not easy. This system also requires the abolition of ambiguity. These models are considered to have the potential to liberate the translation system from modelling the source text tightly (Popescu-Belis et al., 2008).

The Interlingua concept was first suggested by Vauquois (1968) who maintained that using this Interlingua, the contexts of a file could be read abstractly by the machine through understanding the characteristics or nature of the language itself and understanding the semantic meaning of each independent sentences. Since Interlingua initially runs a comprehensive analysis process of the original file data, it allows the machine to locate each meaning of the context and enables it to generate the context in a range of sophisticated languages (Hutchins and Somers, 1992). However, the Interlingua design requires the development of a highly complex MT system or even a Machine Translation prototype, so a limited number of Interlingua machine translation systems have been built and used for commercial purposes (Vauquois, 1968).

Although this may seem to be a more complex exercise than using a single translation, its usefulness becomes obvious in multilingual environments, such as the European Union or the United Nations. For example, if there is a system involving multiple languages such as English, German, Russian, French, Japanese and so forth, the intermediate abstraction format acts as a common factor which can be used to connect any two languages by translating them into the appropriate format.

3.4.5 Statistical Machine Translation

3.4.5.1 Shannon's 1948 noisy communication channel model

The idea of relating statistics to automated language translation dates back to Weaver (1949) who considered the viability of Machine Translation in his paper, and based on his personal experience and Shannon's 1948 theories about communication, he suggested that translation could be viewed as a process that involves decoding.

Natural languages are not precise, and deriving the meaning of a word out of context can be difficult, so he argued that context should be considered when translating. The amount of context to be considered in order to remove ambiguities would depend on the text and the part of speech. This general principle serves as a base on which modern translation systems should be built, otherwise the entire purpose of translation is defeated if the meaning is not relevant to the context.

According to Calderbank and Sloane (2001), Claude Shannon's 1948 paper entitled "A Mathematical Theory of Communication" paved the way for deep-space communication, wireless phones, the Internet, data networks, modems and hard drives, proposing that a noiseless channel could be created in an environment where all possible messages are a definite set. The noisy channel model is also widely used in statistical machine translation studies. Echihabi and Marcu (2003) recommend using this model to improve the quality of MT translating and Chiang (2005) and Calderbank and Sloane (2001) introduce it into phrase-based level MT studies. This would put communication on a solid mathematical platform where communication using different languages could be set up from a global perspective, so that people from different nationalities and geographical regions can communicate securely, clearly and without any ambiguities. Figure 3.3 illustrates how Shannon's noisy communication channel model is adopted in this research. Noise could be found in either inside a communication channel or the external environment and any noise will effect on the quality of MT by giving an unsatisfied translation outcome.

FIGURE 3.3: The MT process in the Noisy Communication Channel.

3.4.5.2 Brown et al.'s 1990 Noisy Channel Model

According to Tiedemann (2003), Statistical Machine Translation is a systematic methodology to generate translations using statistical models. Brown et al. (1990) put forward the idea of a purely statistical approach to Machine Translation. One particular sentence may be translated in many different ways. New sentences are formed after combining the elements found in the sentences generated previously. The aim is to recover these elements to create the translations for new sentences. The elements to be reused are identified using statistical analysis conducted on a large parallel corpus. A probability is assigned to every alignment. The system then utilises the various probabilities observed from the parallel corpus to generate an output, which is the most probable translation for the input sentence. This effective method for MT, devised by Brown et al. (1990), is called the Noisy Channel Model.

Theoretically, the most probable sentence is derived by maximising:

$$P(S/T) \tag{3.1}$$

where S represents the source sentence, and T represents the target sentence. Applying Bayes' Theorem enables the determination of the most probable translation in terms of fluency and faithfulness:

$$P(S/T) = \frac{P(T/S)P(S)}{P(T)}. \tag{3.2}$$

Fluency is measured in terms of a priori probabilities for word sequences in the source language $P(S)$, and faithfulness is measured by the conditional probability of

$$P(T/S). \tag{3.3}$$

The denominator is constant for all translations sought for a given sentence in the target language T, dispensing the need to use it explicitly in the calculations. The fundamental equation for Machine Translation is:

$$P(S/T) = P(T/S)P(S). \tag{3.4}$$

Statistical Machine Translation for the language model requires the modelling of source and target languages, whilst the translation model requires the modelling of the relationship between the two languages.

The statistical evaluation of the fluency of a sentence can be evaluated by calculating the conditional probabilities for all words. The previous words in the sentence can be represented as follows:

$$P(w_1 w_2 .... w_3) = P(w_1) P(w_2 | w_1 w_2 .... w_n - 1). \tag{3.5}$$

The Markov assumption requires that the dependency of a word on the previous word decreases as the distance between them increases, serving as the basis of many statistical natural language processing applications. N-gram probability is the conditional probability of a word based on n number of prior words. It can be used to approximate the probability of a sentence.

Approximation plays an important role in Machine Translation. A corpus containing a set of all possible sentences in a language is not possible but it can include enough features of the lexicon to show the main dependencies. The early Statistical Machine Translation systems used the product of trigram word probabilities to determine the probability of a sentence, which is represented as follows (Katz, 1987):

$$\widehat{P}(w_i | w_{i-2} w_{i-1}) = \begin{cases} P(w_i | w_{i-2} w_{i-1}), & \text{if } C(w_{i-2} w_{i-1} w_i) > 0. \\ P(w_i | w_{i-1}), & \text{if } C(w_{i-2} w_{i-1} w_i) = 0 \textit{ and } C(w_{i-1} w_i) > 0. \\ P(w_i), & \text{otherwise.} \end{cases} \tag{3.6}$$

3.4.6 N-gram Models

N-gram models are quite popular in machine translation and basically consist of a continuing sequence of n number of items from the text under consideration for machine translation, hence the name n-gram model. The value of n can start from one and go upwards and similarly models with an n value of 1 are called unigram models, those with a value of 2 are bigram models and so forth.

Empirical evidence is available to support N-gram modelling. Shannon (1948) highlighted that familiarity with a language enables humans to fill in missing

or incorrect letters, an ability which Manin (1996) referred to as 'tentative word guess'. Brown et al. (1993) successfully conducted experiments by using the trigram language model to to test the performance and accuracy of the translation between English and French. In total, 84% of the output generated was coherent, and 63% of the output was in the same order as the original.

By using higher orders of N-grams, the predictive accuracy of the language can be certainly improved to a significant extent. This is true in theory but in actual practice, this may not be the case as a higher N-gram order would mean increasing the sparse data problem too.

The overall procedure for enhancement of the N-gram factor is something similar to the more familiar process of generation of vocabulary. The output of the N-gram generation is then fed to the language generation as the next step.

3.5 The Corpus: A key element in machine translation

3.5.1 Defining terms

In its broadest and most general sense, the term 'corpus' (from a Latin word literally meaning 'body') has long been used to refer to "A collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject" (OED). However, with the birth of modern corpus linguistics in the late 1960s, signalled by the publication of Computational Analysis of Present-Day American English (1967) which was co-authored by Henry Kucera and W. Nelson Francis, the word 'corpus' acquired a much more specific and narrowly defined sense (Bowker and Pearson, 2002).

According to Xiao (2008), in its more specialised meaning, a corpus does not consist of a randomly assembled group of texts nor is it merely an archive. Instead, Xiao's definition highlights the fact that a corpus (or perhaps it sould be more accurate to say, a good corpus) should exhibit four key features:

A corpus is a collection of (1) machine-readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety.

Many of the definitions which have been produced draw on a similar set of criteria but choose to place slightly differing emphasis on the various dimensions of the corpus. It is worth briefly exploring some aspects of these four criteria highlighted by Xiao.

The phrase 'machine-readable' can be replaced by words or phrases such as 'computerized' (Tognini-Bonelli, 2001) or 'in electronic form' (Bowker and Pearson, 2002), with the latter pointing to the fact that nowadays large quantities of digital text, in the form of newspapers and myriad other publications, is now freely available from online sources in a way that would have been unthinkable even just a decade ago. These are now "amenable to automatic or semi-automatic processing or analysis" (Tognini-Bonelli, 2001, p.2). However, Arabic content still lags far behind that for English according to Nielsen (online).

Xiao's next point regarding the need for authentic texts is an interesting own given the diglossic situation of Arabic, since in one sense, as the previous chapter suggested, many Arabic speakers might consider Modern Standard Arabic less 'authentic' in terms of self-expression than their own variant of this language. The fact that a corpus may contain transcripts of spoken data also raises the issue of the dialectal variants of this language and the impact that this may have on a corpus which is to be used as part of a machine translation system.

This point can be linked to the next two criteria concerning sampling and representativeness. Clearly, for the purposes of corpus linguistics and for the field of lexicography in which Xiao himself works, these are crucial factors. Although other authors often fail to use the specific words 'sample' and/or 'representative' in relation to the content of the corpus, they too indicate similar ideas with the use of the word 'principled' (see, for example, Sinclair, 1991; Biber et al., 1998) or by references to the need for 'explicit criteria'. Thus, Tognini-Bonelli's 2001 definition of corpus stresses that the texts should be "selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language".

Interestingly, in his guide to good practice in developing linguistic corpora, Wynne (2005) takes issue withTognini-Bonelli's definition, arguing that in some highly specialized industries and fields such as biomedical text corpora, these criteria may not necessarily be appropriate or need to be strictly applied.

Perhaps due to his insistence on sampling and representativeness as criteria, Xiao is not prescriptive in his definition concerning the size of a corpus and does not specifically include this as a criterion.

For the purposes of this current study, focusing on corpora in the context of machine translation, the key point to be taken from Xiao's exhaustive examination of the defining characteristics of a corpus is that close attention should be paid when describing its content and the means by which this material was chosen with the aim of ensuring this description is as accurate as possible and that the criteria used when choosing this material is as transparent as possible. This will allow judgements to be made by others, if necessary, on the representativeness of the material in terms of the quantity of items and their characteristics, and on the suitability of the corpus for the purpose to which it has been put.

In the case of this study, it makes sense to use a definition of 'corpus' which approaches this question from a functional perspective. Therefore, this research will use the definition suggested by Bowker and Pearson (2002, p.9) which is as follows:

A large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria [...] which can serve as a basis for linguistic analysis and description.

In the context of machine translation, the overwhelming advantage of using a corpus is that it is expandable, meaning that extra language material can be constantly added to it as a means of keeping it up-to-date and, perhaps more importantly, of improving its accuracy. For example, if the machine detects that a piece of a a translation task contains more physical terms or concepts, when it faces a corresponding sentence, it can prioritise picking the most relevant results in one physical category, over a less likely one in another. However, the development of a corpus usually requires many professional linguists to test and improve the quality of translation. This tends to raise the cost in both finance and labour terms. This also tends to prevent the development of MT because many academic researchers and low-budget users cannot afford the existing commercial MT systems. Therefore, one of the research aims of this thesis is to improve an existing parallel corpus for Arabic and English and extend this so that it can be used for the purposes of low-cost translation.

3.5.2 Types of Corpora

In his discussion of well-known and influential corpora, Xiao (2008) summarised the existing types of corpora in a comprehensive list which also distinguished the various characteristics of these different corpora. Before examining in more depth the characteristics of the parallel corpora which is the type of corpus which forms the basis of this research, a number other types of corpora and their distinguishing features identified by Xiao (2008) are briefly discussed here, together with relevant literature.

The first of the corpora identified by Xiao (2008) is the so-called **treebank**, a term coined by Geoffrey Leech in the 1980s (Kennedy, 2014). This is also known as a parsed corpus because it consists of naturally occurring text that has been annotated with structured mark-tags or part-of-speech tags and can be used to show syntactic and semantic information about a language. Parsing breaks down the linear structure of a sentence in a natural language into a 'tree'. Similar to HTML language, using treebanks in the machine translation process helps to maintain the natural structure of the two corresponding languages. It thus offers a universal analysis and testing platform which is based on statistics (Kennedy, 2014). In machine translation the introduction of the treebank was a milestone in that it highlighted the correlation between the sentence and phrase in each sentence. Figure 3.4 shows an example of a treebank taken from the Penn Treebank Project which is based in the Computer and Information Science Departure at the University of Pennsylvania. The original sentence in this example was "Battle-tested industrial managers here always buck up nervous newcomers with the tale of the first of their countrymen to visit Mexico, a boatload of warriors blown ashore 375 years ago". This was marked with a series of tags such as NP, VP and PP to analyse it in a tree shape (Kennedy, 2014).

Another example is the Qur'anic Arabic Dependency Treebank (QADT) (`http://corpus.quran.com/`) which is an attempt to map out the entire grammar of the Qur'an by linking Arabic words through dependencies. This annotated linguistic resource shows the Arabia grammar, syntax and morphology for every word in the Qur'an.

The term **'historical corpus'** (also known as a diachronic corpus) refers to a corpus which has been assembled for the purposes of studying language change over time, allowing researchers to chart the evolution of particular linguistic items

```
( (S
   (NP Battle-tested industrial managers
       here)
    always
   (VP buck
       up
       (NP nervous newcomers)
       (PP with
           (NP the tale
               (PP of
                   (NP (NP the
                           (ADJP first
                                 (PP of
                                     (NP their countrymen)))
                           (S (NP *)
                              to
                              (VP visit
                                  (NP Mexico))))
                       ,
                       (NP (NP a boatload
                               (PP of
                                   (NP (NP warriors)
                                       (VP-1 blown
                                           ashore
                                           (ADVP (NP 375 years)
                                                 ago)))))
                       (VP-1 *pseudo-attach*)))))))
  .)
```

FIGURE 3.4: The Penn Treebank Project (Kennedy, 2014).

within a language (Evans, 2007). Davies (2010) comparative discussion of two Spanish and one Portuguese historical corpora makes the important point that in any corpus "without an adequate architecture and interface, [...] data is in essence 'trapped', with little if any way of getting the data out" (p. 139).

Thirdly, **learner corpora** as the name suggest are intended to be used primarily by language researchers, teachers and producers of language-learning materials. These corpora usually consist of genuine second-language textual data from students studying a language which can be used for a variety of purposes (Granger, 2002). For example, Cambridge University collects all the textual data produced by its international students during its pre-sessional language courses and during the English Learning Test (ELT). The data gathered in the specially designed Cambridge Learner Corpus is then used to analyse the types of errors made by international English users in order to compare how the words differ from the usage of a native speaker (Nicholls, 2003). Rather than simply using statistical

summaries from linguistic experiments, the corpus serves as a means of identifying recurrent errors in usage or grammatical short-comings which can be used to design targeted teaching and study materials.

Of most direct relevance to this study are the two types of multilingual corpora known respectively as comparable corpora or parallel corpora. The first of these, comparable corpora, is the term used to refer to two or more corpora that are constructed using similar parameters with each containing a different language or a different variety of the same language. Evans (2007) gives the example of CorTec, which consists of English/Portuguese comparable corpora containing examples of technical language from five areas. This corresponding pair of corpora was designed primarily for accurately translating meaning from one language to another and, can also be used for revealing elements of similarity and difference between the two languages (Santos, 2011).

Like comparable corpora, parallel corpora hold collections of texts in two or more languages. According to McEnery and Xiao (2007) parallel corpora contain "the same sampling frame and similar balance and representativeness for both source texts and their translations" (McEnery and Xiao, 2007, p.2). However, the key difference between comparable and parallel corpora is that the latter case, the texts have been aligned so that the user of parallel corpora can see all the examples of a particular search term in one language and view all their translated equivalents in the other language. This alignment can take place at a number of levels from text through to word.

3.5.3 The importance of Arabic/English parallel corpora

The adoption of parallel corpora in education and language has mushroomed since the 1980s, and studies of language comparison, cultural understanding and language translation acknowledge the benefits that continue to be received from its adoption (McEnery and Xiao, 2007; Papageorgiou et al., 1994). Ahmed and Nürnberger (2008) argue that developing and adopting parallel corpora may offer an effective solution that satisfies the increasing demand for Arabic/English translation. Several researchers were devoted their studies to Arabic and English parallel corpora translation (Farghaly and Shaalan, 2009; Resnik and Smith, 2003).

Although there have been many parallel corpora developed between various natural language, in reality these parallel corpora are often difficult to access and are of poor quality. Focusing on the development and adoption of contemporary Arabic/English parallel corpora, researchers need to find solutions to two key issues, namely, how to increase the quality of translation accuracy and to reduce the cost of development.

The quality of Arabic/English translation is affected by limitations including incomplete data, untagged entries, and limited availability of text genres (such as news stories) in the development process of the Arabic/English parallel corpora. Diab (2000) argues that this is because parallel corpora are developed in different technologies using different translation matching metrics. Liu et al. (2014) and Skadiņš et al. (2014) also advise that parallel corpora development in future should use innovative MT metrics to increase the quality of translation results.

Another factor preventing the growth of the use of the Arabic/English parallel corpora in society relates to time and expense, since it requires a lot of human labour to ensure that the pairs of sentences, phrases and words in translation are correct. Consequently, researchers, tend to decide to not make the parallel corpora available for free to the public in general. According to an analysis of existing Arabic/English parallel corpora by the Linguistic Data Corporation (LDC)(see Table 4.2) cost for purchasing parallel corpora can be up to $2,000 USD for just 38,000 words (23,000 in part 1 and 15,000 in part 2) for a small corpus whilst the cost of the most expensive corpora reaches $4,000 USD for 31 million words (LDC: online).

These problems of time and expense have been acknowledged by previous researchers carrying out parallel corpora research between Arabic and English. Skadiņš et al. (2014) suggest that researchers should develop new parallel corpora based on open-source corpus sources in the future.

3.5.4 Enhancing corpus quality

Since the linguistic data that makes up a corpus is often automatically captured from existing online corpora and web-sites (Bolia et al., 2000; Leacock et al., 1998), irrelevant data often finds its way into the corpus during the period of new corpus development. Moreover, McEnery and Xiao (2007) found that variability of types

of data collection software and methods create problems with data quality and a process of data verification may be useful in improving the quality of corpus. Data verification, also known as data cleaning has two purposes: cleaning up and de-duplication (McEnery and Xiao, 2007).

Corpus verification refers to a process of checking the quality of the corpus and establishing whether the collected data is useful for completing pre-planned goals in the corresponding language on various levels such as documentary, sentence, phrase or word (Callison-Burch et al., 2004) since a corpus may also be used by various types of researchers to accumulate statistical evidence to predict a pattern of how people use words or phrases differently (Hunston and Francis, 2000; Palmer et al., 2005).

The verification process of corpora can be implemented by using the written script of XQuery which is the query language for retrieving annotation stored in the rhetorical layer of annotated corpora. Retrieved data is transformed by the script into DOT which is the language for drawing graphs (Gansner and North, 2000). Figure 3.5 shows an example of the structure of a diagram from Gansner and North's research:

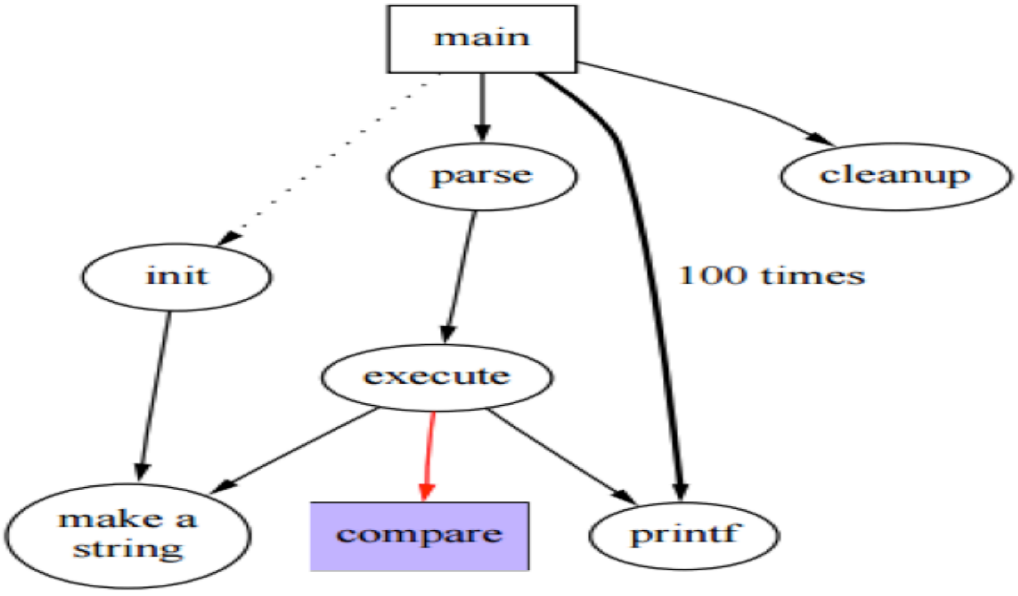

FIGURE 3.5: Gansner and North's Gansner and North's data retrieval process

Firstly, it is necessary to identify the word causing the error in translation. A source word that has two or more meanings is considered to be an ambiguous term (Burgess and Simpson, 1988) and ambiguity is one of the main reasons for

inconsistent translation in MT using corpora (Xiao et al., 2011). For example, "fall" is an ambiguous word that can be translated as " سقط" or " فصل الخريف".

Secondly, the developer needs to identify and collect all corresponding translations of the ambiguous data. Finally, new constraints are developed to teach the machine to identify the satisfactory translation.

Verifying the context is an iterative challenge to every creator of a text collection and dealing with a state of data duplication, in practice, may also entail many experiences with plagiarised, expanded, uncompleted files (Khmelev and Teahan, 2003). Therefore, Khmelev and Teahan (2003) proposed using an R-measure, which measures the normalised sum of the length of the selected contexts (Khmelev and Teahan, 2003). The R-measure is seen as a natural, easy-adopted method that marks duplications within the selected file in a range of number between 0 and 1, if the bigger the number, the more possibility there is that duplicated content exists in the corpus.

The process for constructing and verifying parallel corpora requires extra work. Koehn and Knight (2003) explain how to verify parallel corpora in the process of new corpus building. First, it is necessary to develop criteria in order to examine one-to-one correspondence in each language manually. The criteria divide all translations into several categories. The satisfactory translations are complete and meaningful sentences, while the translations are deemed unsatisfactory because they are usually incomplete, ambiguous or use informal expressions. Secondly, the corresponding translations in pairs are checked manually to identify inappropriate words or phrases from raw texts. Thirdly, new corpora are built with saved XML format files and then tested in experiments.

3.5.5 Parallel corpora and alignment

Lexical-based sentence alignment refers to a translation process that is based on dictionary matching. A dictionary is defined as a static database that saves millions of natural language expressions, and the translation process consists of seeking and making best-match value in the other natural language. The Lexical-based sentence alignment technique used to be adopted massively by industries and business to identify lexical clues and resources and assist in an analytical process, because a dictionary can be relatively easily created and the difficulty of

implementation and maintenance is relevantly low. Its primary advantage is the fast speed of translation for large size files. However, many researchers and users have complained about its poor accuracy. The reason is that a word or phrase in one natural language can be interpreted in many ways that may result in two completely different sentences, one long and the other short, or even in a couple of words, but the dictionary can only recognise the combination of the interpretation saved in the database, it cannot recognise another interpretation here.

In a parallel bilingual corpus, the correct alignment of the various textual elements (i.e. paragraphs, sentences, phrases, words) is an essential job for statistical alignment using the principle of probability theory and using probabilities estimated from the contents of the corpus. A number of different approaches to sentence alignment have been adopted such as sentence length, word co-occurrence, cognates, use of dictionaries and parts of speech etc. to produce a parallel bilingual corpus (Liu et al., 2014).

The flowchart in Figure 3.6 illustrates the general process which relates to how sentence alignment works in the Machine Translation process. Firstly, data that has been collected from Arabic texts and English texts are aligned in the development of an initial parallel corpus which is used for training proposes at the top of the flowchart. In general, data will be checked and aligned manually at this stage. Next, an alignment process is implemented in order to improve the accuracy of the corpora and build it as a parallel corpus that has only the one-to-one translation relationship between target language (in this case, Arabic) and output language (English). Subsequently, the built parallel corpus will be tested and used for training proposes to become an ultimate model of the Statistical Machine Translation (SMT). Therefore, users are enabled to translate their targeted Arabic files to English by adopting the parallel corpus model.

In a paper published early in 1991, Brown et al. (1991) argued that contents stored in a corpus should be logically segmented and aligned in small-sized units. In other words, corpora need to be aligned at the paragraph, sentence, phrase or even word level (Moore, 2002). Focusing on the sentence level, sentence alignment contains five situations in determining whether one sentence or more in one language aligns with one sentence or more in another language: 1 to 0, 1 to 1, 1 to many, many to 1, and 0 to 1. Therefore, the concept of sentence alignment means a logical algorithm supporting a probabilistic model in which it can be predicted that the

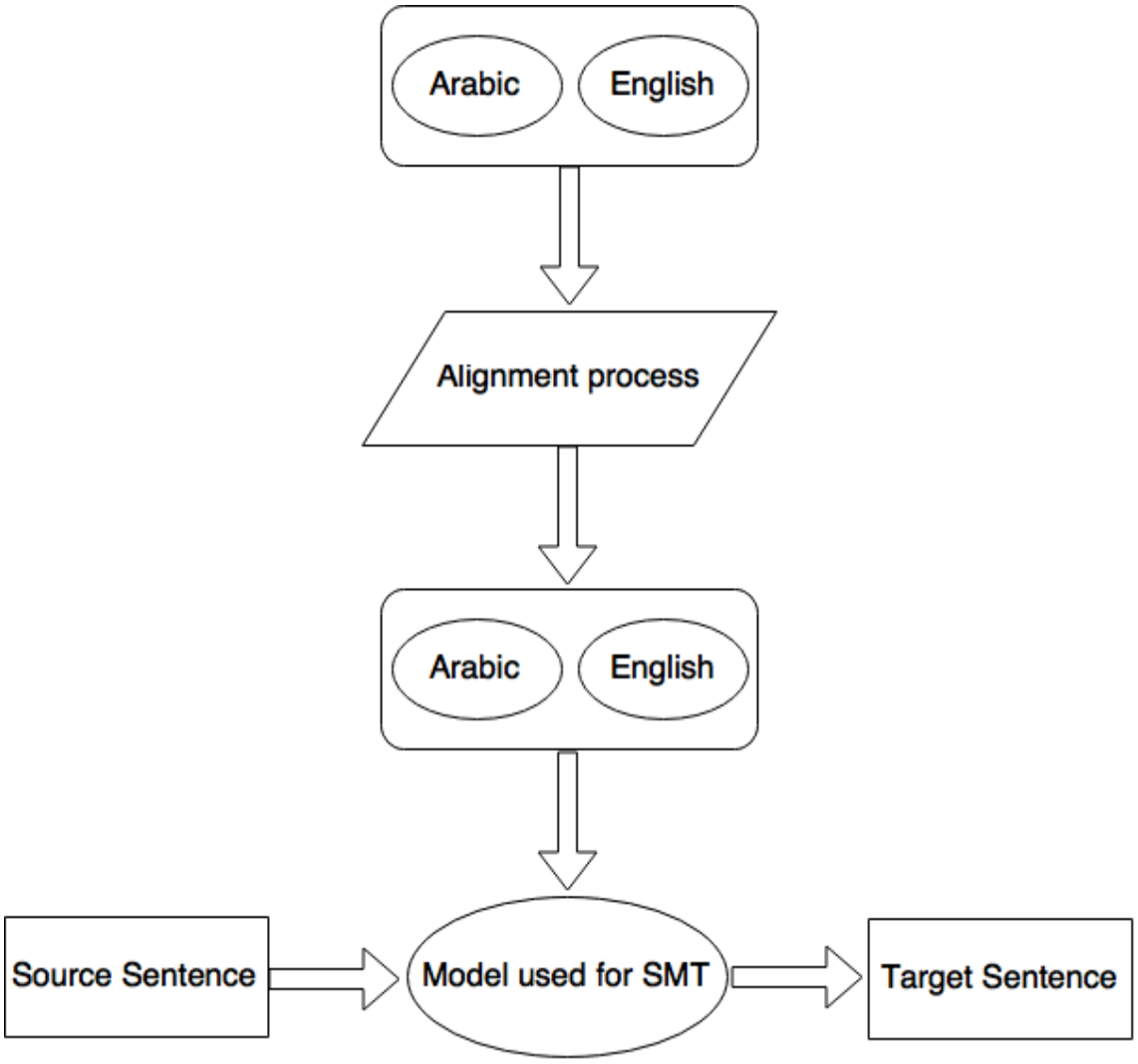

FIGURE 3.6: An ideal SMT Process.

beginning and end of the sentence structure and also maintains the same meaning of original texts Brown et al. (1991).

The sentence length metric assumes that the length for each sentence will be kept the same when it is translated from the source language into the target language. Gale and Church (1993) aligned parallel sentences in English-French and English-German corpora based on a sentence length metric that required calculating the character length of all sentences. A sentence length alignment metric was employed to solve several challenges of a correspondent sentence translation process on the basis of a statistical translation model.

Using an iterative programming mechanism, Gale and Church explored the steps needed to ensure the maximum likelihood of alignment of parallel corpora on a sentence level. Gale and Church (1993) achieved overall accuracies of 97% for

English-German and 94% for English-French. Wu (1994) aligned English-Chinese corpora by using sentence length values and reached an accuracy of 95%. The result also shows that it is possible for a MT process combined with sentence alignment can reach a 96% accuracy rate whilst this remains at 80% in more complicated trilingual corpora circumstances, including English, French and German.

Kay and Röscheisen (1993) developed a program that combined word and sentence alignment and calculated word probabilities by using the dice co-efficient. Haruno and Yamazaki (1996) used a similar method plus a bilingual dictionary for aligning English-Japanese corpora. Papageorgiou et al. (1994) used a sentence alignment metric based on the highest matching part of speech tags and matches restricted to nouns, adjectives and verbs, and reached 99% accuracy. Simard et al. (1992) used cognate-based approaches and found that sentence length difference worked well for sentence alignment. However, Melamed (2000) pointed out that because results were only reported for a relatively easy bilingual text, comparing two algorithms' performances in the literature is difficult.

In addition, Brown et al. (1993) calculated sentence length by using the number of words instead of the number of bytes or characters, which generated similar accuracies between 96% and 97%. Five translation metrics and processes that are developed based on simple statistical model were discussed to compare advantages of sentence alignment, and Brown et al. (1993) proposed a word-by-word algorithm that aligns sentences by calculating the frequency and probability of the word combination in natural languages respectively. Corpora of English and French were mainly collected from the Canadian Parliament and share some similar challenges with the Arabic and English experiments conducted in this thesis. For example, Figure 3.7 shows the process of how each word is aligned. Some words in French need to be translated by a phrase or many words in English, thus the MT process should identify whether a word is one-to-one or one-to-many translation, and which words correspond and how many.

3.5.6 Latest developments in quality enhancement

In the last decade, there have been some new proposals for sentence alignment for parallel bilingual corpora (Yu et al., 2012). One disadvantage of existing sentence alignment algorithms is that they are less effective when linking corresponding

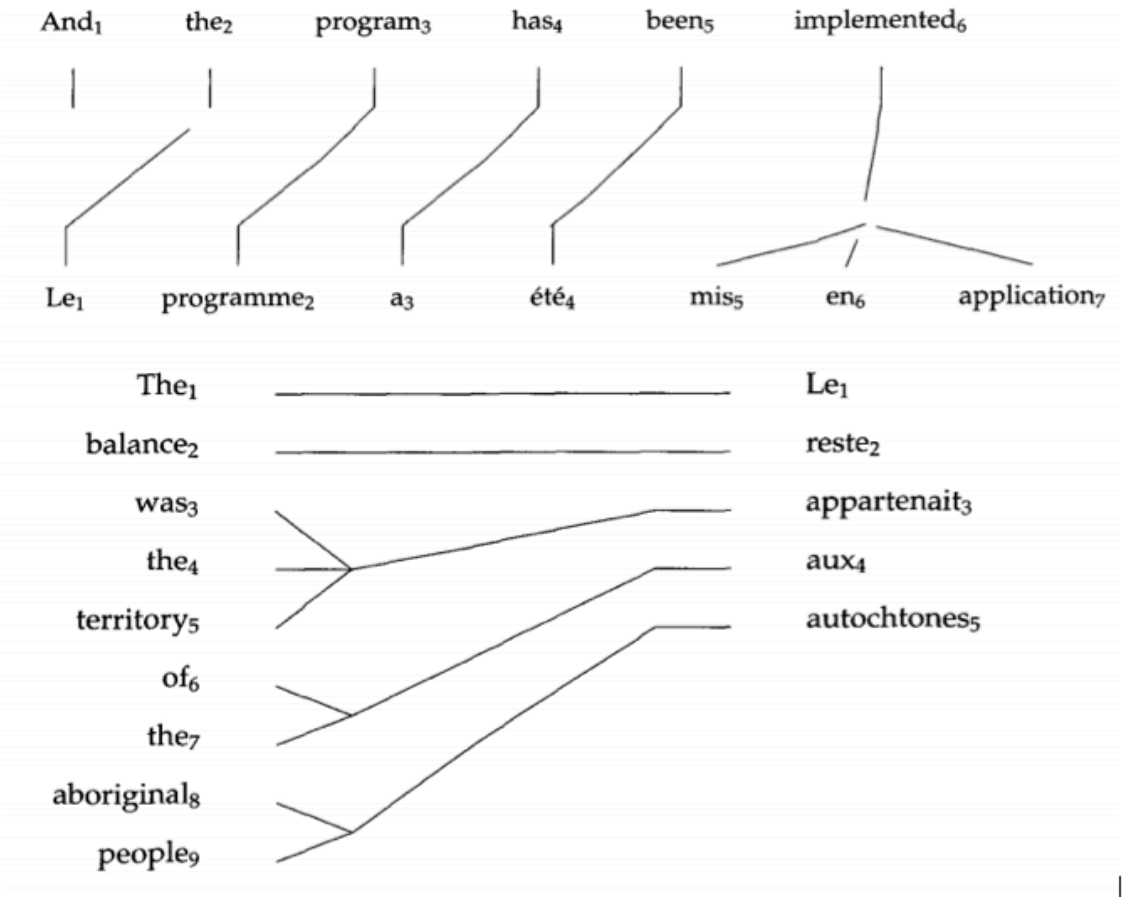

FIGURE 3.7: Alignment between independent English and French words respectively (Brown et al., 1993).

sentences if they are one-to-many or many-to-one mutual translations (Kutuzov, 2013).

Translation scripts in parallel corpora enable data to be compressed and then aligned at various levels such as sentence, phrase and word level. Two classifications that involve sentence length ratio distance metrics have been introduced in former studies (Ganitkevitch et al., 2011). However, they also acknowledged that the sentence length ratio-based alignment have disadvantages of phrase misalignment, and aligning multiple words with one unit. Therefore, they encouraged more MT studies to use a compression-based algorithm. The idea of using compression code length as adopted in this dissertation is that a sentence alignment metric hinges on the premise that the compression of co-translated text (i.e. documents, paragraphs, sentences, clauses, phrases) should have similar code lengths (Behr et al., 2003). This is based on the notion that the information contained in the co-translations will be similar. Since compression can be used to measure the information content, the research can simply look at the ratio of the compression

code lengths of the co-translated text pair to determine whether the text is aligned. That is, if you have a text string (i.e. document, paragraph, sentence, clause or phrase) in one language and its translation in another language, then the ratio of the compression code lengths of the text string pair should be close to 1.0.

Chapter Six in this thesis examines how the Prediction by Partial Matching algorithm can be used in preparing parallel corpora for MT.

3.6 Arabic Encoding

Arabic has been encoded for computers using various methods to represent and interpret the language (Mahmoud, 1994). The letters, diacritics and Arabic-Indic digits which are standard in Arabic are encoded by Arabic area encoding. Contextual forms that are in Arabic are not encoded. The letter variants which are often used for writing non-Arabic languages such as African languages are also encoded by the Arabic Supplement range (Kherallah et al., 2009).

Additional Quranic annotations are also encoded by the Arabic Extended-A range. The other forms of Arabic diacritics and the contextual letter forms are encoded by the Arabic Presentation Forms-B range.

The characters that are used in Arabic mathematical expressions are encoded by the Arabic Mathematical Alphabetical Symbols. This ECMA Standard cannot be used with some other ECMA Standards for 8-bit single-byte coded graphic character sets(Lantsov and Petrovskii, 2010). If there is a requirement that characters from more than one ECMA Standard, or coded characters are to be used and also there is a requirement for using the techniques of code extension, it is preferable to use the equivalent coded character sets from ISO/IEC rather than making use of the Standard ECMA-43 at level 2 or level 3 (Aliprand, 1992; ECMAScript and Association, 2011). Figure 3.8 shows Arabic Encoding types (Ecma, 2011).

3.6.1 UTF-8 encoding

UTF-8 is an abbreviation for the Unicode Transformation Format and 8 bit is the variable. The unit size for Arabic letters in UTF-8 is two bytes and the

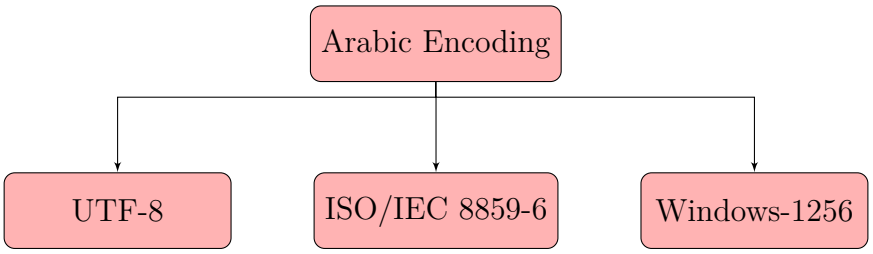

FIGURE 3.8: Arabic Encoding.

range of UTF-8 encoding for Arabic is between (U+0600) and (U+06FF). UTF-8 is variable-width encoding and enables representation of the characters in the Unicode character set. UTF-8 is one of the encoding forms of Unicode that is considered to be in the byte-oriented form and is interpreted in the sequence of bytes. UTF-8 can be defined in the form of supplementary characters.

UTF-8 encoding enables users to script and use the script language. It also provides support for symbols in mathematics and technical forms. UTF-8 was designed for the purpose of providing compatibility with the ASCII standard. It was also initiated to avoid problems that used to arise due to the complications caused in the byte order marks in UTF-16 and UTF-32. It provides additional support allowing scientific forms and equations to be exchanged. When using UTF-8 encoding, it is more appropriate to use Unicode. This is also preferable in backward usage and in environments that are more compatible and are supposed to be used in ASCII, such as Unix. UTF-8 is a technique that is used for making use of Unicode under systems such as Unix and Linux for this purpose, and both benefit from having an awareness of UTF-8 built in (Kuhn, 2013).

3.6.2 ISO(8859-6)

The International Organization for Standardization (ISO)/IEC 8859, published in 1987, comprises the 8-bit single-byte coded graphic character sets and is one of the parts of the ASCII-based standard character encodings. The main motivation for publishing this was to incorporate the languages which use the Arabic alphabets and formulate them in the 8-bit single-byte coded graphic character sets. It includes the encoding of the letters but does not cater for the pre-shaped letter forms. There are different parts of ISO/IEC 8859 and each part is comprised of 191 graphic characters. These sets have been dedicated to be used for specific groups of languages (IOFS, 1999).

3.6.3 Windows-1256

Windows-1256 is used for writing Arabic and other languages which make use of Arabic script in Microsoft Windows. Windows-1256 provides encoding for all the letters of the Arabic languages, including single characters.

Windows-1256 provides encoding for characters but not glyphs (i.e. small graphic symbols). It also does not cater for the visual forms of the letter variants that are isolated, initial, final or medial or are ligatured-letter shaped. The Arabic letters in Windows-1256 are encoded to be in the C0-FF range in alphabetic order.

There are also representations of some Latin characters. A Windows-1256 table is provided for developers to shift their applications to Unicode. The use of Unicode is preferable in any code page as there is a lot of language support and it is less unclear than the code pages (GGDC, 2014).

3.7 Parallel Corpora Evaluation

Language resources including properly compiled corpora of texts and careful utterances play a major role natural language processing research (Koehn, 2005). Parallel corpora such as texts, their translations and utterances provide good linguistic data across many languages. It plays a major role in machine translation systems as well as in cross language information retrieval. The Europarl corpus is one of the most frequently used (Koehn, 2005). It involves a collection of resources which includes 11 European languages derived from the European parliament proceedings (Hajlaoui et al., 2014). Other examples include the Bible, OPUS (Tiedemann, 2004), Czech–English and Hungarian–English.

Building any parallel corpus for natural or written languages requires enriched balance of texts. This balance of texts may be a problem hence the need for use of an opportunistic approach. For example, the balance becomes a problem when the key users are humans such as foreign language translators, linguistics teachers or students rather than applications used in machine translation systems. The correctness of the parallel units such as texts of the language in alignment with their translations determines the parallel's corpora quality (jaan Kaalep and Veskis, 2007).

In machine translation, the quality depends on the availability of parallel data. Metrics applied in evaluating machine translation systems emphasize more on the quality of parallel data used from the source language. Koehn (2005) explains that parallel corporal acquisition takes the following steps: data access, document alignment, sentence splitting, normalization, tokenization and sentence alignment.

To meet the quality with regard to content and translation required, data has to be chosen with care following the intended applications. Use of word alignment and machine learning algorithms is among the methods of evaluating parallel corpora. It assumes that parallel sentences includes many words that align compared to non-parallel sentences.

The quality of the output of machine translation systems allow analysis of the quality of the parallel texts (jaan Kaalep and Veskis, 2007). Recent studies (Hajlaoui et al., 2014) show that the performance of these systems varies greatly with the source language. Overlapping may sometimes occur in the parallel corpus. In such a case, including both overlapping parts should be avoided. Also, calculating quantitive measurements and assessing the purpose of its suitability are among the metrics of evaluating the corpus quality. Moreover, developing a model with machine learning for comparing a pair of sentences as either parallel or not could be of great importance.

Conversion of texts into a set of binary numbers is essential for machine learning algorithms to function properly. Many parallel corpora have been declared suitable for different purposes but many of them have not been evaluated (Tiedemann, 2004). Also, many of them have only partially been evaluated on how suitable they are for machine translation (Koehn, 2005). Therefore, once parallel corpora have been built, there needs complete and thorough evaluation for them to be considered useful in machine translation systems.

The diversity of the language and the nature of corpora provides the opportunity to evaluate the effectiveness of the machine translation systems in place. They also help to tackle problems and challenges from the technical and linguistic point of view (Samy et al., 2006). Researchers ought to investigate ways to improve machine translation systems for diverse languages. It is highly recommended that minority languages should be included in parallel corpora despite the texts shortage in comparison with the expanded languages. Also, it is important to test the

built parallel corpora for a much wider set of languages to verify that each language is applicable and to identify whether the evaluation of its quality correlates with human justification and judgment.

3.8 Summary

Though a series of translation models has been developed and adopted, the unsatisfactory quality and high cost of machine translation are always the two most commonly highlighted limitations. In order to enhance the quality and effectiveness of existing Arabic/English machine translation models and systems, this literature review began by exploring the development of machine translation models and systems, and identifying their respective limitations, as a means of selecting the most appropriate model for this present study. It focused in particular on the principles underlying statistical machine translation (SMT) development as this is viewed as being the most promising method for multi-respondent translation between Arabic and English. It also examined how concepts originally developed in the post-World War II period by Shannon (1948) and Weaver (1949) were developed and extended by Brown et al. (1990), to form the basis of our contemporary understanding of statistical models for machine translation.

The second part of the chapter outlined the different forms that corpora may take, with specific emphasis on parallel corpora and their key role in successful machine translation. This part of the literature review also considered the vital importance of verifying corpora and enhancing their quality by the process of text alignment. It concluded by examining the various ways of encoding Arabic within a computer-based system.

In the following chapters of the thesis, the focus shifts to the developmental aspect of the project, charting the development, verification and enhancement by experimental means of Arabic/English parallel corpora designed to be used for machine translation.

Chapter 4

Building parallel corpora for Arabic and English

4.1 Introduction

It is necessary to use text corpora for training purposes for machine translation systems in order to build more effective and robust systems. Unfortunately, many of the current Arabic corpora in the marketplace are of low accuracy and expensive. Seldom do they reach the fundamental requirements of academics and researchers. This chapter, therefore, discusses how two test parallel corpora for Arabic and English were developed from existing data in order to facilitate the experimental analysis in the following chapter.

In previous studies, researchers have adopted different corpora to support their translation-related research. The sizes of selected corpora differ substantially, varying from less than 1 million to 80 million words (Alansary et al., 2008).

Various researchers have suggested that future work in corpora research should seek cheaper resources to be adopted since they admit a large bulk of research funding is used for purchasing the corpora. Moreover, there are constant difficulties in either replicating the research process or conducting further research because most of the purchased corpora are only authorised for one-time use (Habash, 2008).

Al-Sulaiti and Atwell (2006) have identified the most frequently-used Arabic corpora in previous natural language processing research including translation (see Table 4.1).

Table 4.1 shows a large variation in size and also application. This demonstrates that 90% of existing corpora are used for commercial purposes (such as available for purchase (LDC, 2013)) and up to three quarters of corpora are created by researchers in Universities and other Higher Education institutions. The studies also claim that the satisfactory text has already been manually divided from the rejected ones in all of the corpora, although in many cases, verification checks such as the ones proposed in this dissertation show this to be far from the case. Also the accuracy of the existing corpora is fixed till the developers decide to employ more workers to improve them.

The size of the corpora is another limitation which has prompted the creation of a new parallel corpus for this research. For example, 70 percent of the corpora in the list contain fewer than 3 million words, and are hence too small to adopt, whereas the larger ones only cover a restricted set of semantic categories that occur in natural language. Therefore, there is an urgent need to create further free Arabic and English corpora in order to improve accuracy and reduce costs.

This chapter is organised as follows: first, it discusses existing Arabic/English parallel corpora. It is shown that these are quite expensive with fees in the thousands of dollars and/or are of poor quality. Secondly, the existing Arabic/English corpora, particularly in parallel corpora, are introduced to describe the development and uses of the corpora on the market in general. The creation of the parallel corpora for Arabic and English respectively is then discussed.

The purpose of this chapter is to create the corpora which are used later in the experimental analysis. The hope is that these corpora will also prove useful for other researchers.

No	Name of Corpus	Source	Medium	Size	Purpose	Material
1	Buckwalter Arabic Corpus 1986-2003	Tim Buckwalter	Written	2.5 to 3 billion words	Lexicography	Public resources on the Web
2	Leuven Corpus 1990 - 2004	Catholic University Leuven, Belgium	Written and spoken	3M words (spoken: 700,000)	Arabic-Dutch / Dutch- Arabic learner's dictionary	Internet sources, radio & TV, primary school books
3	Arabic Newswire Corpus (1994)	University of Pennsylvania LDC	Written	80M words	Education and the development of technology	Agence France Presse, Xinhua News Agency, and Umma Press
4	CALLFRIEND Corpus (1995)	University of Pennsylvania LDC	Conversational	60 telephone conversations	Development of language identification technology	Egyptian native speakers
5	NijmegenCorpus (1996)	Nijmegen University	Written	Over 2M words	Arabic-Dutch / Dutch-Arabic dictionary	Magazines and fiction
6	CALLHOME Corpus (1997)	University of Pennsylvania LDC	Conversational	120 telephone conversations	Speech recognition produced from telephone lines	Egyptian native speakers
7	CLARA (1997)	Charles University, Prague	Written	50 words	Lexicographic purposes	Periodicals, books, internet sources from 1975-present
8	Egypt (1999)	John Hopkins University	Written	Unknown	Machine Translation	A parallel corpus of the Qur'an in English and Arabic
9	Broadcast News Speech (2000)	University of Pennsylvania LDC	Spoken	More than 110 broadcasts	Speech recognition	News broadcast on the radio by the Voice Of America.
10	DINAR Corpus (2000)	Nijmegen Univ., SOTETEL-IT, co-ordination of Lyon2 Univ	Written	10M words	Lexicography, general research, NLP	Unknown
11	*An-Nahar* Corpus (2001)	ELRA	Written	140M words	General research	*An-Nahar newspaper* (Lebanon)
12	*Al-Hayat* Corpus (2002)	ELRA	Written	18.6M words	Language Engineering and Information Retrieval	*Al-Hayat* newspaper (Lebanon)
13	Arabic Gigaword (2003)	University of Pennsylvania LDC	Written	Around 400M	Natual language processing, information retrieval, language modelling	Agence France Presse, *Al-Hayat* news agency, *An-Nahar* news agency, Xinhua news agency
14	E-A Parallel Corpus (2003)	University of Kuwait	Written	3M words	Teaching translation & lexicography	Publications from Kuwait National Council
15	General Scientific Arabic Corpus (2004)	UMIST, UK	Written	1.6M words	Investigating Arabic compounds	http://www.kisr.edu.kw/science/
16	Classical Arabic Corpus (CAC) (2004)	UMIST, UK	Written	5M words	Lexical analysis research	www.muhaddith.org
17	Multilingual Corpus 2004	UMIST, UK	Written	10.7M words (Arabic 1M)	Machine Translation	IT-specialized websites
18	SOTETEL Corpus	SOTETEL-IT, Tunisia	Written	8M words	Lexicography	Literature, academic and journalistic material

TABLE 4.1: Arabic Corpora Summary (sources from: http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm)

4.2 Existing Arabic/English Parallel Corpora

For more than six decades both governmental and non-governmental institutions have researched how to apply technology to the problems inherent in translation. Developing high quality corpora requires an understanding of the developments that have previously taken place in this field, as well as those areas needing improvement. Before a new corpus can be created, it is necessary to revisit the advantages and disadvantages of previous Arabic/English parallel corpora.

Corpora	Size (words)	Price (US $)
ISI Arabic/English Automatically Extracted Parallel Text	31M	$4000
Ummah Arabic English Parallel News Text	2M	$3000
Arabic/English Parallel Translation	42K	$3000
Multiple Translation Arabic (Part 1)	23K	$1000
Multiple Translation Arabic (Part 2)	15K	$1000
Arabic Newswire English Translation Collection	551K	$1500
Arabic News Translation Text (Part 1)	441K	$3000
GALE Phase 1 Arabic Broadcast News Parallel Text (Part 1)	90K	$1500
IGALE Phase 1 Arabic Broadcast News Parallel Text (Part 2)	56K	$1500
UN Bidirectional Multilingual	1M	$4000

TABLE 4.2: Parallel Arabic/English Corpora as provided by the LDC (LDC, 2013).

The use of parallel Arabic/English corpora to train statistical MT models provides an effective way of building MT systems. However, currently Arabic/English corpora used throughout the world are still very limited. These limitations include incomplete data, untagged entries, and limited text genres being available (such as news stories). Worse still, most of the corpora are not available for public use. The corpora are provided by Linguistic Data Corporation (LDC), was approximately 35 million words. The corpora for Arabic and English that were analysed are enumerated in Table 4.2.

Table 4.2 shows that the only large size corpus is the ISI Arabic/English Automatically Extracted Parallel Text that contains 31 million words, whereas all the others are less than 2 million. It is also the most expensive corpus to buy, costing $4,000. Even the price of the smallest corpus, the Multiple Translation Arabic (Part 2), is $1,000 to buy. Two reasonable explanations for why building parallel corpora is so expensive have been identified: (1) the high cost of source decoding and (2) high cost of human judgement (Ambati and Vogel, 2010).

Franz et al. (2001) and Diab (2000) note that the enormous costs involved are one critical motivation that drove them to build open-source or crowd-knowledge-based parallel corpora. The money helps those building the corpora to obtain sources for the corpora, ensure authorisation for use and filter and tag raw scripts in corresponding languages. On the other hand, the money and time costs of using professional linguists are also very large during the corpora-building process. Since existing corpora are not able to recognise satisfactory sentences automatically, thousands of language specialists are required to check and test the accuracy of corpora manually. If the parallel corpora builders decide to import more raw data as corpora into the manual checking process needs to be repeated again to ensure the corpora maintain the same quality level (Liu et al., 2014).

To develop high quality corpora, it is important to understand the latest developments in this field and where improvement is still needed. Due to the high costs of corpora, it is critical that the quality of the data is as accurate as possible. The research described in this thesis has resulted in the quality of existing parallel corpora being improved. The improved data contains 58 million words in total which is greater than any of the currently-available corpora listed in the table. Of course, the hope is that larger sized parallel corpora will result in improved accuracy in translation for MT systems derived from the data.

Given that existing corpora are quite expensive with fees in the thousands of dollars, these are clearly unaffordable for students and many researchers. Moreover, many of the currently-available corpora are available from a single entity, the Linguistic Data Consortium (LDC), a venture sponsored by the University of Pennsylvania. If the improved data corpora produced by this research gains popularity, the hope is that it will provide an important alternative to this costly data.

4.3 Building an Arabic/English parallel corpus from an existing and a new source (Corpus A)

This section describes the development process of building a large extended and improved bilingual Arabic/English parallel test corpus.

It is critical to decide where to identify valuable resources to adopt before starting to build a corpus. Both academic use and commercial parallel corpora suggest that the World Wide Web (WWW) offers a good opportunity to seek translated text because it contains up-to-date natural language expressions in multiple languages and is also used by international users every day. The corpus in this research is a test corpus, which was built to facilitate experimental evaluation in later chapters from two WWW-based sources: *Al Hayat* and OPUS. The new large corpus, which will be called Corpus A, was first created containing 58 million words collected from two online sources: *Al Hayat* (http://www.alhayat.com) and OPUS (http://opus.lingfil.uu.se) (shown in Table 4.3) with permission obtained from the owners of the data.

Source	**Website**
Al Hayat	http://www.alhayat.com
OPUS – an open source parallel corpus	http://opus.lingfil.uu.se

TABLE 4.3: The parallel test Corpus A sources.

Source	**Arabic Sentences**	**Arabic Words**	**English Sentences**	**English Words**
Al Hayat	2,822	81,854	203,359	121,505
OPUS	3,638,420	58,380,784	27,775,663	30,808,480

TABLE 4.4: The sentences and word counts for the parallel test Corpus A sources.

The first source, *Al Hayat*, is a daily international newspaper with both English and Arabic versions, so it is seen as a good source as it contains translations of articles for both languages. The collected raw data from this source was 2,822 sentences and 203,359 words in total (of which 81,854 words were Arabic and 121,505 were English).

Additionally, all collected linguistic words were manually checked again to eliminate or reduce any basic errors, because several news articles contained more than one language in a sentence. The accuracy of these bilingual sentences may not be recognised by MT and can only be verified by manual checking.

To build a new parallel corpus successfully, Skadiņš et al. (2014) says that four challenges of text conversion have to be overcome: These are data cleansing and filtering, linguistic pre-processing and sentence alignment, in sequence. To ensure a high-quality parallel corpus, the developer of the corpus needs to employ specialists

in two languages at the stage of text conversion, making this stage of creating a new parallel corpus expensive and time-consuming. Furthermore, the success of a new parallel corpus development also depends on whether the raw data is aligned properly or not.

Firstly, the process of building Corpus A began by checking the contents of the Al Hayat newspaper website. Investigation revealed that there are two Al Hayat websites, one in Arabic, one in English, but both sharing the same structure to store and organise their contents. It is difficult to collect textual data by using standard data collection tools, so the author decided to manually collect data by himself. Each article was manual copied and pasted into a text file editor from both the Arabic and English websites and saved in text format. Subsequently, the author segmented the text file into parallel sentences followed by adding appropriate XML markup tags then deciding by a manual analysis on the labeling of sentences as "satisfactory" or "unsatisfactory". Meanwhile, the more data from the OPUS website was accessed because one of the major goals was to involve further sources in the new extended parallel corpus.

Secondly, the collected raw textual data needed to be checked and filtered to produce a relatively pure and clean corpus. Each Arabic script was segmented by sentence unit and paired with the sentence of the same meaning in English from the *Al Hayat* website. The author also combined these corpora with the OPUS download corpora and saved them in an integrated file in XML format. It took the author one year to complete this step and ensure all pairs of translations were tagged with unique ID numbers. A sample sentence in the parallel test corpus A is shown in Figure 4.1. The sample shows the sentence ID number, and then the Arabic and English texts.

Thirdly, the author used Java and Eclipse software to segment Corpus A into several small pieces because it was too large to open and edit otherwise, and also because there was no special requirement regarding the sequence of the corpus.

The steps involved in developing the new extended corpus are summarised in Table 4.5. The data was collected from two different bilingual sources and then all the collected documents were processed manually by the author and then divided into thirteen categories. This was done in order to facilitate further experimentation – the large corpus was also split into several different semantic classifications. Subsequently, source data was cleaned by verifying the ambiguous sentences, pairing

the most satisfactory corresponding translations, and all verified sentences were separated and segmented on a sentence level. Finally, all processed files were transformed and saved in XML format.

Step	Title	Description
1	Collection	Manually cut and paste from *Al Hayat* website
2	Collection	Concatenate data from OPUS
3	Classification	Manually classify data into 13 categories
4	Cleaning	Clean data by using PPMD5 (see Chapter 6)
5	Sentence Segmentation	Apply software to segment data
6	XML tagging	Tag data by XML

TABLE 4.5: The processing steps for parallel test Corpus A.

A second source, OPUS, (shown in Table 4.3) was used to make up the new extended and improved corpus. This is an existing open source parallel corpus that provides a large collection of translated texts from the Internet. Since all texts and data from the OPUS were collected automatically by machine, the quality of the original source was lower than expected, and was not of high quality (Tiedemann, 2009).

Consequently, a lot of problems were found with the OPUS corpus, such as Arabic sentences which were not translations for the English sentences, or some missing translations, as shown in Table 4.6.

Table 4.6 provides ten examples which illustrate the many errors that were found in the OPUS corpus:

- Sentence number 1: The translation is completely wrong.
- Sentence number 2: The translation has omitted the "The Conducts that are carried out at the direction of the state or under its control".
- Sentences numbered 3 to 9: The translation is completely wrong.
- Sentence number 10: The translation is fragmentary and is missing the phrase "A common implementation of the transition plan between the UN-AMSIL and United Nations country team."

The completely wrong translations include problems in varying degrees such as:

1. An alignment contains several sentences or a sentence aligns with half or an incomplete sentence from Arabic to English.
2. Specific information was not translated in the first place such as year, list and bullet points.
3. The automatic Machine Translation software could not recognise the abbreviations and acronyms in a segment, for example, UNAMSIL / United Nations Country Team Transition Plan (example 10).

As shown in Table 4.4, the OPUS collection consists of Arabic and English documents at the sentence level. It contains 3,638,420 sentences, or 58,380,784 words in total (27,775,663 of which are Arabic and 30,808,480 English).

The quality of all aligned texts was manually checked by the researcher and then the corpus was divided into thirteen categories that enabled the author to control the size of corpora in the later experiments and allowed further experiments on a category basis. The final text was converted into XML format.

Table 4.7 shows a statistical summary of Corpus A. It includes various categories of text such as Books, Business, Cinema, Conferences, Crimes, Decisions, Economy, Geographies, Issues, Law, Politics, Reports and Stories.

Partitioning the corpus into various classifications is a unique characteristic of Corpus A. This was done in order to make it easier during the experiments in the later chapters to discover badly translated sections of the corpus since this was created from different sources. For example, experiments showed that overall certain categories were more poorly translated than others and therefore this separation facilitated identification of problem areas in the corpus.

Segmenting the large corpus A into small categories also allows the builder or maintainer to update the corpus by changing only parts of it. However, though Corpus A enables running of large-scale experiments, the pilot experiment results indicated that the accuracy of translation was lower than expected. One known reason is the raw source of OPUS that contains numerous errors and the research required a more accurate corpus than the initial version of Corpus A.

Example	Arabic	English
1	3 - ليس في هذه الاتفاقية ما يمس حقوق والتزامات أي شخص بمقتضى القانون الذي يحكم الصكوك القابلة للتداول.	(g) A letter of credit or independent guarantee.
2	التصرفات التي يتم القيام بها بناء على توجيهات الدولة أو تحت رقابتها.	Conduct directed or controlled by a State.
3	الاحتياجات من الوظائف.	a Some of the extrabudgetary posts may not be available for the full biennium due to changing requirements of the funds and programmes for services rendered by the United Nations.
4	(ز) خطاب اعتماد أو ضمانة مستقلة.	(f) Bank deposits.
5	باء - برنامج العمل().	1 Subprogramme 2 of programme 24 of the biennial programme plan and priorities for the period 2006-2007.
6	تخطيط البرامج وميزنتها.	Office of Programme Planning, Budget and Accounts Organizational structure and post distribution for the biennium 2006-2007
7	شعبة الحسابات.	b Outposted to the Department of Economic and Social Affairs for financial matters related to technical cooperation.
8	كما لا يوجد ربط شبكي مستمر مع وكالات الأمم المتحدة في المنطقة دون الإقليمية.	Furthermore, the reports of subregional offices are often provided as hard copies to delegates to intergovernmental committees of experts as they arrive for the meeting, depriving them of sufficient time to study them and conduct an informed discussion.
9	جيم - أهداف السياسات العامة التي ترمي إلى جعل السياحة أكثر استدامة 5.	Item 5 (c) of the provisional agenda∗
10	ألف تنفيذ الخطة الانتقالية المشتركة بين بعثة الأمم المتحدة في سيراليون والفريق القطري للأمم المتحدة.	Implementation of the UNAMSIL/United Nations country team transition plan.

TABLE 4.6: Some examples of unsatisfactory translations that were found manually from the raw OPUS texts.

The final numbers of Arabic and English characters and words in each of the categories are also shown in Table 4.7. The Cinema classification is the largest one in both Arabic and English because different movies have different storylines and it contains not only the relevant words used by the film industry, but also items and catchphrases which are not easily classified elsewhere. Though the Crimes and Books classifications are the two smallest ones in Table 4.7, they still contain more than one million English words and the corresponding number of Arabic words.

Categories	**Arabic Characters**	**English Characters**	**Arabic Words**	**English Words**
Books	10,574,252	7,242,426	931,836	1,079,699
Business	26,367,126	17,987,925	2,289,276	2,624,274
Cinema	61,557,926	36,482,892	7,919,902	8,127,509
Conferences	21,696,083	15,129,972	1,879,527	2,215,857
Crimes	10,147,866	6,473,170	933,842	1,005,221
Decisions	15,863,975	10,822,315	1,397,181	1,605,851
Economy	25,962,438	17,760,514	2,266,424	2,599,651
Geographies	16,096,053	10,924,063	1,392,099	1,595,115
Issues	11,390,107	6,937,792	1,051,195	1,042,316
Law	16,083,105	10,936,231	1,407,292	1,597,873
Politics	23,427,958	15,675,917	2,035,969	2,304,233
Reports	15,960,285	10,819,195	1,388,457	1,590,056
Stories	29,703,105	20,294,105	2,882,663	3,420,825
Total	284,830,279	187,486,517	27,775,663	30,808,480

TABLE 4.7: Character and word counts for parallel test corpus A.

The text in parallel test Corpus A has been segmented by sentence and has been produced in XML. A sample sentence in parallel test Corpus A is shown in Figure 4.1. The sample shows the sentence ID number, and then the Arabic sentence and English translations. In Figure 4.1, <sentence> is the opening tag for the Arabic and English sentence which has <id> of 254 in this sample. Furthermore, <Arabic> specifies the Arabic sentence and <English> specifies the English sentence.

However, a primary purpose of this research was to develop a higher quality parallel corpus. Therefore, it was necessary to look for a solution to improve the poor quality of the original translations and to further improve the quality of the OPUS source by using sentence-matching metrics discussed later in this thesis. The sentence-matching metrics use a new generation compression PPM-based alignment approach to filter and align the units of analysis in a matrix as a general approach.

```
<sentence>
<id>254</id>
<Arabic>
طال الحديث الى ان استولى الليل على باريس.
</Arabic>
<English>
The conversation dragged on until the night took over Paris.
</English>
</sentence>
```

FIGURE 4.1: A sample from parallel test corpus A between Arabic and English.

Finally, the author also separated a much smaller random sample of the good translations into a satisfactory dataset, and the bad ones into an unsatisfactory dataset. This was done because the value of the corpora in both datasets was to be tested in future experiments as detailed in later chapters.

Corpus A was used as a source to build a second corpus, Corpus B. Corpus B contains a small subset of Corpus A that was judged to be in one of two categories: satisfactory translation; and unsatisfactory translation. This corpus is discussed in more detail in the next section.

4.4 Building a small Arabic/English parallel test corpus (Corpus B) from Corpus A

The second smaller test corpus, Corpus B, has been developed in order to aid the experimental evaluation described in later chapters.

This corpus was created by choosing at random 10,000 translations taken from Corpus A which were judged to be satisfactory and 2,000 translations judged unsatisfactory. These were manually selected from Corpus A and formed the validated reference data for later experiments (see Table 4.8). This corpus was needed in order to test the efficiency of various compression-based metrics and sentence alignment tests, as discussed in later chapters. Table 4.8 shows ten example sentences deemed satisfactory as samples from Corpus B.

Example	Arabic	English
1	أعرف أن الموت حق، إلا ان من حقي أن أحزن، أن أتألم، أن أسترجع ذكريات عزيزة على قلبي مع الأمير الراحل.	I know that death is ineluctable, but I have the right to be saddened, to feel sorrow, and to recall some memories that are dear to my heart with the late Prince.
2	عدت الى لندن من زيارة سريعة الى بيروت، ثلاثة ايام فقط، ووجدت على مكتبي قصاصات جمعها لي قسم الارشيف من مصادر المعلومات التي اتوكأ عليها، وبينها دور بحث وميديا اميركية وبريطانية.	I returned to my office in London from a quick three-day visit to Beirut, and found on my desk clippings gathered for me by the Archive Department from the usual sources I rely on, including British and American think-tanks and the media.
3	في روسيا ونيجيريا وكينيا الأرقام أفضل لإسرائيل، ولكن في مصر الرقم هو 85 إلى سبعة، وفي باكستان 50 إلى تسعة، وفي إندونيسيا 61 إلى ثمانية.	In Russia, Nigeria, and Kenya, the figures are more favorable to Israel. However, in Egypt, the figure was 85 percent to 7, in Pakistan 50 to nine, and in Indonesia, 61 to eight.
4	كنا صحافيين صغاراً ننظر الى غسان تويني كمثل يحتذى وكل منا يأمل أن يسير في خطاه.	When we were young journalists, we used to look up at Ghassan Tueni as a role model, and we all hoped to walk in his footsteps.
5	فكل هذه الظروف كانت موجودة في ظل أسعار نفط ارتفعت إلى أكثر من 120 دولاراً للبرميل وهي الآن سائدة في ظل أسعار تراجعت.	All of these conditions existed when prices rose to more than $120 a barrel, and these conditions continue to exist as prices fall.
6	ولعلهم اليوم ينظرون بارتياح إلى مواسم حصاد أزمات ترسيم الحدود. ففي القارة الإفريقية ما يزيد على عشرات الصراعات التي تخفت وتظهر، على خلفية التوزيع العرقي والجغرافي والديني حتى.	They are perhaps today looking with relief at the harvest season of crises of border delineation. Indeed, there are in African continent dozens of conflicts that wane then reemerge, on a background of ethnic, geographic and even religious distribution.
7	والراجح أنه كان يضع في اعتباره أن المشكلة لا تكمن دائماً في عدم وجاهة القرارات الصادمة للرأي العام مثل رفع أسعار البنزين بقدر ما يطاول أسلوب عرضها والدفاع عنها.	It is likely that he took into account the fact that the problem does not always lie in the irrelevance of the decisions that are shocking to the public – such as increasing the price of gasoline – as much as in the way they are put forward and defended.
8	ما يتطلب من المسؤولين عن السياسات البترولية في الشرق الاوسط التنبه الى أخطار نابعة من الغموض السياسي، كما على الدول المنتجة الكبرى في الشرق الاوسط التنبه الى عدم استنزاف الاحتياط المتوافر لديها.	Therefore, this requires officials responsible for oil policies in the Middle East to take note of the risks resulting from political uncertainty, while major oil-producing countries in the Middle East must also pay attention to not deplete their hydrocarbon reserves.
9	مقال الجريدة يميل الى الجزرة، ويتحدث عن نجاح البرنامج السعودي لإعادة تأهيل المتطرفين، وعن برامج مثله في سري لانكا وسنغافورة والفيليبين.	The newspaper piece is tipped in favour of the carrot approach, and speaks of the success of the Saudi program to rehabilitate extremists, and similar programs in Sri Lanka, Singapore and the Philippines.
10	البطولة أسوأ مهنة في العالم، والبطل يُذكر يوماً أو إثنين ثم يُنسى، ولا يبقى غير ألم أفراد أسرته الأقربين ومعاناتهم سنين.	Heroism is the worst profession in the world and a hero is remembered for a day or two but is soon forgotten, and all that is left after that is the sorrow of his close relatives and their long suffering.

TABLE 4.8: Some examples of the sentences deemed satisfactory of the parallel corpus B

Table 4.9 shows ten example sentences deemed unsatisfactory as samples from Corpus B.

Example	Arabic	English
1	2 - إطار قانوني لضمان المعاملة المنصفة للمنشآت والموظفين التابعين	Main features of the building to house the permanent Secretariat 8 Basis for placing the office facilities at the disposal of the permanent Secretariat 9 Responsibility for maintenance, repair, and utilities 9 Office facilities furnished and equipped by the host Government 9 Duration of the arrangements regarding office space 9
2	3 - الأحكام بما في ذلك أي قيود تنطبق على تشغيل الأشخاص الذين يعولهم الأعضاء 9	Description of facilities and conditions 9
3	6 - أساس وضع التسهيلات المكتبية تحت تصرف الأمانة الدائمة 12	d. International conference facilities 12
4	خدمات الدعم المقدمة في مركز فيينا الدولي	The Division also provides limited administrative support to offices of other United Nations entities located at the Vienna International Centre, such as the Office of the United Nations High Commissioner for Refugees, the United Nations Office for Project Services, and to the United Nations Interregional Crime Research Institute located in Turin, Italy.
5	الاحتياجات من الموارد	28F.17 Resources amounting to $842,900 reflecting an increase of $169,200 provide for three posts, including one new P-3 post for the Secretary of the Joint Appeals Board/Joint Disciplinary Committee, Vienna and related non-post resources.
6	اعتماد مقررات اتفاقية فيينا من قِبل الاجتماع السابع عشر لمؤتمر الأطراف في اتفاقية فيينا	the seventh meeting of the Conference of the
7	حاء حاء - مشروع المقرر 17/حاء حاء: الاجتماع الثامن لمؤتمر الأطراف في اتفاقية فيينا	with a firm date to be announced as soon as possible.
8	وسأغدو ممتنا لو عملتم على تعميم هذه الرسالة باعتبارها من وثائق مجلس الأمن.	(Signed) Ali Said Abdella Minister
9	(ه) النظر في تمويل التكنولوجيا في إطار التنمية المستدامة والتكيف.	The engagement of the insurance industry is essential as risk insurance could be a passive option for financing these technologies that redistributes the risks between different actors.
10	* الأهداف الإنمائية المتفق عليها دوليا بما فيها تلك الواردة في الإعلان بشأن الألفية.	Notes that UNCDF has efficiently and effectively addressed the specific needs of least-developed countries through its local governance and microfinance programmes, thereby playing a clear role in achieving the Millennium Development Goals (MDGs) * at the local level;

TABLE 4.9: Some examples of the sentences deemed unsatisfactory from the parallel corpus B

4.5 Summary and Discussion

This chapter describes the process of building two high quality test corpora. The main goal was to build a high-quality and inexpensive parallel test corpus for testing the effectiveness of the PPM compression metric in sentence alignment experiments. The bonus outcome of the parallel test corpus is that a new high-accuracy and cheap parallel corpus in Arabic and English has been produced.

The chapter began with an investigation of 20 existing parallel corpora and summarised their advantages and disadvantages to inform the process of building a parallel corpus. Habash (2008) found that seven of the existing ten corpora are used commercially and their prices are quite high. The cost of human judgement of corpora text quality has been recognized as one of the major limitations preventing improvement of quality and cost decreasing.

Moreover, it was found that the existing corpora are of poor quality and further work was needed to produce a parallel corpus of sufficient quality that can be used to train an MT system.

Inspired by the suggestions of Skadiņš et al. (2014), the study decided to build a new improved and extended parallel test corpus by using the semantic scripts from the *Al Hayat* newspaper website, and to enlarge the size of the corpus by combining this with OPUS, an open-source corpora. *Al Hayat* is an international newspaper targeting readers with an interest in politics and OPUS is a large-scale open-source online corpus. The author of this thesis used several crawling and sentence alignment approaches to extract the raw data from their advanced search engines and convert the format into standard XML format. Then the original data was cleaned and filtered through a manual check because automatic MT does not easily recognise some particular types of expression in Arabic such as years, ratio and bullet point lists.

The corpus which was produced (Corpus A) has 58,584,143 words from two selected sources (27,775,663 in Arabic and 30,808,480 in English) and the component data was classified into thirteen classifications including Books, Business, Cinema, Conferences, Crimes, Decisions, Economy, Geographies, Issues, Law, Politics, Reports and Stories. The most significant benefit of classification is that it allows results in each to be examined to see if there are differences in quality or not.

However, the initial experimental result showed that the quality of Corpus A is lower than expected. One main reason for the low-accuracy of the data is that the original corpus from the OPUS website contained many mistakes due to its open-source character.

Therefore, substantial further improvement in the quality of Corpus A is required. The purpose of the remaining chapters is to develop new techniques based on the PPM compression scheme to aid in the process of improving the quality further.

A smaller sub corpus of Corpus A, Corpus B, containing 10,000 sentences judged to be satisfactory translations and 2,000 sentences judged unsatisfactory was created. This was done in order to aid in the experimental evaluation of the new alignment techniques as discussed in subsequent chapters.

The creation of these two new parallel test corpora, A and B, provides an important contribution to existing MT study, specifically Arabic/English parallel corpora study. Firstly, it increases the number of high-quality corpora for this language pair.

The new parallel test corpora can be adopted in various experimental research such as sentence alignment, phrase and word alignment, and corpora comparison and so on. Secondly, the higher quality large scale corpora manual tests help to improve accuracy in the results of later experiments. Quality of the experimental result in the later stage relies on the accuracy of the original data. In the other word, the accuracy of original corpora is the key that leads the sentence alignment experiments to success.

Finally, the manual process of construction allowed the author to discover problems and learn more about issues and principles of the parallel corpora. The newly-acquired skills and experience helped the author to produce more accurate alignment in order to identify the best thresholds in later experiments.

Chapter 5

Analysing and verifying Arabic/English parallel corpora

5.1 Introduction

This chapter describes how the Prediction by Partial Matching (PPM) compression scheme was used in this project as a method for sentence alignment based on compression code length levels to check the level of quality of sentence pairs in a parallel Arabic/English corpus. The purpose in this case is to increase the accuracy of sentence alignment between the two sets of corpora.

This chapter begins by considering the nature of the PPM-based compression scheme, and further discussing the different compression algorithms for code length ratio distance metrics, the aim being to better understand the principle of the PPM approach to sentence alignment. It addition, it also explains the reason why adoption of a compression-based approach could improve accuracy in comparison to other methods which have previously been employed for this purpose.

The performance of this new compression-based approach is then evaluated using a series of experiments which focus on the data compression ratio. Most of these experiments are conducted using the material in Corpus A. Once the level of reliability of this PPM compression approach has been verified using these experiments, the quality of the translation is then analysed in terms of the sentence length ratio formula. The results of this analysis are presented in more detail in a series of tables and charts which provide illustrative examples of both satisfactory

and unsatisfactory translation of extracts taken from the Arabic/English parallel corpora. Following this, the results from the new corpora, containing both satisfactory and unsatisfactory aligned samples, are reviewed.

The chapter concludes by re-considering the translation samples which were rejected in order to assess the quantity of the raw translation material in the corpora which cannot be used, and furthermore, to highlight the reasons why it would be inappropriate to use this material for MT purposes at this stage.

The Prediction by Partial Matching (PPM) data compression scheme was first proposed by Cleary and Witten (1984) who originally referred to this as Partial String Matching. It can be used to predict the next symbol or character in a stream from a fixed order context. The context models are adaptively updated as the text is processed sequentially using an online process. Experiments conducted by Teahan (1998) and later Alhawiti (2014) have shown that for both English and Arabic text order 5 models (i.e. those which use fixed order contexts of length 5) perform best at compressing text using the PPMD variant of the algorithm developed by Howard in 1993, which was based on the PPMC variant devised by Moffat in 1990 (Wu, 2007).

The main difference between the four variants of the PPM data compression scheme, namely PPMA, PPMB, PPMC and PPMD, relates to the calculation of the *escape* probability when the model needs to back off to lower order models if a symbol is not predicted by a higher order model.

Formally, the estimation of the escape probability for PPMD is $e = \frac{t_d}{2T_d}$ and for the symbol probability it is $p(\varphi) = \frac{2c(\varphi)-1}{2T_d}$ where: d is the current coding order of the model; φ is an upcoming symbol ($\varphi = x_{n+1} \in A$); s_d is the current context $s_d = x_n, \ldots, x_{n-d+1}$; $c_d(\varphi)$ is the number of times that the symbol φ occurs in the context s_d; t_d is the total number of unique symbols that occur after the context s_d; T_d is the total number of times that the context s_d has occurred; $T_d = \sum c_d(\varphi)$, e is the escape probability; and $p(\varphi)$ is the probability of the upcoming symbol φ.

This thesis uses PPMD with $d = 5$ since, as previously stated, this variant performs better than PPMA, PPMB and PPMC and experience has shown that this is most effective coding order for the model in the case of compressing English and Arabic text. Table 5.1 shows how the probabilities are estimated using PPMD when the model has been trained using the sample Arabic text string of “ سبيللسلسبيلا ”.

Table 5.1 illustrates the predictions, frequency counts c and probability estimates p for the order $k = 3$, $k = 2$, $k = 1$, $k = 0$ and $k = -1$ PPMD contexts (where k is the order of the model or context length). For example, only one symbol has been predicted in the single order 3 context – this has occurred once in the training text, and therefore its probability estimate that it will occur again is 3/4 and the probability estimate that a previously-unseen symbol in this context will occur instead is 1/4. In this case, therefore the use of lower order models is needed in order to estimate the probability of the unseen symbol.

As the example shows, the model will keep on escaping down until it encounters a context where either the symbol has been seen before or the symbol will be encoded using the default $k = -1$ context. In this context, every symbol is estimated as being of an equal probability $1/|A|$ where $|A|$ is the size of the alphabet.

Order $k = 3$			Order $k = 2$			Order $k = 1$			Order $k = 0$		
Prediction	c	p	Prediction	c	p	Prediction	c	p	Prediction	c	p
ل→سبي	2	3/4	ي→سب	2	3/4	ب→س	2	3/8	→س	4	7/30
									→ب	2	3/30
									→ي	2	3/30
						→ل	2	3/8	→ل	6	11/30
									→ا	1	1/30
→ esc	1	1/4	→ esc	1	1/4	→ esc	2	2/8	→ esc	5	5/30
ل→بيل	1	1/4	ل→بي	2	3/4	ي→ب	2	3/4	Order $k = -1$		
→ا	1	1/4							Prediction	c	p
→ esc	2	2/4	→ esc	1	1/4	→ esc	1	1/4	→ A	1	$1/\|A\|$
ل→يلل	1	1/2	ل→يل	1	1/4	ل→ي	2	3/4			
			→ا	1	1/4						
→ esc	1	1/2	→ esc	2	2/4	→ esc	1	1/4			
س→للل	1	1/2	ل→لل	1	1/4	ل→ل	2	3/12			
			→ا	1	1/4	→س	3	5/12			
						→ا	1	1/12			
→ esc	1	1/2	→ esc	2	2/4	→ esc	3	3/12			
ل→للس	1	1/2	ل→لس	2	3/6						
			→ب	1	1/6						
→ esc	1	1/2	→ esc	2	2/6						
س→لسل	2	3/4	س→سل	2	3/4						
→ esc	1	1/4	→ esc	1	1/4						
ل→سلس	1	1/4									
→ب	1	1/4									
→ esc	2	2/4									
ي→لسب	1	1/2									
→ esc	1	1/2									

TABLE 5.1: Processing "سيبللسلسبيلا" using the PPMD model.

5.2 Code length ratio distance metric for matching sentences

In this context, the term 'code length' refers to the size (in bytes) of the compressed output file produced by the PPM compression algorithm. When using PPM to compress either Arabic or English text, the code length is a measure of the cross-entropy of the text, which is the average size (in bytes) per character for the compressed output string. Theoretically, the cross-entropy is estimated using the following equation:

$$H(S) = -\frac{1}{k}\log_2 p(S) = -\frac{1}{k}\sum_{i=1}^{k} -\log_2 p(x_i|x_1 \ldots x_{k-1})$$

where $H(S)$ is the average number of bits to encode the text and k is the order of the model (in the case of the models used in this thesis, $k = 5$).

Note that the compression code length (which is the number of bits required to encode the text string losslessly) can be expressed simply as $nH(S^L)$, so that it can be unambiguously decoded.

The ratio of the compression code lengths of the parallel text strings for the languages English (E) and Arabic (A) is defined as follows:

$$R(S^E, S^A) = \frac{n}{m} \times \frac{H(S^E)}{H(S^A)}$$

where S^E is an English text string with length n and S^A is an Arabic text string with length m. The code length ratio (CR) is defined as:

$$CR = \max\left\{R^{E,A}, R^{A,E}\right\}$$

In their experiments with compression-based methods for English/Chinese sentence alignment in parallel corpora, Liu et al. (2014) showed that code length ratio is a more effective distance metric for aligning sentences in these two languages than a distance metric which is based on sentence length. The primary purpose of the research conducted in this thesis was to investigate whether this

would also be the case when applying code length ratio as a distance metric for sentence alignment with an Arabic/English parallel corpora.

5.3 Sentence length ratio distance metrics for matching sentences

Automatically generated bilingual corpora often contain a large number of noisy sentence pairs. Consequently, researchers have sought to devise and test various methods which can be employed for filtering out these noisy sentences from parallel corpora (Khadivi and Ney, 2005). However, for the experiments discussed in Chapter Six, a new technique has been devised to carry out this filtering process. This is a new hybrid method based on a combination of the compression code length ratio (CR) and the standard sentence length ratio (SLR) described by Mújdricza-Maydt et al. (2013), which is particularly effective when applied to Arabic/English sentence pairs and helps to produce a high-quality corpus. The SLR for a pair of Arabic/English sentences in parallel corpora can be calculated as follows:

$$SLR = \max\left\{\frac{L^A}{L^E}, \frac{L^E}{L^A}\right\}$$

where:
L^A is the length of the text for Language A (Arabic), and
L^E is the length of the text for Language E (English).

5.4 Experimental Evaluation

As detailed in Chapter Four, the parallel Arabic/English Test Corpus A was created for the purposes of conducting a number of experimental evaluations. These experiments are described in details and discussed in the following sections.

5.4.1 Compression experiment one

Some preliminary compression experiments were conducted to determine if the compression CR measure would prove to be effective as a metric for determining the level of quality of translation between Arabic and English sentence pairs from the corpus.

As previously stated, standard PPM is an adaptive technique with its language models starting from null when the beginning of a text string is processed. The context frequency counts, from which the probability estimates are made, are then updated as the text string is processed sequentially. In the case of longer text strings (such as paragraphs or whole documents), the PPM algorithm will usually have enough text in order to train its models effectively so that higher order contexts are used for the majority of predictions, leaving less need to escape down to lower order contexts.

One obvious concern when using PPM code lengths for sentence alignment is that sentences may not be long enough in order for more reliable probability estimates to be made for the CR calculation. However, a simple expedient which can be used to overcome this particular challenge is to prime the PPM models prior to the compression.

A large corpus that is representative of each of the languages in the pair can be used for this priming prior to the compression being performed Suitable examples would be the Brown Corpus for English (one million words) and the Corpus of Contemporary Arabic (one million words). This priming text is used to 'train' the models. Liu et al. (2014) found this approach to be very effective in their study which used compression code length-based metrics for sentence alignment between Chinese and English texts.

The purpose of the preliminary experiments described here in this section was to determine how effective this priming of the PPM models proved to be for compressing Arabic sentences, and also to gauge the effectiveness of the primed PPM compression method as a sentence-matching metric.

A key requirement of using the CR metric is that the compression code lengths in the pair of languages in question should be similar for sentences that are meant to function as co-translations of each other.

The intuition is that if the sentences are satisfactory co-translations, then each one of the pair should convey exactly the same amount of information. Since compression code length has been found to be an effective method for measuring information (see Teahan, 1998 for details of several studies in this area), then it might be expected that roughly 50% of the compression code lengths of sentences in one language would be longer than those in the other, and vice versa. Clearly, this correlation would not be expected for sentence lengths.

English sentences are commonly shorter than their co-translation counterparts in other languages, although this was not the case when compared with the examples of the Arabic sentences which are reported below in this section. However, this should not be the case for compression code lengths if our intuition about the correlation between information in each language pair is correct. If the compression code lengths are found not to correlate, then the reason for this is more likely to be the result of a less effective compression algorithm being used for one language, resulting in a less accurate estimate of the information contained in the sentence.

In a preliminary experiment, 20 sample sentence pairs in Arabic and English were randomly chosen from the test Corpus A. This sample is shown in Table 5.2. The experiment consisted of compressing each sentence using the PPM compression scheme and calculating the compression code length ratio (CR), as explained in section 5.3, with the aim of measuring the information in each of the sample sentences to be tested.[1]

[1]Some small problems were noted with the English translations of sentences identified here as 9 and 10 as follows: ID 9. You will have something new to listen to it = You will have something new to listen to. ID 10. The militaries are not more persistent on the civil, democratic and secular state. = The military are no longer persistent on the civil, democratic and secular state.

ID	Arabic Sentences	English Sentences
1	موضوعي اليوم جديٌ ولكن أبدأه بطرفة قديمة إستدراجاً للقارئ.	My topic today is a serious one, but I will begin with an old anecdote, to lure the reader in.
2	الوقوف في الجانب الصحيح من التاريخ محاولة لتبرير الحروب العادلة.	Standing on the right side of history represents an attempt to justify just wars.
3	كنت أهاذرها إلا أنها فكرت، وسألتني هل أعتقد حقاً أن البكاء وسيلة أفضل لكسب الأصوات.	I was joking with her, but she took it seriously and asked me whether I really believed that crying was a better way to win votes.
4	هكذا الدنيا، جْنازِة أو جْوازِة كما يقول اللبنانيون.	Such is life, a wedding or a funeral, as the Lebanese say.
5	هذا الرجل يقول: إنه يعرف ما لا يعرف قضاة لجنة الانتخابات.	This man is saying that he knows something the judges on the Election Commission do not know.
6	فلندّعِ مجدداً ريادتنا الربيع، ونحصي الخيبات، ومرارات صيفٍ بائس.	So let us once again claim to be the precursors of the spring, count the disappointments and tally the bitterness of a wretched summer.
7	وأن الذين توجهوا بعدها إلى القصر تصوروا أن الرجل يجلس خلفه في انتظارهم!	Those who subsequently headed to the palace, truly imagined that the man was sitting there, waiting for them!
8	هو أخيراً ارتاح، بعد رحلة الآلام والآمال والنكبات والانتصارات، وترك لنا جميعاً مثلاً يُحتذى.	He has finally rested, after a journey of pains, hopes, disasters and triumphs, and left us all an example to be followed.
9	سيكون هناك شيء جديد تسمعه.	You will have something new to listen to it.
10	العسكريون أكثر تمسكاً بالدولة المدنية الديموقراطية والعلمانية.	The militaries are not more persistent on the civil, democratic and secular state.
11	ويجري دعم النظام الكامل لمجموعة الأدوات بالنسبة لمستعمليه.	The full tool pack system is being supported for those using it.
12	واللجنة مدعوة إلى تقديم تعقيبات على هذه المسائل.	The Commission is invited to provide input on those issues.
13	ويمكن قصر عملية الجمع على المدن الرئيسية، إذا توفرت بيانات لاستقراء الأسعار على المستويات الوطنية.	Collection can be limited to major cities if data are available to extrapolate prices to national levels.
14	أن تشارك البلدان في عملية استعراض البيانات على الصعيد الإقليمي.	Countries are to be engaged in the regional data review process.
15	ولن تستخدم البيانات إلا في تحليل البيانات من جانب المكتب العالمي.	The data would be used only for data analysis by the Global Office.
16	ومن المسلم به أن قاعدة البيانات النهائية ستحتوي على بيانات أكثر تفصيلا من البيانات التي ستنشر.	It is recognized that the final database will contain data in greater detail than that to be published.
17	وتشعر مناطق كثيرة أخرى أن أصحاب المصلحة المعنيين بها سيطلبون بيانات أكثر تفصيلا.	Many of the other regions feel their stakeholders will require data in much more detail.
18	وأعدت استمارات جمع البيانات وأرسلت إلى المناطق.	Data collection forms have been prepared and sent to the regions.
19	وأجرى واحد وعشرون بلدا حساباتها لمتوسطات الأسعار للربع الأول.	Twenty-one countries computed average prices for the first quarter.
20	وتستخدم المنطقة نظام المكتب الإحصائي للجماعات الأوروبية بدلا من مجموعة الأدوات.	The region is using the Eurostat system instead of the tool pack.

TABLE 5.2: The 20 sample sentence pairs from Test Corpus A used in compression experiment one.

The results of three experiments which applied the order 5 PPMD (PPMD5) compression code (i.e. using a fixed order context of length 5) to these sample sentences (1-20) are presented in Table 5.3. For each of the sample sentences (1-20) the table lists the sentence length (Sent. Length) for both Arabic and English characters (bytes). It then provides the number of bytes produced as a compressed output by running three different variants of the PPMD5 compressor. The first of these is Without Training (WOT), meaning that in this initial variant no prior priming of the text was carried out before applying PPMD5 compression. In the second case, the With Training (WT) variant refers to PPM compression with text priming. For this experiment, the PPM model was trained prior to compressing the text, using the Brown Corpus for the English text and the Corpus of Contemporary Arabic for the Arabic text. Finally, the third variant, With Training and Pre-Processing (WTPP), used the same initial priming approach as for the WT variant, but then, in addition, it adopted a pre-processing algorithm to convert the UTF-8 encoded Arabic text into a number string before the PPMD5 compressor was applied. This approach is described in detail in Alhawiti (2014).

Considering the results of three variants, for the Arabic/English sentence pair referred to here as ID 1 (i.e. the first sample sentence in Table 5.2), the WOT variant for PPMD5 required 69 bytes to compress the Arabic sentence, and a similar number to compress its English equivalent. In contrast, the sentence lengths are very different – the Arabic sentence numbering 59 characters (bytes) as opposed to 95 for its English counterpart. This variant appears to produce significantly better compression in the case of the Arabic text and therefore leads to a better estimate of the information contained in the Arabic sentence.

Results for the three experiments, shown in Table 5.3, show that there is a clear mismatch between the sentence lengths for the two languages, and as might be expected given their differences in grammatical and syntactical structures, the English sentence length is longer than that for Arabic in all cases. This provides clear evidence that metrics based on well-established techniques in Information Theory, such as compression code length-based metrics, have merit since they lead to better correlation.

The WOT variant does a surprisingly good job of matching the sentences, with the Arabic byte size being closer to that for English. In this experiment, the number of bytes of the compressed English output is greater than the number of bytes of the compressed Arabic output in only six cases (ID 3, 5, 6, 7, 9 and 18) and is

ID	Sent. Length		PPMD5 Codelength					
			WOT		WT		WTPP	
	Arabic	English	Arabic	English	Arabic	English	Arabic	English
1	59	95	69	69	32	29	26	29
2	68	82	62	62	31	22	24	22
3	84	132	83	88	41	35	31	35
4	53	59	55	50	32	19	27	19
5	58	94	59	64	25	24	20	24
6	67	136	70	93	39	37	31	37
7	72	110	72	76	33	30	26	30
8	93	123	87	86	43	37	35	37
9	27	45	36	38	13	14	11	14
10	63	83	61	61	28	23	20	23
11	59	65	59	53	173	20	18	20
12	49	60	55	45	140	16	16	16
13	99	106	89	73	282	27	31	27
14	64	65	63	52	183	15	17	15
15	66	68	58	54	188	18	20	18
16	95	104	81	70	269	27	27	27
17	81	89	81	65	230	23	25	23
18	48	66	51	53	136	16	17	16
19	62	68	61	55	175	18	25	18
20	79	65	77	52	231	20	22	20

TABLE 5.3: Compression results for the sample sentences after running three variants (WOT, WT, and WTPP) of the PPMD5 compression code.

of equal size to its Arabic counterpart in a further three cases (ID 1, 2 and 10). For the WT variant, the opposite story is now the case – the size in bytes of the compressed Arabic is greater than the compressed English equivalent in all but one case (ID 9) and the differences in size are also much greater, for example in the case of ID13, 282 bytes for the Arabic as opposed to just 27 for the English.

This indicates that the compression method being used for the Arabic text is probably not as accurate as is the case for its English equivalent given that, as Teahan (1998) notes, the use of PPM for compressing text in the latter language has been fine-tuned over the course of many years. Alhawiti (2014) aimed to address this problem in his research on the compression of Arabic text and found that using pre-processing techniques significantly improves PPM-based compression for Arabic, in many cases by over 25%. In the set of experiments carried out for this study, when these pre-processing techniques were applied as the WTPP variant, then a more mixed set of results was achieved for the sample of language pairings, with nine higher compression code lengths for Arabic, nine for English, and two of equal length.

5.4.2 Compression experiment two

In order to investigate these results further and to confirm whether a compression method for Arabic text produces compression code lengths that correlate well with those produced by the compression method for English text, a further series of experiments were conducted using the WTPP variant of PPMD5. This time, this compression code was applied to the whole of the parallel Arabic/English Test Corpus A, and the results of this experiment are shown in Table 5.4, listed separately for each of the 13 subject categories (column 1) into which texts had been allocated, with an average also provided for the purposes of comparison.

Results shown in column 2 show as a percentage the amount of sentence pairs for which the Arabic sentence length is greater than that of their English counterpart. The highest percentage proved to be 35.43% for the Cinema category with the lowest recorded being only 7.93% for text which was categorised as Decisions. These two results compared with an overall average of 16.56% covering all 13 subject categories.

By contrast, the results in the third column, which lists the percentage of sentence pairs for which the Arabic compression code length is greater than that for their English counterparts, show a very different pattern. The comparison is now more even, with the average being 55.14%, and most of the results being clustered around this mark. The only higher exception was the Crimes category (62.48%) and two lower exceptions were Stories and Cinema at 48.79% and 44.94% respectively.

Producing results by subject category in this way serves to highlight areas which may merit further investigation. Differences of this kind may point to specific linguistic phenomena, generic or stylistic conventions which vary significantly in Arabic and English texts and may help to explain this type of discrepancy within a particular subject category.[2] For example, American journalist Joseph Braude (2011: 31), writing with reference to newspaper reporting of violent crime in the

[2]Although the texts in the corpus are likely to be journalistic accounts of crimes, Tahani Alghureiby's attempt to account for the lack of interest in Crime Fiction as a genre in the Arabic-speaking world also offers some interesting insights into the factors, amongst them linguistic, which she claims have limited the spread of this genre amongst Arabs despite its immense popularity amongst a western audience. See Tahani Alghureiby 'The Curious Case of Crime Fiction in Arabic Literature', Arab World English Journal, 4 (May), 2015, 155-166.

Moroccan press, notes a number of significant differences in the framing of these articles which he relates specifically to Arab cultural issues. This also illustrate the value of parallel corpora for those pursuing linguistic research.

Categories	**% of Arabic sentence lengths that are greater**	**% of Arabic compression code lengths that are greater**
Books	8.55	54.96
Business	16.58	56.93
Cinema	35.43	44.94
Conferences	17.09	56.26
Crimes	21.88	62.48
Decisions	7.93	52.74
Economy	16.80	57.02
Geographies	14.79	58.97
Issues	16.30	53.71
Law	15.40	56.73
Politics	16.23	55.25
Reports	16.53	58.06
Stories	11.77	48.79
Average	16.56	55.14

TABLE 5.4: Comparison of Arabic sentence lengths (%) or compression code lengths (%) that are greater than their English sentence equivalents in Test Corpus A.

5.5 Analysing the quality of translations in Test Corpus B

In an initial experiment to estimate the quality of the parallel test corpus, 10,000 translations judged satisfactory and 2,000 translations judged unsatisfactory were manually selected from parallel Test Corpus A in order to form a second Test Corpus (Test Corpus B) which was then used to form the validated reference data for the experiments relating to quality.

Using the categories of satisfactory and unsatisfactory, Table 5.5 compares the sentence length and compression code length of Arabic sentences to their English counterparts. Results show that in Test Corpus B for those examples judged to be satisfactory, the percentage of Arabic sentence lengths that are greater than the sentence lengths of their English equivalent was 9.62% whilst in the unsatisfactory

category this rose dramatically to 83.25%. On the other hand, when this comparative analysis was carried out for Arabic compression code length, whilst the unsatisfactory category at 83.20% closely resembled the result previously obtained for sentence length, in the satisfactory category for compression code lengths the 48.22% obtained was considerably larger than the previous result comparing Arabic and English sentence length.

Categories	**% of Arabic sentence lengths that are greater**	**% of Arabic compression code lengths that are greater**
satisfactory	9.62	48.22
unsatisfactory	83.25	83.20

TABLE 5.5: Comparative analysis of Arabic/English sentence lengths and compression code lengths for samples judged satisfactory and unsatisfactory in Test Corpus B.

In an experiment to estimate the quality of the parallel test corpus, 10,000 translations judged satisfactory and 2,000 translations (Corpus B) judged unsatisfactory were manually selected from parallel test Corpus A, and checking by Arabic linguists.

To determine the difference between satisfactory and unsatisfactory, various thresholds were applied to the data, firstly focusing solely on the SLR, then applying the same threshold to the CR alone. When the calculated distance metric exceeded the threshold values which had been set, then the equivalence of the Arabic/English sentence pair was judged to be unsatisfactory. All other sentence pairs were placed in the satisfactory category.

The results of accuracy for the filtering process against the validated reference data is shown in Table 5.6. The table shows the threshold values that were originally set and applied to the SLR calculations (column 1). The accuracy results (the percentage of correct satisfactory/unsatisfactory classifications made by SLR or CR) are then provided in the subsequent columns. These have been presented as satisfactory and unsatisfactory translations of sentence pairs, with the average results for both SLR and CR provided in the final columns.

Thus Table 5.6 shows, for example, that a SLR with a threshold of 2.50 or higher is able to accurately classify 100% of the sample of satisfactory translations whereas the threshold for this to occur in the case of CR needs to be set at 2.25. With regards to the smaller sample of unsatisfactory translations, 100% of these will

be identified using SLR if the threshold is set at 1.5 or less, whereas at the same low threshold setting of 1.25 the highest accuracy which could be obtained for CR was 97.45%, which would mean that most sentence pairs would be rejected. The calculation that results in producing the highest accuracy for all sentence pairs (whether satisfactory or unsatisfactory) occurs when a threshold of 2.25 is applied to SLR settings.

Threshold	**10000 satisfactory translations**		**2000 unsatisfactory translations**		**Average**	
Values	**SLR**	**CR**	**SLR**	**CR**	**SLR**	**CR**
1.25	20.29%	89.91%	100%	97.45%	60.15%	93.68%
1.50	62.35%	97.86%	100%	78.05%	81.18%	87.96%
1.75	88.15%	99.24%	99.95%	43.40%	94.05%	71.32%
2.00	96.5%	99.76%	99.35%	24.85%	97.93%	62.31%
2.25	98.90%	100%	98.45%	15.00%	98.68%	57.50%
2.50	100%	100%	97.20%	11.40%	98.60%	55.70%
2.75	100%	100%	70.25%	7.35%	85.13%	53.68%
3.00	100%	100%	50.40%	4.95%	75.20%	52.48%
3.25	100%	100%	31.85%	3.30%	65.93%	51.65%
3.50	100%	100%	19.85%	2.15%	59.93%	51.08%

TABLE 5.6: Comparison of translation accuracy for a range of threshold values

Figure 5.1 represents the tendencies of the sample of satisfactory versus unsatisfactory translations for the Arabic/English sentence pairs in Test Corpus B, showing the effect of applying the range of different threshold values for SLR.

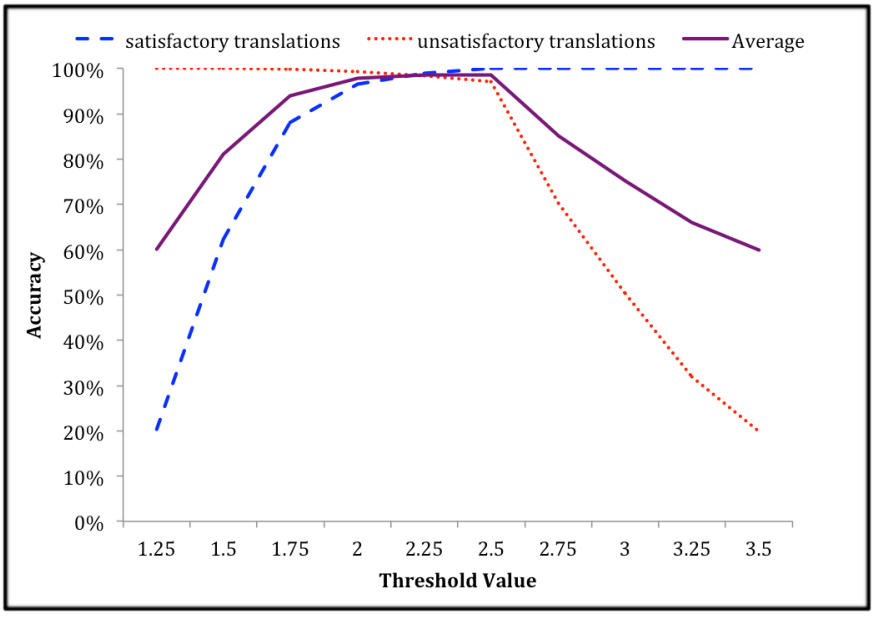

FIGURE 5.1: Satisfactory versus unsatisfactory trends for translation accuracy in Test Corpus B after applying a range of threshold values for SLR.

5.6 Analysing translations rejected as unsatisfactory

In order to ensure the best possible alignment at sentence level between the language pairs in the corpora, the CR threshold value should be used to filter out all those sentence pairs judged to be unsatisfactory. In theory, the lower the value of the threshold is, the higher the corpus quality will be. However, this presents something of a problem given that there are still many unsatisfactory sentence pairs in Corpus A because if the threshold value is set too low, large numbers of sentences pairs will be eliminated from the parallel corpora, meaning that there will be less material available to support training for machine translation purposes.

Therefore, in order to develop a better corpus, it is important at this stage to understand why some translations have been categorised as unsatisfactory as a result of the translation process. Table 5.7 illustrates four examples which were deemed to be unsatisfactory translations because a single sentence in Arabic needs to be aligned with two English ones to create equivalence.

Three examples illustrating the reverse situation with Arabic/English sentence pairs judged unsatisfactory are presented in Table 5.8 since in this case a single English sentence needs to be aligned with two Arabic sentences to achieve linguistic equivalence in translation.

Arabic		English	
Sentence Length	**One Sentence**	**Sentences Length**	**Two Sentences**
86	فالمجلس يدرس حالياً اصدار اعلان دستوري جديد، ما يعني ان الاستفتاء الدستوري جاء ناقصاً.	98	The council is currently looking into the possibility of making a new constitutional announcement.
		62	This implies that the constitutional referendum was deficient.
102	مضى يوم اعتقدت فيه أن سورية ستكون سنغافورة الشرق الأوسط وأجدها اليوم البوسنة 1995، والحولة سريبرنيتسا.	101	The day I believed that Syria could become the Singapore of the Middle East has long since been gone.
		111	Today, I find Syria to be like the Bosnia of 1995, and the Houla massacre to be the equivalent of Srebrenica's.
86	كما يدرس تشكيل الهيئة التأسيسية لوضع دستور دائم بعدما كان البرلمان شكل مثل هذه الهيئة.	129	In addition, the council is looking into the formation of a founding committee in order to come up with a permanent constitution.
		54	The parliament had already formed a similar committee.
77	حتى البرلمان نفسه الذي شغلت انتخاباته البلاد على مدى شهور جرى حله بحكم قضائي.	77	The entire country had been busy for months with the parliamentary elections.
		82	However, this resulting parliament has been dissolved through a legislative order.

TABLE 5.7: Four examples of translations deemed unsatisfactory because a single Arabic sentence needs to be aligned with two English sentences to produce equivalence.

5.7 Discussion and Summary

On one level it is perhaps only to be expected that examples are found of one-to-two and two-to-one Arabic/English translations as Emery (1991:130) notes that is it possible to find considerable variation in Arabic discourse depending on the context and the type of discourse: "the tension between prolixity and brevity continues to run through Arabic writing. On the one hand, Arabic discourse is often characterised as having an abundance of synonymy [...] on the other hand, Arabic may exhibit [...] an austerity and economy unmatched in English". However, when the examples listed in Tables 5.7 and 5.8 are reviewed, certain textual features emerge which cast further light on the mismatch which has led to the translations being judged as unsatisfactory. In the case of the first English example, the division into two sentences could have been avoided by using the type of relative clause which would usually be employed in English, transforming two sentences into one:

Arabic		English	
Sentence Length	**Two Sentences**	**Sentence Length**	**One Sentence**
173	ويشمل ما ترتب إجمالا عن هذه العملية على الباب 28 واو نقل ثلاثة وظائف (وظيفة واحدة بالرتبة ف-3 وواحدة بالرتبة ف-2/1 وواحدة بفئة الخدمات العامة (الرتبة الرئيسية)) من الباب 16.	317	Its net effect on section 28F comprises a redeployment of three posts (1 P-3, 1 P-2/1 and 1 General Service (Principal level)) from section 16, United Nations Office on Drugs and Crime, to this section (totalling $606,400), together with related functions previously exercised by the administrative services of UNODC.
190	مكتب الأمم المتحدة المعني بالمخدرات والجريمة، إلى هذا الباب (تبلغ إجمالا 400 606 دولار)، إلى جانب مهام ذات صلة كانت تتولاها من قبل الدوائر الإدارية التابعة للمكتب المعني بالمخدرات والجريمة.		
144	فعلى سبيل المثال، فإنه بالنسبة للمسائل التي من المتوقع أن تستخلص فيها الاستقصاءات المكتوبة التي يشارك فيها مسؤولون حكوميون ردودا ذاتية أو رسمية.	251	For example, for issues where written surveys of government officials may be expected to elicit a subjective or official response, alternative means such as informal interviews, surveys of citizen groups, or document analysis may be more insightful.
134	قد تكون وسائل بديلة من قبيل اللقاءات غير الرسمية والاستقصاءات التي يشارك فيها جموع المواطنين أو تحليل للوثائق أنفذ بصيرة في هذا الصدد.		
159	والمبادئ التوجيهية للاتفاقية الإطارية فيما يتصل باستعراض قوائم الجرد، وهي المبادئ التي اعتمدت في عام 1999 (المقرر 3/م أ-5) ونقحت في عام 2002 (المقرر 19/م أ-8).	183	UNFCCC review guidelines adopted in 1999 (decision 3/CP.5) and revised in 2002 (decision 19/CP.8) help to ensure that reviews are conducted consistently in a technically sound manner.
85	تساعد المبادئ على ضمان إجراء عمليات الاستعراض بطريقة متسقة وسليمة من الناحية التقنية.		

TABLE 5.8: Three examples of translations deemed unsatisfactory because a single English sentence needs to be aligned with two Arabic sentences to produce equivalence.

Sentence 1: *The council is currently looking into the possibility of making a new constitutional announcement.*

Sentence 2: *This implies that the constitutional referendum was deficient.*

The two separate sentences could have been rendered as one single sentence without any loss of meaning:

The council is currently looking into the possibility of making a new constitutional announcement which implies that the constitutional referendum was deficient.

Although this makes little if any difference to the sentence length in characters, it would produce a sentence-to-sentence alignment. Similarly, in example four in

the same table, the use of a different conjunction could have produced a single sentence:

Sentence 1: *The entire country had been busy for months with the parliamentary elections.*

Sentence 2: *However, this resulting parliament has been dissolved through a legislative order.*

These could have been rewritten as:

Although the entire country had been busy for months with the parliamentary elections, the resulting parliament has been dissolved through a legislative order.

However, the English in all four examples is not glaringly inadequate or an inaccurate rendering of the Arabic and does not seem to present any particular irregularities or represent any particular genre. On a purely linguistic level, then, these texts could be judged satisfactory but simply fail to align at the sentence level.

However it is striking in the case of the three examples in Table 5.8 in which one English sentence is translated into two Arabic sentences, that these sentences share some common features. Firstly, all of them are lengthy at 317, 251 and 183 characters respectively. Secondly, they all consist of multiple clauses, some of which are themselves embedded within other clauses. In examples one and three, the complexity is enhanced by the use of parenthetical clauses. Example one has a particularly convoluted sentence structure which has double parentheses inside an already over long and poorly structured sentence:

Its net effect on section 28F comprises a redeployment of three posts (1 P-3, 1 P-2/1 and 1 General Service (Principal level)) from section 16, United Nations Office on Drugs and Crime, to this section (totalling $606,400), together with related functions previously exercised by the administrative services of UNODC.

It is noticeable that in two of these cases the texts are clearly produced by agencies of the United Nations and can perhaps be considered part of a phenomenon which Mohammed Bagher Roozgar (2008) refers to as 'hybrid' texts. This type of writing in English is normally used within the multinational, multilingual environments of organizations such the United Nations or the European Union. Texts of this kind tend to have a very specific in-house audience and are aimed solely at the

staff of the organization or for meetings of particular bodies within it. He notes that they pose a unique set of challenges for the translator. Clearly in these cases the Arabic translator struggled to make sense of their complexities.

The examples in Tables 5.7 and 5.8 both have implications in terms of the material which is included in parallel corpora. In the first set of examples, the two-to-one equivalence poses problems with sentence alignment although the language and the translation is clearly satisfactory. In the second set of examples, although the English appears to be the original source language, it cannot help but produce unsatisfactory parallel texts due to its specific nature, indicating that perhaps a further category is needed within corpora for this new type of hybrid English text.

The other general point that this analysis highlights is the crucial role to be played by the skilled linguist in verifying the satisfactory or unsatisfactory nature of parallel texts. However, other more basic verification tasks which have not been described here also need to be carried out to improve the accuracy of the parallel corpora although they are often overlooked. These include for example, a check on document sizes. It is crucial for example to ensure there are no zero byte documents, and also to remove any unusually large documents, if appropriate.

This chapter has described the experimental methods which were used to firstly analyse the parallel Arabic/English Test Corpus A and then verify the degree of accuracy of its content. The process of verification is essential in order to ensure that the quality of the corpus is as high as possible. One of the most useful types of verification task which can be conducted on parallel corpora which have been manually-aligned like these is to ensure that the texts are correctly aligned at the level of whole documents, paragraphs and sentences.

A method to improve quality was developed based on the combination of two distance metrics, sentence length ratio (SLR) and compression code length ratio (CR). A threshold mechanism which determines when either the SLR or CR values have been exceeded can be used for the purposes of attempting to identify unsatisfactory translations. Experiments with a small sample of sentence pairs from the parallel Test Corpus A, manually judged to be satisfactory or unsatisfactory translations, also shows that a combination of both SLR and CR distance metrics performs better than using a single distance metric by itself.

Chapter 6

A New hybrid method for sentence alignment

The purpose of this chapter is to discuss a new hybrid method for sentence alignment that was created in order to improve the overall quality of the corpus that was discussed in Chapter Four (Test Corpus A). The chapter begins by introducing the development process for the new metric for sentence alignment. The second section describes the process of creating a hybrid of a sentence-length and the new PPM compression-based sentence alignment metric. Next, the focus shifts to discussing how to program the code length ratio distance metric for the purpose of sentence matching. This is also compared with data in the other metric for sentence matching, the sentence length ratio. Finally, the chapter presents the results of the experiment in graphical form.

A co-authored paper based on this chapter was presented at the Fifth International Conference on Computer Science, Engineering and Applications in Dubai in January 2015, entitled "A new hybrid metric for verifying parallel corpora of Arabic/English".

6.1 A new hybrid method for filtering parallel corpora

Previous research has agreed that sentence-based parallel corpora represent important resources. There are three main classifications, namely sentence length-based,

lexical-based and a combination of these two (Kay and Röscheisen, 1993). Brown et al. (1991) were the first to suggest that sentence length could be used to align corpus for the purposes of MT. Based on Brown's study, Gale and Church (1993) discussed the possibility of many-to-one or even many-to-many sentence alignment by introducing lexical information. Subsequent research on parallel corpora and multiple corpora also discussed various way of combining the traditional sentence length method with lexical methods.

As discussed in previous chapters, a possible solution to this low-accuracy limitation is to improve the algorithm and the metric that are used in a sentence alignment process by a compression-based one such as PPMD5. After validating the reliability of the new PPMD algorithm and the positive results of preliminary experiments with Test Corpus A, in the previous chapter, a new method has been developed for aligning sentences. The new hybrid method leverages the advantages of both sentence length method and code length ratio method and, a series of experiments have been conducted with Corpus B to identify the most appropriate threshold values. Finally, the result is presented as a filtering process in a flowchart to describe how to conduct the sentence alignment process in steps.

This chapter describes an experiment to improve the quality of parallel Test Corpus A. Experiments were performed using the validated reference data in Corpus B to determine the best thresholds and combinations for the CR and SLR metrics in order to accurately filter out the unsatisfactory translations. For the CR calculations listed there, the WTPP with PPMD5 variant (which was primed on the CCA corpus) was used to compress the Arabic text sentences, whereas standard PPMD5 (primed on the Brown corpus) was used for the English text sentences.

On the basis of the insights gained from the PPM compression experiments described in the previous chapter, this section discusses two sentence quality checking metrics that include the CR distance metric and SLR distance metric. Understanding the principles of both metrics helped the researcher to verify the quality of new parallel corpora.

A range of thresholds were applied firstly using only SLR, secondly using only CR, and thirdly by applying the same threshold to both SLR and CR together. If the distance metric calculated exceeded the threshold value(s), then the translation sentence pair was judged to be unsatisfactory; if not it was judged to be satisfactory.

The accuracy of the filtering process was compared to the validated reference data for test Corpus B and the results are shown in Table 6.1. Threshold values that were used for both the SLR and CR calculations are shown in the first column. The accuracy results are then provided in the subsequent columns. (This is the percentage of correct classifications made by the SLR, CR or SLR+CR metrics where a correct classification is made when the metric at a specific threshold judges the sentence pair to be satisfactory or unsatisfactory and this matches the validated reference judgment). The results are split into the satisfactory and unsatisfactory sentence pairs, with the average results provided in the final columns.

Table 6.1 shows, for example, that SLR with a threshold of 2.5 or higher is able to accurately classify 100% of the satisfactory translations whereas the threshold where this occurs for CR is 2.25. For the unsatisfactory translations, 100% of these will be identified using SLR if the threshold is set at 1.5 or less, whereas the highest accuracy for CR is 97.45% when the threshold is set as low as 1.25 (meaning most sentence pairs will be rejected). The only calculation that results in an average accuracy of 100% for all sentence pairs (both satisfactory and unsatisfactory) occurs when both SLR and CR are combined together with a threshold of 2.5.

Figure 6.1 shows the tendencies for the classification of the satisfactory and unsatisfactory translations for test Corpus B using different threshold values when using both SLR CR. (This compares with Figure 5.1 when only a single metric and threshold for SLR was used).

Threshold	**10000 satisfactory translations**			**2000 unsatisfactory translations**			**Average**		
Values	**SLR**	**CR**	**SLR&CR**	**SLR**	**CR**	**SLR&CR**	**SLR**	**CR**	**SLR&CR**
1.25	20.29%	89.91%	17.11%	100%	97.45%	100%	60.15%	93.68%	58.56%
1.50	62.35%	97.86%	61.10%	100%	78.05%	100%	81.18%	87.96%	80.55%
1.75	88.15%	99.24%	87.58%	99.95%	43.40%	100%	94.05%	71.32%	93.79%
2.00	96.5%	99.76%	96.3%	99.35%	24.85%	100%	97.93%	62.31%	98.15%
2.25	98.90%	100%	98.90%	98.45%	15.00%	100%	98.68%	57.50%	99.45%
2.50	100%	100%	100%	97.20%	11.40%	100%	98.60%	55.70%	100%
2.75	100%	100%	100%	70.25%	7.35%	72.30%	85.13%	53.68%	86.15%
3.00	100%	100%	100%	50.40%	4.95%	51.80%	75.20%	52.48%	75.90%
3.25	100%	100%	100%	31.85%	3.30%	32.80%	65.93%	51.65%	66.40%
3.50	100%	100%	100%	19.85%	2.15%	20.35%	59.93%	51.08%	60.18%

TABLE 6.1: Comparison of accuracies for different threshold values when using the different sentence matching metrics on test Corpus B.

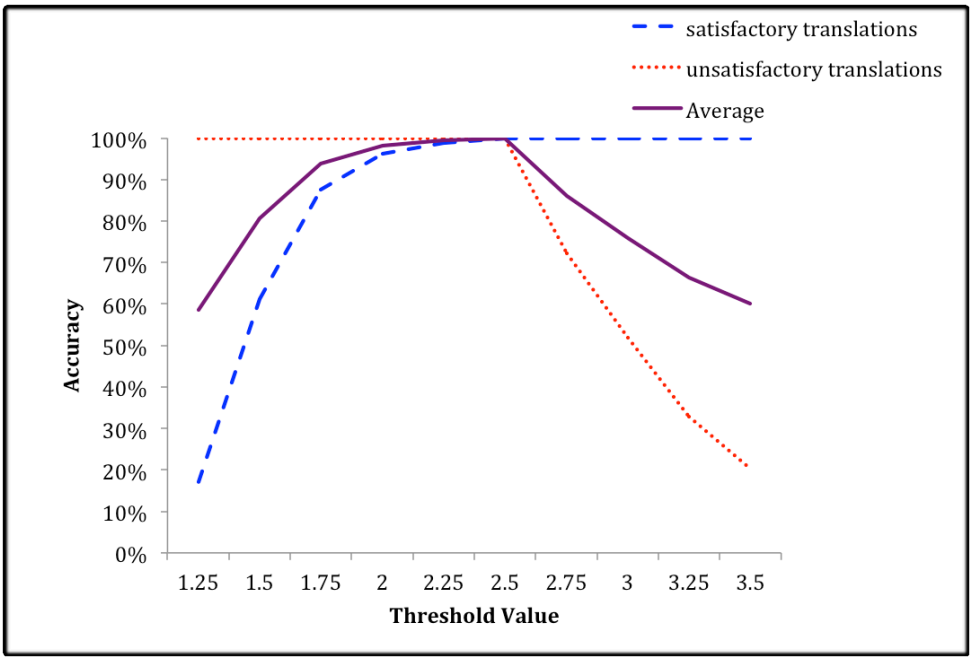

FIGURE 6.1: Tendencies of satisfactory and unsatisfactory translations for test Corpus B with different threshold values for SLR&CR.

CR\SLR	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
1.25	64.39%	86.73%	95.39%	96.6%	96.73%	96.75%	95.98%	95.58%	95.53%	95.48%
1.50	65.76%	88.24%	97.39%	99.3%	99.65%	99.76%	94.29%	91.34%	89.64%	89.11%
1.75	65.79%	88.26%	97.44%	99.39%	99.8%	99.96%	89.64%	82.69%	77.36%	74.49%
2.00	65.79%	88.28%	97.46%	99.41%	99.83%	99.99%	88.64%	80.04%	72.71%	68.09%
2.25	65.79%	88.28%	97.48%	99.43%	99.84%	**100%**	87.95%	78.9%	70.53%	65.08%
2.50	65.79%	88.28%	97.48%	99.43%	99.84%	**100%**	87.35%	77.98%	69.3%	63.78%
2.75	65.79%	88.28%	97.45%	99.2%	99.31%	99.18%	86.15%	76.58%	67.68%	62.03%
3.00	65.79%	88.28%	97.45%	99.13%	99.14%	98.75%	85.58%	75.9%	66.85%	61.05%
3.25	65.79%	88.28%	97.45%	99.1%	99.06%	98.63%	85.33%	75.6%	66.4%	60.55%
3.50	65.79%	88.28%	97.45%	99.1%	99.06%	98.6%	85.15%	75.33%	66.1%	60.18%

TABLE 6.2: The matrix for threshold values of SLR and CR from 1.25 to 3.5.

A further experiment was conducted to investigate whether different threshold values are more effective when using the combined SLR&CR technique. Table 6.2 displays the results of the experiment on the overall accuracy averages on the same sample (10,000 satisfactory translations and 2,000 unsatisfactory translations) used in the previous experiment. Table 6.2, the SLR threshold value is shown across the top row, and the CR threshold value is shown down the left column, both ranging from 1.25 to 3.50. The table shows that 100% accuracy is achieved using threshold values of 2.50 for SLR and 2.25 or 2.50 for CR.

Another experiment was devised to determine how much of the parallel Test Corpus A would be classified as satisfactory or unsatisfactory using various CR threshold values (from 1.25 to 3.50) when the SLR threshold value was set at 2.5. The

results of this experiment are shown in Table 6.3. The table shows the number classified in each category (in the columns labelled "Amount") and the corresponding percentages for Test Corpus A. For example, a threshold value of 2.50 for both SLR and CR would result in 8.18% of Test Corpus A being labelled unsatisfactory (and therefore these candidates could be removed from Test Corpus A in order to improve its quality).

Threshold	**Satisfactory translations**		**Unsatisfactory translations**	
CR	**Amount**	**Percentage**	**Amount**	**Percentage**
1.25	1,313,387	72.14%	507,234	27.86%
1.50	1,559,275	85.65%	261,346	14.35%
1.75	1,626,973	89.36%	193,648	10.64%
2.00	1,650,374	90.65%	170,247	9.35%
2.25	1,665,709	91.49%	154,912	8.51%
2.50	1,671,768	91.82%	148,853	8.18%
2.75	1,674,677	91.98%	145,944	8.02%
3.00	1,675,700	92.04%	144,921	7.96%
3.25	1,676,166	92.07%	144,455	7.93%
3.50	1,676,311	92.07%	144,310	7.93%

TABLE 6.3: Distribution of satisfactory and unsatisfactory translations for the parallel corpus A when SLR threshold is set at 2.5.

Figures 6.2, 6.3, 6.4 and 6.5 show correlations for the sentence length and code length metrics for Test Corpus A. Figures 6.2 and 6.3 illustrate the sentence lengths and code lengths of Arabic and English sentences classified as unsatisfactory for Test Corpus A and show an obvious split in the plot due to 1:2 and 2:1 type mismatches.

In contrast, Figures 6.4 and 6.5 illustrate sentence length and code length of Arabic and English for the translations classified as satisfactory for Test Corpus A, showing a strong correlation between both sentence lengths and compression code lengths.

For defining what constitutes a satisfactory translation in this case, it was decided if the values of SLR were less than 2.5 and less than 2.25 in the case of CR for a pair of translation sentences, then it is classified as a satisfactory translation; if not, it is classified as an unsatisfactory translation.

The unsatisfactory translations might be caused by errors in alignment between Arabic and English sentences which may include non-literal translations and therefore result in significant differences between the sentence pair. English sentences

containing websites or acronyms such as USA (United States of America), or UNCTAD (United Nations Conference on Trade and Development) might also lead to mistranslations (Khadivi and Ney, 2005).

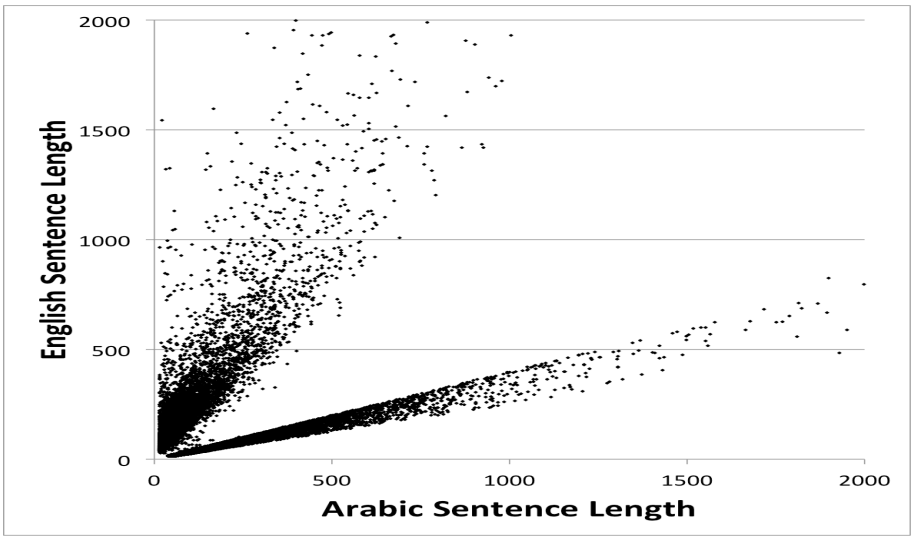

FIGURE 6.2: Sentence length distribution for unsatisfactory translations.

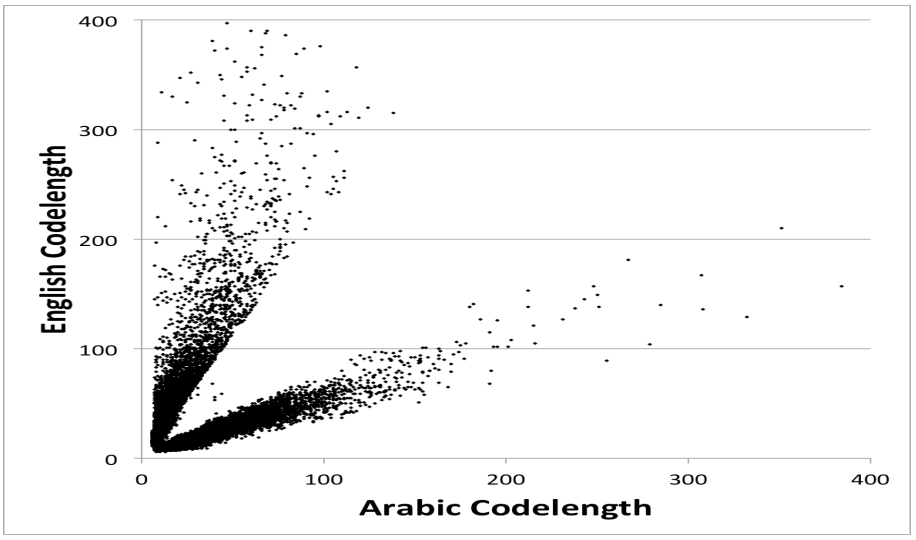

FIGURE 6.3: Code length distribution for unsatisfactory translations.

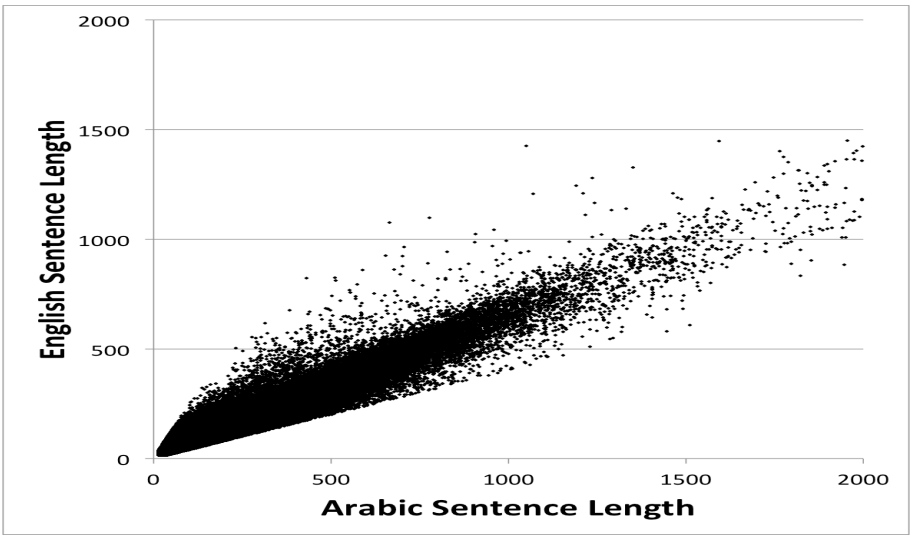

FIGURE 6.4: Sentence length distribution for satisfactory translations.

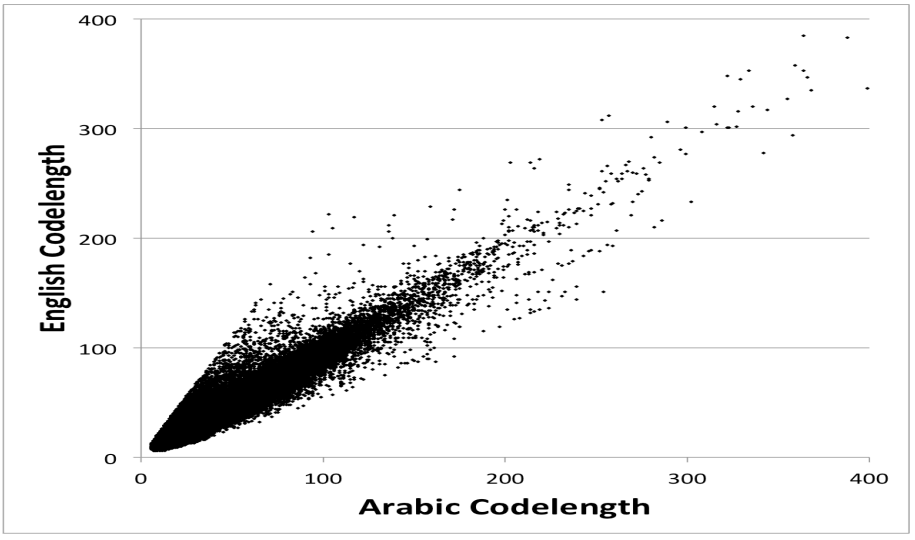

FIGURE 6.5: Code length distribution for satisfactory translations.

6.2 Filtering Process

The process shown in Figure 6.6 shows the flowchart of how parallel Test Corpus A was analysed in this manner. The flowchart shows the separate pathways for processing the English and Arabic sentences in the corpus. The experiment runs the compression test (PPMD5) on both the English and Arabic sentence pairs through the SLR comparison horizontally and the result then goes into the decision point. In the meantime, the experiment runs the comparison again, this time adopting the PPM metric. It encodes the sentence length as a ratio for a second comparison and in order to obtain a CR value. Additionally, the threshold value is also calculated for the purposes of judging whether a translation result should

be classified as "satisfactory" or as "unsatisfactory". If either the CR threshold of 2.25 is exceeded, or the SLR threshold of 2.5 is exceeded then the sentence pair is rejected and then removed from the final parallel corpus.

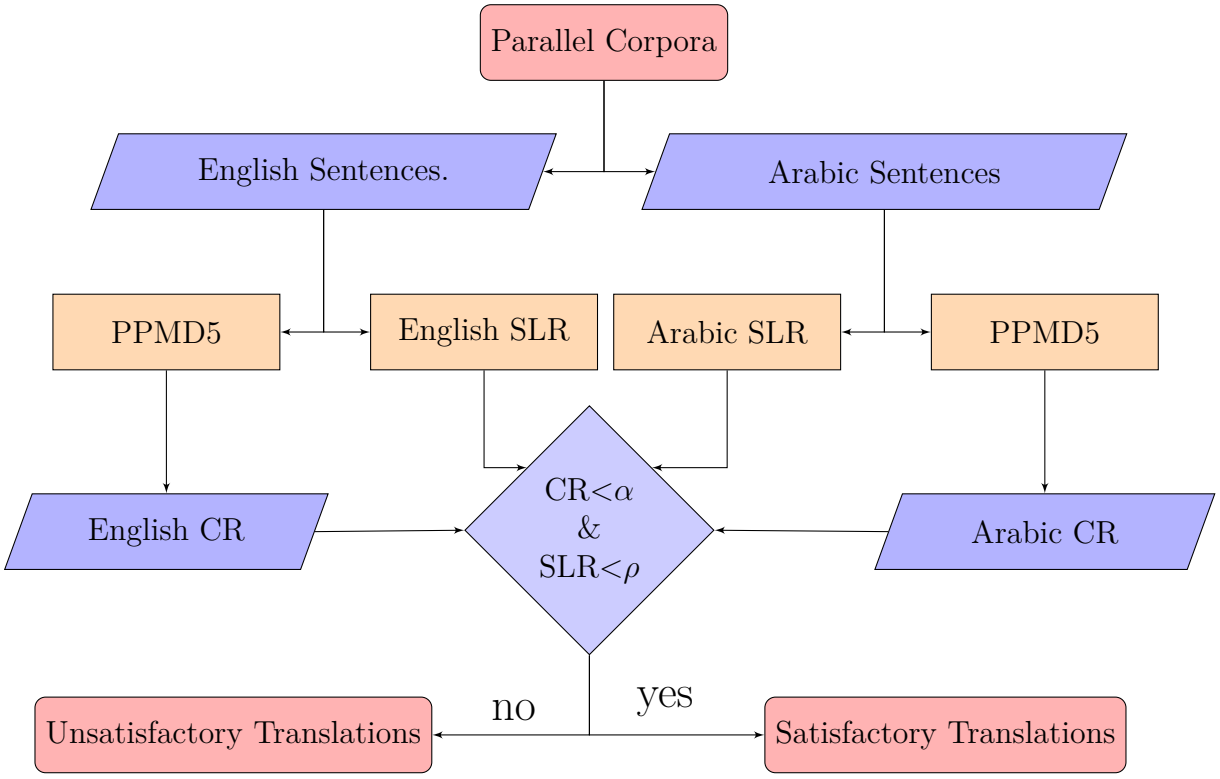

FIGURE 6.6: Flow chart showing how the parallel corpus A was analyzed, best results for $\alpha = 2.25$ and $\rho = 2.5$.

Categories	**Satisfactory**		**Unsatisfactory**	
	Arabic Words	**English Words**	**Arabic Words**	**English Words**
Books	902736	1045722	29100	33977
Business	2184044	2472623	105232	151651
Cinema	5720706	6905019	2199196	1222490
Conferences	1819518	2079142	60009	136715
Crimes	896003	969711	37839	35510
Decisions	1359057	1561628	38124	44223
Economy	2199040	2492437	67384	107214
Geographies	1355636	1550623	36463	44492
Issues	902498	1023844	148697	18472
Law	1352025	1527696	55267	70177
Politics	1856148	2134871	179821	169362
Reports	1337124	1515481	51333	74575
Stories	2823946	3341531	58717	79294
Total	24708481	28620328	3067182	2188152

TABLE 6.4: Word counts for Corpus A by using the new model.

Table 6.4 shows final word counts from Test Corpus A after filtering by using the new method. This table can be compared to Table 4.7 which lists the previous word counts for Test Corpus A. The Cinema category is the largest in either of the categories: satisfactory (Arabic 5,720,706 and English 6,905,019) or unsatisfactory (Arabic 2,199,196 and English 1,222,490). It is followed by several categories that including Stories, Economy and Business in the satisfactory category, while the second largest unsatisfactory category is Business but its size is significantly smaller. A possible reason is that the Cinema or Stories is a composite field that includes a variety of texts such as written documents, legal files and also includes many informal texts. In contrast, when the data is collected from a specific language domain such as Business or Economics the discourse tends to be more specialised and formulaic, making for a greater level of correspondence in the two languages.

Table 6.5 shows how many sentences were deleted (as opposed to words) from each category and the total in Test Corpus A after using the new model. All corpora are divided into thirteen categories, the largest one being Cinema with 2,087,072 corresponding sentences. When the PPM as used to test the Cinema corpus under both CR and SLR conditions 1,784,882 sentences were identified as satisfactory and 302,190 as unsatisfactory. The smallest categories are Crime and Issues with

Categories	**Untreated Sentences**	**Satisfactory Sentences**	**Unsatisfactory Sentences**
Books	57738	55472	2266
Business	187792	180660	7132
Cinema	2087072	1784882	302190
Conferences	151548	146894	4654
Crimes	75088	72528	2560
Decisions	82154	79634	2520
Economy	189394	183500	5894
Geographies	114150	111940	2210
Issues	74526	71524	3002
Law	113512	110506	3006
Politics	153166	147992	5174
Reports	113038	109578	3460
Stories	242064	238784	3280
Total	3641242	3293894	347348

TABLE 6.5: Final sentence counts for Corpus A by using the new filtering method.

75,088 and 74,526 words respectively and the results show that over 96% sentences were classified as satisfactory while only 4% were rejected as unsatisfactory.

6.3 The accuracy and error-rate of the new hybrid method

The accuracy and error-rate of the new hybrid sentence alignment method was determined by manually checking a random sample of 100 sentence pairs from Corpus A classified as "satisfactory" and 100 sentence pairs from Corpus A classified as "unsatisfactory". The accuracy-rate was found to be 99% and the error-rate to be 1% for the final filter at predicting "satisfactory" co-translations and for sentence pairs judged to be "unsatisfactory", the accuracy rate was found to be 96% and the error-rate to be 4%. The errors in these classifications are shown in Tables 6.6 and 6.7.

Arabic Sentences	**English Sentences**	Comments
- الرعاية الاجتماعية، وحماية الطفل والأسرة؛	Social welfare, child protection and family protection; Health care, health insurance;	Error due to there is no trans- lation for the following words: " Health care, health insurance". In this case, SLR = 1.98 < 2.5 and CR = 1.5 < 2.25, therefore the new hybrid alignment method classified this sentence pair to be satisfactory.

TABLE 6.6: The sentence pairs that were erroneously classified in the random sample as satisfactory by the hybrid method.

Arabic Sentences	**English Sentences**	Comments
وتبلغ نسبة فقراء الريف/الحضر في البوسنة والهرسك 16 مقابل 15 في المائة.	In the Federation of Bosnia and Herzegovina the rural/urban ratio stands at 16 to 15 per cent.	SLR = 2.98 > 2.5 and CR = 1.09 < 2.25, therefore the new hybrid alignment method classified this sentence pair to be unsatisfactory.
358- وبعد تنفيذ قوانين الملكية في جمهورية صربسكا والاتحاد، ينبغي تناول وتنظيم قطاع الإسكان، بوصفه مشكلة ملحة للبلد بأسره، من خلال قانون على مستوى الدولة.	After the implementation of property laws in Republika Srpska and the Federation, the area of housing, as an urgent problem for the whole country, should be defined and regulated by the law at the State level.	SLR = 4.69 > 2.5 and CR = 1.05 < 2.25, therefore the new hybrid alignment method classified this sentence pair to be unsatisfactory.
371- ولفرقة الخبراء برنامج أنشطة، وفي الوقت نفسه فإن المشاورات وتبادل الآراء مع الخبراء من ميثاق الاستقرار مستمرة.	The expert team has a programme of activities, while consultations and an exchange of opinions with the experts from Stability Pact are going on.	SLR = 2.86 > 2.5 and CR = 1.03 < 2.25, therefore the new hybrid alignment method classified this setence pair to be unsatisfactory.
411- بلغ متوسط النسبة المئوية للأطفال في مستوى التعليم الابتدائي الذين تم تحصينهم في عام 2000 ضد داء السل 91 في المائة، أي أقل من نسبتهم البالغة 92.5 في عام 2001.	The average percentage of primary children vaccinated with BCG in 2000 was 91 per cent and is falling when compared to 92.5 per cent in 2001.	SLR = 3.5 > 2.5 and CR = 1.19 < 2.25, therefore the new hybrid alignment method classified this sentence pair to be unsatisfactory.

TABLE 6.7: The sentence pairs that were erroneously classified in the random sample as unsatisfactory by the hybrid method.

6.4 Summary and Discussion

This chapter describes how the quality of the newly-developed parallel Test Corpus A for Arabic and English was improved. This chapter has described a new hybrid method of checking the extent to which the sentences match each other in a parallel corpus. The method is based on combining two distance metrics, Sentence Length Ratio (SLR) and Compression code length Ratio (CR). A threshold mechanism can be used to filter out unsatisfactory translations when either the SLR or CR values have been exceeded. Experiments with a small sample of sentence pairs from a test Arabic/English corpus (Corpus B) containing ground truth judgments, which were manually judged to be satisfactory or unsatisfactory translations, show that a combination of both SLR and CR distance metrics is better at classification than a single distance metric by itself.

As explained in the previous chapter, there are also other important verification tasks that are often overlooked, not described here that need to be done. For example, a single check on document sizes is crucial (e.g. ensuring no zero byte documents, and removing unusually large documents if appropriate). Checking for self-plagiarism (ensuring that documents do not contain strings repeated in other documents) is also essential (especially for corpora containing news stories since it is a common practice for these types of documents to contain material copied from other news stories). It has been found that the compression code length metric described here is also effective at classifying the quality of translation not just at the sentence level, but also at the document, paragraph and clause levels, and these should also be checked when verifying a parallel corpus in the future.

Chapter 7

Conclusion and Future Directions

7.1 Introduction

This chapter begins by reviewing the achievements of this research in general. Secondly, it reviews the initial aims and objectives. Next, it highlights the most important results of the research and the contributions which the experimental process has made to the practical field. Finally, it offers a series of suggestions to assist future studies avoid repeating mistakes and instead focus on promising avenues. Issues discussed here might inspire future researchers to find better solutions to improve the quality of the corpora or reduce the cost of deployment further. The personal and professional recommendations also highlight a discussion of the design and development of the new parallel corpora as well as future applications of the research.

7.2 Review of aim and objectives

The main aim of this study was to investigate novel compression-based methods for aligning parallel corpora between Arabic and English. Its secondary aim was to build a high-quality, low-cost parallel corpus for Arabic and English. These aims have been successfully achieved.

Corpus A, the new large-size parallel corpus, has proven to be of high quality in terms of accuracy of translation. This corpus is available for free by contacting AIIA.[1]

All the objectives of this research mentioned in Chapter One were successfully achieved in the research. First of all, a review of prior parallel corpora was completed in Chapter Four, which concluded that the main problem with existing Arabic/English parallel corpora is the comparatively poor level of quality.

Next, a high quality parallel corpus, Corpus A, was produced, as well as a subset corpus, Corpus B. These were created on the basis of texts from the bilingual newspaper website, *Al Hayat*, and OPUS, an open-source online corpus. Details of the development of these corpora are discussed in chapter five.

Thirdly, experiments conducted on the newly developed parallel corpora, have shown that the PPMD alignment methods are effective at identifying the correct threshold level for distinguishing between satisfactory and unsatisfactory sentence pairs in the corpus.

Finally, a hybrid method sentence alignment was developed based on using both the sentence length metric (SLR) and the compression code length metric (CR) and this was used to improve the quality of the new developed Corpus A.

7.3 Review and Conclusions

This research project developed a new compression-based approach for the alignment of parallel corpus for Arabic/English use. The thesis began by introducing the linguistic specificity of the Arabic language and the difficulties it presents for encoding as a natural language which varies very considerably from English in terms of text directionality, graphic representation, syntax and grammar, as detailed in Chapter Two. The development of parallel corpora to assist machine translation, therefore, is seen as an effective way to solve this problem.

In the literature review presented in Chapter Three, various methods of sentence alignment for parallel corpora were discussed include sentence length. The problems with using such methods were highlighted. One key issue is that for dissimilar

[1] Artificial Intelligence and Intelligent Agents research group, email contact: w.j.teahan@bangor.ac.uk.

languages (such as Arabic and English), sentence lengths clearly do not correlate, so the accuracy of translation does not significantly increase. Moreover, it is difficult to access and re-use previously-developed Arabic/English parallel corpora because they are expensive to purchase.

The researcher investigated whether this situation could be remedied by adopting the concept of Prediction by Partial Matching (PPM) for sentence alignment for MT. PPM is a new sentence matching metric that uses the entropy of sentences to determine the accuracy of comparison. Unlike sentence length comparison, the essence of entropy comparison is to measure whether the information matches between sentences.

Following the methods that had been proposed in existing literature, the researcher prepared the raw materials for development of the new parallel corpora. First and foremost, the resources included in new corpora determine their quality and scale. In this project, two online sources, *Al Hayat* and OPUS, were chosen by the researcher to offer a large number of up-to-date text sources in both English and Arabic. *Al Hayat* is a corpus based on a daily international good quality newspaper and its corpus contains large quantities of text and accurate English translations. The researcher finally gathered 2,822 sentences (equal to 203,359 words) from *Al Hayat*. Meanwhile, the other Internet-based open-source corpus also drew the researcher's attention because it contains a larger source than *Al Hayat* although some components within it contained incomplete sentences, informal expressions and complex sentence structures. Eventually, 3,638,420 sentences (equal to 58,380,784 words) were collected and all the data were fetched and transformed into a unified XML format.

The new corpora were also subjected to compression and comparison experiments, the results of which helped the researcher to develop a new hybrid approach to sentence alignment. Over the years, much of the literature regarding parallel corpora development has proposed a number of sentence alignment approaches such as word probability calculation, the highest matching of speech tag and sentence length measurement (Kay and Röscheisen, 1993; Papageorgiou et al., 1994; Simard et al., 1992). However, later studies have shown that there are various disadvantages with those existing sentence alignment approaches, the most significant being low-efficiency. The approach used in this research assumes that information contained in pair translations is similar, and that compression code

length ratio should be a value that trends to 1.0. Similarly, the best threshold values of the new hybrid corpora are identified for both CR and SLR metrics.

Experiments were conducted on a small scale trial of 10,000 translations taken from the collected data, to measure performance of the CR compression code length and SLR metrics. Results showed that the optimal threshold value is 2.25 for CR and 2.5 for SLR approaches respectively.

Thus, for example, SLR with a threshold of 2.5 or higher is able to accurately classify 100% of the satisfactory translations whereas the threshold for this for CR is at 2.25; the highest accuracy for CR is 97.45% when the threshold is set as low as 1.25 (meaning most sentence pairs will be rejected). The only calculation that results in an average accuracy of 100% for all sentence pairs (both satisfactory and unsatisfactory) occurs when both SLR and CR are combined together with a threshold of 2.5. Results showed that different threshold values tended to apply to satisfactory and unsatisfactory translation in the case of text in the small scale trail.

7.4 Contributions

The main contributions of this research are the development of a new hybrid metric for alignment and the creation of a high quality corpus. The best threshold values for both CR measurement and the SLR approach for the new corpus have been identified. These were established by using a subset of the corpus being analysed. This new approach also helps researchers or developers to improve new parallel corpora easily. The new hybrid sentence alignment has been demonstrated and applied during the new parallel corpora development to help with distinguishing good translations from bad. This establishes a solid foundation for future development.

Secondly, the previous literature acknowledged that many of the existing parallel Arabic/English corpora in the marketplace are high in cost and low in accuracy. This research has developed an inexpensive parallel corpus of Arabic and English which will assist in future development in MT research. The most important aspect of the new corpus is that the quality of co-translations has been improved significantly by using the PPMD-based compression algorithm. Figure 6.6 illustrates

this process using a flowchart to show how the parallel corpus A was analysed in this manner.

Finally in practical terms, this more accurate alignment method will encourage more researchers and students, and even perhaps some private developers, to contribute by expanding parallel corpora.

7.5 Recommendations for Future Research

During the process of this doctoral research project, a series of questions have arisen that merit further investigation. The issues are summarised as follows:

- Further research is needed to confirm that the information in one language is similar to that in the other.
- The sources of the new parallel corpora are still narrow, covering just 13 categories. Further sources should be obtained to expand the scale of the corpus. An appropriate target for an even larger corpus would be 100 million words, in order to obtain greater amounts of training data.
- The quality of the raw material still has room for improvement. Currently, raw materials are manually filtered, but unfortunately, it is difficult to eliminate all problems with raw data via individual work. Furthermore, manual work is seen as a low-effectiveness approach to control quality and may decrease reliability. Therefore, it is highly recommended that an assistant programme or auto-check mechanism should be added to control and improve the quality of raw materials in the first place.
- The research could be extended to align at the phrase and word levels in the future. The current parallel test corpora has only completed alignment at the sentence level.
- Once this is done, the resulting parallel corpus could be used to train an MT system.

References

Abb, B., Buschbeck-wolf, B., Tschernitschek, C., 1996. Abstraction and underspecification in semantic transfer. In: In Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96). Montreal, pp. 56–65.

Ahmed, F., Nürnberger, A., 2008. Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes. Vol. 8. In Proceedings of the 12th European Machine Translation Conference (EAMT), Hamburg, Germany, pp. 6–11.

Al-Batal, M., 1992. "diglossia proficiency: The need for an alternative approach to teaching". in A. Rouchdy (ed.) The Arabic language in America Detroit, MI: Wayne State Press, pp. 284–304.

Al-Sulaiti, L., Atwell, E., 2006. The design of a corpus of contemporary Arabic. Vol. 11. International Journal of Corpus Linguistics, John Benjamins Publishing Company, pp. 135–171.

Al-Toma, S. J., 1969. The problem of diglossia in Arabic: A comparative study of classical and Iraqi Arabic. Cambridge, MA:Harvard University Press.

Alansary, S., Nagi, M., Adly, N., 2008. Building an international corpus of Arabic (ICA): Progress of compilation stage. In Proceedings of In 7th International Conference on Arabic Language Resources and Tools in Cairo, Egypt.

Alhawiti, K., 2014. Adaptive models of Arabic text. Ph.D. thesis, Bangor University.

Aliprand, J. M., 1992. Arabic script on RLIN. Vol. 10. In Proceedings of Library hi tech, MCB UP Publishing Ltd, pp. 59–80.

Ambati, V., Vogel, S., 2010. Can crowds build parallel corpora for machine translation systems? In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, pp. 61–65.

Ayari, S., 1996. "diglossia and illiteracy in the Arab world". Vol. 9. Language, Culture and Curriculum, pp. 243–253.

Bar-Hillel, Y., 1964. Demonstration of the nonfeasibility of fully automatic high quality translation. In Proceedings of Language and information, New York: Academic Press, pp. 174–179.

Behr, F., Fossum, V., Mitzenmacher, M., Xiao, D., 2003. Estimating and comparing entropy across written natural languages using PPM compression. In Proceedings of Proceedings of the Data Compression Conference, John Benjamins Publishing Company, p. 416.

Beland, R., 2001. Deep dyslexia in the two languages of an Arabic/French bilingual patient. Vol. 82. Cognition, pp. 77–126.

Biber, D., Conrad, S., Reppen, R., 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge, Cambridge University Press.

Bolia, R. S., Nelson, W. T., Ericson, M. A., 2000. A speech corpus for multitalker communications research. Vol. 107. The Journal of the Acoustical Society of America, AIP Publishing LLC, pp. 1065–1066.

Bowker, L., Pearson, J., 2002. Working with Specialized Language: A Practical Guide to Using Corpora. London & New York, Routledge, p. 9.

Braude, J., 2011. The Honored Dead: A story of friendship, Murder and the Search for Truth in the Arab World. New York Spiegel & Grau.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., Roossin, P. S., 1990. A statistical approach to machine translation. Vol. 16. Journal of Computational linguistics, MIT Press, pp. 79–85.

Brown, P. F., Lai, J. C., Mercer, R. L., 1991. Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 169–176.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., Mercer, R. L., 1993. The mathematics of statistical machine translation: Parameter estimation. Vol. 19. Journal of Computational linguistics, MIT Press, pp. 263–311.

Burgess, C., Simpson, G. B., 1988. Cerebral Hemispheric Mechanisms in the Retrieval of Ambiguous Word Meanings. Vol. 33. Proceedings of Brain and Language, Published by Elsevier Inc., pp. 86–103.

Calderbank, R., Sloane, N. J., 2001. Obituary: Claude Shannon (1916–2001). Vol. 410. Journal of Nature, Nature Publishing Group, pp. 768–768.

Chiang, D., 2005. A hierarchical phrase-based model for statistical machine translation. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 263–270.

Cleary, J. G., Witten, I., 1984. Data compression using adaptive coding and partial string matching. Vol. 32. Journal of Communications, IEEE Transactions on, IEEE, pp. 396–402.

Davies, M., 2010. Creating Useful Historical Corpora: A Comparison of CORDE, the Corpus del Español, and the Corpus do Portugués. Proceedings of In Diacron ía de las lenguas iberorromances: nuevas perspectivas desde la lingüística de corpus, ed. Andrés Enrique-Arias. Frankfurt/Madrid.

Diab, M., 2000. An unsupervised method for multilingual word sense tagging using parallel corpora: A preliminary investigation. In Proceedings of the ACL-2000 workshop on Word senses and multi-linguality, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 1–9.

Dorna, M., Emele, M. C., 1996. Semantic-based transfer. In Proceedings of the 16th conference on Computational linguistics, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 316–321.

Durrett, G., Pauls, A., Klein, D., 2012. Syntactic Transfer Using a Bilingual Lexicon. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics Publishing, Jeju Island, Korea, pp. 1–11.

Echihabi, A., Marcu, D., 2003. A noisy-channel approach to question answering. Vol. 1. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 16–23.

Ecma, 2011. Standard ECMA-144: 8-Bit Single-Byte Coded Character Sets - Latin Alphabet No. 6. Accessed: 2013-01-10.
URL http://www.ecma-international.org/publications/standards/Ecma-144.htm

ECMAScript, E. C. M. A., Association, E. C. M., 2011. ECMAScript Language Specification. Ecma International Publishing.

Elbeheri, G., Everatt, J., Reid, G., Al-Mannai, H., 2006. "Dyslexia assessment in Arabic". Vol. 10. Journal of Research in Special Educational Needs, pp. 143–152.

Emery, P., 1991. Lexical incongruence in Arabic-English translation. Vol. 3. Translators' Journal, Babel, pp. 129–137.

Evans, D., 2007. Corpus building and investigation for the Humanities. University of Birmingham, accessed: 2015-07-17.
URL http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/Intro/Unit1.pdf

Farghaly, A., Shaalan, K., 2009. Arabic Natural Language Processing: Challenges and Solutions. Vol. 8. In Proceedings of ACM Transactions on Asian Language Information Processing (TALIP), ACM, New York, NY, USA, pp. 14:1–14:22.

Ferguson, C. A., 1959. Diglossia. Vol. 15. Word, pp. 325–340.

Franz, M., McCarley, J. S., Ward, T., Zhu, W.-J., 2001. Quantifying the utility of parallel corpora. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM New York, NY, pp. 398–399.

Frías Conde, X., 2000. Some Parallels between Arabic and Romance Languages. Vol. 1. IANUA, pp. 14–31, accessed: 2011-11-10.
URL http://www.romaniaminor.net/ianua/Ianua01/01Ianua02.pdf

Gale, W., Church, K., 1993. A program for aligning sentences in bilingual corpora. In Proceedings of ACL'93 29th Annual Meeting, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 177–184.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., Van Durme, B., 2011. Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 1168–1179.

Gansner, E. R., North, S. C., Sep. 2000. An Open Graph Visualization System and Its Applications to Software Engineering. Vol. 30. Softw. Pract. Exper., John Wiley & Sons, Inc., New York, NY, USA, pp. 1203–1233.

GGDC, 2014. Windows Code Pages - Windows 1256. Accessed: 2013-01-10.
URL http://msdn.microsoft.com/en-us/goglobal/cc305149.aspx

Ghazzawi, S., 1986. The Arabic Language. Center for Contemporary Arab Studies Georgetown University, Washington, DC.

Guidère, M., 2002. Toward corpus-based machine translation for standard Arabic. Translators and Computers, accessed: 2015-07-17.
URL http://translationjournal.net/journal/19mt.htm

Habash, N., 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics, pp. 57–60.

Habash, N., 2012. Machine Translation and Arabic Language Issues. AMTA Conference, San Diego, October 28, 2012, accessed: 2015-07-17.
URL http://www.amta2012.amtaweb.org

Habash, N., Sadat, F., 2006. Arabic preprocessing schemes for statistical machine translation. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 49–52.

Haeri, N., 2000. “form and ideology: Arabic sociolinguistics and beyond”. Vol. 29. Annual Review of Anthropology, pp. 61–87.

Hajlaoui, N., Kolovratník, D., Väyrynen, J., Steinberger, R., Varga, D., 2014. DCEP -Digital Corpus of the European Parliament. Proceedings of the Ninth

International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014., pp. 3164–3171.

Haruno, M., Yamazaki, T., 1996. High-performance bilingual text alignment using statistical and dictionary information. In Proceedings of Proceedings of the 34th Annual Meeting of Association for Computational Linguistics, Association for Computational Linguistics Stroudsburg, PA, USA, pp. 131–138.

Holes, C., 1995. Modern Arabic: Structures, functions, and varieties. London: Longman.

Homiedan, A., 1997. Machine translation. In Proceedings of King Saud University Periodical, King Saud University, p. 10.

Hunston, S., Francis, G., 2000. Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. Amsterdam, John Benjamins.

Hutchins, W. J., 1986. Machine translation: past, present, future. Ellis Horwood Chichester.

Hutchins, W. J., 2001. Machine translation over fifty years. Vol. 23. In Proceedings of Histoire epistémologie langage, Histoire Épistémologie Langage, pp. 7–31.

Hutchins, W. J., Somers, H. L., 1992. An introduction to machine translation. Vol. 362. London: Academic Press.

Imamura, K., Okuma, H., Watanabe, T., Sumita, E., 2004. Example-based Machine Translation Based on Syntactic Transfer with Statistical Models. Proceedings of Coling 2004, COLING, pp. 99–105.

IngilizceTurkce.Gen.Tr., 1998. The Advantages and Disadvantages of Machine Translation. Accessed: 2015-10-02.
URL http://www.omniglot.com/language/articles/machinetranslation.htm

IOFS, 1999. Iso/iec 8859-6:1999 information technology – 8-bit single-byte coded graphic character sets – part 6: Latin/Arabic alphabet. Accessed: 2013-01-10.
URL http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=28250

jaan Kaalep, H., Veskis, K., 2007. Comparing parallel corpora and evaluating their quality. In: In Proceedings of MT Summit XI. pp. 275–279.

Katz, S. M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. Vol. 35. In Proceedings of IEEE Transactions on Acoustics, Speech and Signal Processing, IEEE, pp. 400–401.

Kay, M., Röscheisen, M., 1993. Text-translation alignment. Vol. 19. In Proceedings of Computational Linguistics, MIT Press Cambridge, MA, pp. 121–142.

Kennedy, G., 2014. An introduction to corpus linguistics. Routledge Publishing.

Khadivi, S., Ney, H., 2005. Automatic filtering of bilingual corpora for statistical machine translation. Vol. 3513. In Proceedings of Natural Language Processing and Information Systems, Springer, pp. 263–274.

Kherallah, M., Bouri, F., Alimi, A. M., 2009. Toward an on-line Arabic handwriting recognition system based on visual encoding and genetic algorithm. Vol. 22. In Proceedings of Engineering Applications of Artificial Intelligence, Springer Vienna, pp. 153–170.

Khmelev, D. V., Teahan, W. J., 2003. A Repetition Based Measure for Verification of Text Collections and for Text Categorization. SIGIR '03. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, New York, NY, USA, pp. 104–110.

Kitchen, A., Ehret, C., Assefa, S., Mulligan, C. J., 2009. Bayesian phylogenetic analysis of semitic languages identifies an early bronze age origin of semitic in the near east. Vol. 276. In Proceedings of the Royal Society B: Biological Sciences, Royal Society Publishing, pp. 2703–2710.

Klavans, J., Gonzalo, J., 2006. Recent Advances in Natural Language Processing and Information Retrieval. The 38th Annual Meeting of the Association for Computational Linguistics.

Koehn, P., 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. Vol. 1. In Proceedings of The 10th Machine Translation Summit X, Phuker, Thailand, pp. 79–86.

Koehn, P., 2009. Statistical machine translation. Cambridge University Press.

Koehn, P., Knight, K., 2003. Empirical Methods for Compound Splitting. EACL '03. Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1, Association for Computational

Linguistics, Stroudsburg, PA, USA, pp. 187–193.
URL http://dx.doi.org/10.3115/1067807.1067833

Kuhn, M., 2013. UTF-8 and Unicode FAQ for Unix/Linux. Accessed: 2013-01-10.
URL http://www.cl.cam.ac.uk/~mgk25/unicode.html

Kutuzov, A., 2013. Improving English-Russian sentence alignment through pos tagging and damerau-levenshtein distance. In Proceedings of Association for Computational Linguistics, Association for Computational Linguistics, pp. 63–68.

Lantsov, V. A., Petrovskii, Y. A., 2010. Classification of encoding of text message characters. Vol. 69. In Proceedings of Telecommunications and Radio Engineering, Begell House, NY.

LDC, 2013. Linguistic Data Consortium. Accessed: 2013-10-23.
URL http://catalog.ldc.upenn.edu

Leacock, C., Miller, G. A., Chodorow, M., 1998. Using Corpus Statistics and WordNet Relations for Sense Identification. Vol. 24. Proceedings of Computational Linguistics, MIT Press, pp. 147–165.

Liu, W., Chang, Z., Teahan, W., 2014. Experiments with compression-based methods for English-Chinese sentence alignment. In Proceedings of Second International Conference on Statistical Language and Speech Processing (SLSP), Springer International Publishing, pp. 14–16.

Mahmoud, S., 1994. Arabic character recognition using fourier descriptors and character contour encoding. Vol. 27. In Proceedings of Pattern Recognition, Taylor & Francis, Inc. Bristol, PA, USA, pp. 815–824.

Manin, D., 1996. The right word in the left place: Measuring lexical foregrounding in poetry and prose. In Proceedings of language and literature, John Benjamins Publishing Company.

Mansour, M. A., 2013. The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus. Vol. 3. International Journal of Humanities and Social Science, pp. 81–90.

McEnery, A., Xiao, R., 2007. Parallel and comparable corpora: What are they up to? Translating Europe. Incorporating Corpora: Translation and the Linguist, Multilingual Matters Publishing, p. 3.

Melamed, I., 2000. Models of translational equivalence among words. Vol. 26. In Proceedings of Computational Linguistics, MIT Press Cambridge, MA, pp. 221–249.

Moore, R. C., 2002. Fast and accurate sentence alignment of bilingual corpora. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, Springer-Verlag, pp. 135–144.

Mújdricza-Maydt, E., Körkel-Qu, H., Riezler, S., Padó, S., 2013. High-precision sentence alignment by bootstrapping from wood standard annotations. Vol. 99. In Proceedings of The Prague Bulletin of Mathematical Linguistics, ISSN (Online), pp. 5–16.

Munteanu, D. S., Marcu, D., 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Vol. 31. In Proceedings of Computational Linguistics, MIT Press Cambridge, MA, USA, pp. 477–504.

Nida, E., 1975. Language structure and translation. Vol. 8. In Proceedings of Language in Society, Stanford University Press, pp. 300–303.

Palmer, M., Gildea, D., Kingsbury, P., 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Vol. 31. Proceedings of Computational Linguistics, MIT Press, pp. 71–106.

Papageorgiou, H., Cranias, L., Piperidis, S., 1994. Automatic alignment in corpora. In Proceedings of The 32nd Annual Meeting of Association of Computational Linguistic, Association of Computational Linguistic, pp. 334–336.

Popescu-Belis, A., Renals, S., Bourlard, H., 2008. Machine Learning for Multimodal Interaction: 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers. Springer-Verlag.

Rasheed, Z. T., 2008. Arabic is the tie that binds. Al-Jazeera, 28/01/2008, accessed: 2015-07-17.
URL http://www.aljazeera.com/focus/arabunity/2008/01/2008525185325418882.html

Resnik, P., Smith, N. A., 2003. The Web as a Parallel Corpus. Vol. 29. In Proceedings of Computational Linguistics CL, MIT Press Journals, pp. 349–380.

Roozgar, M. B., 2008. 'hybrid texts, sources and translation'. Accessed: 2015-09-08.
URL http://www.translationdirectory.com/articles/article1655.php

Saiegh-Haddad, E., 2004. The Relevance of Linguistic and Sociolinguistic Features of Arabic Diglosssia to the Acquisition of Various Reading Skills in Arabic. British Dyslexia Association International Conference 2004, accessed: 2015-01-10.
URL http://www.bdainternationalconference.org/2004/presentations/sat_s6_c_1.shtml

Samy, D., S, A. M., Guirao, J. M., Alfonseca, E., 2006. Building a parallel multilingual corpus (arabic-spanish-english). In Proceedings of Language Resources and Evaluation Conference.

Santos, A., 2011. Contributions for Building a Corpora-Flow System. Msc dissertation, University of Minho.

Shannon, C., 1948. A mathematical theory of communication. Vol. 7. Bell System Technical Journal, ACM New York, NY, pp. 379–423, 623–656.

Shannon, C., Weaver, W., 1949. The mathematical theory of communication. University of Illinois Press.

Simard, M., Foster, G., Isabelle, P., 1992. Using cognates to align sentences in bilingual corpora. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in (MT), IBM Publishing, pp. 67–81.

Sinclair, J., 1991. Corpus, Concordance, Collocation. Oxford, Oxford University Press.

Skadiņš, R., Tiedemann, J., Rozis, R., Deksne, D., may 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In: Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland.

Soudi, A., van den Bosch, A., Neumann, G., 2007. Arabic Computational Morphology: Knowledge-based and Empirical Methods, 1st Edition. Springer Publishing Company, Incorporated.

Stubbs, M., 1996. Text and corpus analysis: Computer-assisted studies of language and culture. Oxford, OX, UK and Cambridge, Mass., USA, Blackwell Publishers.

Taghipour, K., Khadivi, S., Xu, J., 2011. Parallel corpus refinement as an outlier detection algorithm. In Proceedings of MT Summit XIII, NA.

Teahan, W., 1998. Modelling English text. Ph.D. thesis, University of Waikato.

Tiedemann, J., 2003. Combining clues for word alignment. In: In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 339–346.

Tiedemann, J., 2004. Word to word alignment strategies. COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland.

Tiedemann, J., 2009. News from OPUS : A collection of multilingual parallel corpora with tools and interfaces. In: Recent Advances in Natural Language Processing V. Vol. V. John Benjamins, pp. 237–248.

Tognini-Bonelli, E., 2001. Corpus Linguistics at Work. Proceedings of Computer learner corpora, second language acquisition and foreign language teaching, John Benjamins Publishing Company, p. 55.

UN, 2013. UN official languages. Accessed: 2013-10-21.
URL https://www.un.org/en/aboutun/languages.shtml

UNDP, 2003. Arab human development report 2003: Building a knowledge society. New York: UNDP, Regional Bureau for Arab States, accessed: 2015-07-17.
URL http://www.miftah.org/Doc/Reports/Englishcomplete2003.pdf

Vauquois, B., 1968. A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In: IFIP Congress (2). DBLP computer science bibliography, pp. 1114–1122.

Versteegh, K., 2001. The Arabic Language. Edinburgh University Press.

Weaver, W., 1949. Translation in machine translation of languages. Technology Press of MIT and Wiley and Sons, New York.

Wu, D., June 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Las Cruces, New Mexico, USA, pp. 80–87.

Wu, P., 2007. Adaptive models of Chinese text. Ph.D. thesis, University of Wales, Bangor.

Wynne, M., 2005. Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books.

Xiao, R. Z., 2008. Well-known and influential corpora. In: Corpus Linguistics: An International Handbook. Handbooks of Linguistics and Communication Science, Mouton de Gruyter, Berlin.

Xiao, T., Zhu, J., Yao, S., Zhang, H., 2011. Document-level Consistency Verification in Machine Translation. Proceedings of the 13th Machine Translation Summit (MT Summit XIII), International Association for Machine Translation, pp. 131–138.

Yamabana, K., Apr. 13 2006. Translation system, translation communication system, machine translation method, and medium embodying program. US Patent App. 11/235,333. Accessed: 2013-01-10.
URL http://www.google.com.lb/patents/US20060080079

Yu, Q., Max, A., Yvon, F., 2012. Revisiting sentence alignment algorithms for alignment visualization and evaluation. In Proceedings of LREC Workshop on Building and Using Comparable Corpora, ELRA – European Language Resources Association, pp. 10–16.

Zughoul, M., Abu-Alshaar, A., 2005. English/Arabic machine translation: A historical perspective. Vol. 3. Translators' Journal, pp. 1022–1041.