The Utilization of Parallel Corpora for the Extension of Machine Translation Lexicons

Jeanne Pienaar and G.D. Oosthuizen,
*Department of Computer Science, University of Pretoria,
Pretoria, South Africa*

Abstract: There has recently been an increasing awareness of the importance of large collections of texts (corpora) used as resources in machine translation research. The process of creating or extending machine translation lexicons is time-consuming, difficult and costly in terms of human involvement. The contribution that corpora can make towards the reduction in cost, time and complexity has been explored by several research groups. This article describes a system that has been developed to identify word-pairs, utilizing an aligned bilingual (English-Afrikaans) corpus in order to extend a bilingual lexicon with the words and their translations that are not present in the lexicon. New translations for existing entries can be added and the system also applies grammar rules for the identification of the grammatical category of each word-pair. This system limits the involvement of the human translator and has a positive impact on the time, cost and effort needed to extend a bilingual lexicon.

Keywords: ALIGNMENT, BILINGUAL CORPORA, CORPUS, EXTENSION, LEXICON, MACHINE TRANSLATION, MONOLINGUAL CORPORA, PARALLEL CORPORA

Opsomming: Die benutting van parallelle korpusse vir die uitbreiding van masjienvertalingsleksikons. Onlangs was daar 'n toenemende bewustheid van die belangrikheid van groot versamelings tekste (korpusse) wat as bronne in die navorsing van masjienvertaling gebruik word. Die proses om masjienvertalingsleksikons te skep of uit te brei is tydrowend, kompleks en duur in terme van menslike betrokkenheid. Die bydrae wat korpusse kan maak tot die vermindering van koste, tyd en kompleksiteit is deur verskeie navorsingsgroepe ondersoek. Hierdie artikel beskryf die ontwikkeling van 'n stelsel wat gebruik maak van 'n afgepaarde tweetalige (Engels-Afrikaanse) korpus vir die identifisering van woordpare met die doel om 'n bestaande tweetalige leksikon uit te brei met hierdie woorde en hul vertalings wat nie in die leksikon voorkom nie of om nuwe vertalings vir bestaande inskrywings by te voeg. Die stelsel pas ook grammatikareëls toe vir die identifisering van die grammatikale kategorie van elke woordpaar. Die stelsel beperk die betrokkenheid van die menslike vertaler en het 'n positiewe impak op die vermindering van tyd, koste en moeite in die uitbreiding van 'n tweetalige leksikon.

Sleutelwoorde: AFPARING, EENTALIGE KORPUSSE, LEKSIKON, MASJIENVERTALING, PARALLELLE KORPUSSE, TWEETALIGE KORPUSSE, KORPUS, UITBREIDING

1. Introduction

The term *corpus* originates from the Latin word "body" and can be defined as a sufficiently large collection of text samples of the written or spoken language. The utilization of a corpus as a source of information on the nature and understanding of language was delayed as a result of criticism by Noam Chomsky in the late 1950s. He viewed corpora as being inadequate in terms of size and representativeness. His view became the conformed opinion of the next generation of theoretical linguists. With the advance of computer technology and the availability of large collections of texts (corpora) in machine-readable form, the potential of using machine-readable corpora has been recognized in a variety of areas such as lexical knowledge acquisition, grammar-construction and machine translation. A large number of computerized corpora are available which vary in design, size and research purpose.

The purpose of this article is, firstly, to give an overview of the state of the art regarding the use of corpora in machine translation. Secondly, it describes the development of a system that utilizes an aligned bilingual (English-Afrikaans) corpus for the identification of word-pairs in order to extend a bilingual lexicon with the words and their translations that are not present in the lexicon, or for adding new translations for existing entries.

In the next section, various approaches which utilize the information contained in corpora, will be outlined. The purpose of section 3 is to identify some problems that are experienced in the use of corpora such as corpus size, corpus representativeness as well as vocabulary- and word frequency-related problems. The aim of section 4 is to give an overview of the alignment of bilingual corpora. In section 5, the alignment method utilized for aligning a sample of the South African Hansards is explained. The problems experienced during the preparation of the corpora for the project, as well as problems encountered with the alignment process, are described. The aligned bilingual corpus was used as a resource for the task of extending a bilingual lexicon. In section 6, the focus is on the method utilized to extend a bilingual lexicon with the aligned English-Afrikaans corpus as a resource. The development of the system is described and the results obtained from various experiments are discussed. The aim of the final section, section 7, is to give an idea of some future research topics utilizing bilingual corpora as resources, and to end with a few concluding remarks.

2. Applications of Computer Corpora

There are several approaches to using computer corpora as components in machine translation systems. Corpora can broadly be divided into two types: monolingual and bilingual. In this section, an overview of these two types of corpora will be given, and the various applications of both monolingual and bilingual corpora will be provided.

2.1 Monolingual Corpora

At the most basic level a monolingual corpus is an important resource for a linguist in the determination of language-usage in a given domain (Arnold *et al.* 1994). Monolingual corpora have proved to be excellent sources of lexical information as well as limited world knowledge and are used extensively in the production of monolingual machine translation lexicons. Large text corpora allow for a detailed study of how a word is used, thus enabling the evaluation of the accuracy of lexicon entries by comparing it with evidence of how the word is utilized in the real world. Very large corpora provide reliable statistics on occurrence and co-occurrence of lexical categories and word-senses. Terms or idioms in context can also be identified.

Some machine translation systems require the automatic grammatical analysis of texts (tagging[1] and parsing[2]) as a first stage of analysis and the monolingual corpus usually serves as input for the tagging process. The utilization of computer programs for automatic tagging and parsing have a positive impact on speed and consistency. An important use of annotated corpora is the provision of probability statistics for probabilistic language processing systems.

2.2 Bilingual Corpora

Recently, the interest of some research groups in machine translation and linguistics has shifted to bilingual corpora. There are currently very few large machine-readable bilingual corpora. The South African Hansards, proceedings of the South African parliamentary debates, which by law were published in English and Afrikaans are examples of bilingual corpora.

Bilingual corpora are especially useful if the user can view translation segments. In bitext (or multitext) the text is aligned in such a way that within each bilingual (or multilingual) segment the two texts are translations of each other. A properly aligned bilingual corpus is an exceedingly important and valuable source of information and can be used in several ways to contribute to direct

[1] The tagging process associates with each word in a text or corpus of texts a tag identifying its lexical category and occasionally its syntactic features or subcategories (Garside *et al.* 1987).

[2] Parsing (syntactic analysis) is the process of recognizing a sentence and simultaneously compiling a representation of its structure. Thus, a sentence will be partitioned into its constituent phrases, subphrases and lexical categories.

automatic machine translation with a degree of human assistance. Several alignment algorithms have been developed for aligning bilingual texts and attention will be paid in section 4 to different approaches to aligning sentences in bilingual corpora.

Bilingual corpora are important resources for the construction as well as the extension of bilingual lexicons, as they are rich repositories of information about actual language-usage. A lexicon can be created or updated by deriving lexical information from a corpus and lexical frequency lists can be constructed from raw (untagged) corpora. Bilingual corpora give detailed information on the properties of words and on the selection restrictions of verbs, which can provide a basis for the identification of the senses of the occurrences of a given word in a corpus. Examples of possible uses of a word are provided by a corpus and although a word's semantic type has been identified, the facts provided by a bilingual corpus can be utilized to at least confirm a classification according to a lexical definition and possibly also add supplementary attributes that are absent from a lexicon.

An annotated (tagged) aligned bilingual corpus together with a probabilistic model could be used to automatically provide equivalent terms in the two languages. These terms can then be automatically compiled into the relevant formalism for lexical entries in a machine translation system. Thus, an important and possibly complex channel of information transfer exists between corpora and lexicons.

Today, computer technology provides fast access to large memories and data sources. Therefore, methods based on the access of large machine-readable corpora can be investigated. Some of the corpus-based machine translation systems utilize the corpus as a database and examples of this type of corpus-based machine translation are sublanguage machine translation systems, example-based methods/systems and knowledge-based methods/systems.

Other corpus-based machine translation approaches use statistical and probabilistic techniques for the analysis of the source language text and the generation of the target language text. An example of this type of corpus-based machine translation is the statistics-based approach of Brown *et al.* (1990). Example-based translation and statistically-based translation are so-called "empirical" approaches which apply relatively low-level statistical or pattern-matching techniques. The term "empirical" is used to refer to the fact that whatever linguistic knowledge is used by the system, is derived empirically by examining the real texts and not by relying on the knowledge of linguists.

An example of the first type of corpus-based machine translation is a sublanguage machine translation system where the texts are written in a particular sublanguage for a specific subject domain. The sublanguage corpus is utilized as a database and is searched for a source language string which is similar or identical to the string to be translated.

Example-based and knowledge-based methods are not purely corpus-based approaches to machine translation as they are not limited to a particular sublanguage or to specific corpora. However, large corpora of texts in source languages and target languages must be accessed to obtain the necessary knowledge. The acquired knowledge will be used in the translation of previously unseen texts. The problem of acquiring and managing required syntactic and semantic knowledge forces these approaches to be applied within specific domains. The incorporation of the large volume of knowledge tends to make the system large and complicated and, therefore, not cost-effective.

In 1949 Warren Weaver suggested the application of statistical techniques for the translation of text from one natural language into another (Weaver 1955). Unfortunately, efforts in this direction were soon abandoned for various theoretical and philosophical reasons. Over the last few years there has been a trend back towards the application of statistical techniques and methods in the analysis and generation of text. The growing availability of bilingual, machine-readable texts has also stimulated the interest in the utilization of these methods (Armstrong 1994).

Statistical methods have been successfully applied to lexicography and to natural language processing. However, the success of statistical-based approaches to speech research in recent years is the primary reason for the upsurge in applying statistical methods to machine translation. Statistically-based translation is entirely statistical and probabilistic. No grammatical information is incorporated as explicit rules for the analysis of source language texts and the generation of target language texts.

The essence of this approach is the alignment of the sentences in the two languages and the calculation of the probability that one word in a sentence of the source language text corresponds to two, one or zero words in the sentence of the target language text (Hutchins 1992). The knowledge-acquisition problem is eliminated, since linguistic information is not explicitly encoded, but the general suitability of this method may be in doubt as it requires very large amounts of good quality bilingual or multilingual data.

Bilingual corpora can play an important role in the evaluation of machine translation systems. The corpus can be used to test the experimental system on real data. The tests on the corpus will help to uncover errors, limitations as well as potential areas which need improvement. The corpus can be utilized to evaluate the linguistic quality of the translation by comparing it to the target language part of the corpus. With the help of the corpus it will be possible to make sure that phenomena which appear in theoretical linguistics do indeed occur in the real world.

However, certain words and syntactic structures will not be found in the corpus. The vocabulary of the corpus is limited and the translation provided for the source text of the corpus may not be the one and only correct option. Therefore, the corpus must only be used as a guide for the evaluation of machine translation systems. The role of the corpus can thus be seen as aid in the definition of research goals and for the provision of material for system testing.

In this section, the applications of monolingual and bilingual computer corpora in machine translation were outlined. The utilization of corpora in machine translation is not problem-free and unrestricted, and in the next section some of the problems will be identified.

3. Problems with Corpora

In contrast to the era of Chomsky, the late 1950s, the computers of today are more powerful, faster and have hundreds and millions of bytes of storage. It is therefore possible to utilize very large corpora for research purposes. The question of the importance of corpus size inevitably arises. As the corpus size increases, the number of new types of words decreases and less effort is needed to search larger and larger corpora. However, a large portion of all word-types encountered occur only once. The problem is how large the corpus must be to capture all the words of a language as a corpus of more or less 100 million words will only produce less than half the theoretical total of word-types (Sebba 1991).

It is important for the corpus to be representative of the totality of texts from which it is drawn. At present, no statistical or other models exist for the determination of the representativeness of a corpus. One of the problems in selecting a representative sample is that in order to make valid conclusions, the sampling must be random. It is required that the procedure of composing a random sample should be objective and it would seem impossible to obtain a representative sample from a corpus. However, presently no actual criteria exist for the selection of a "representative" corpus and thus far corpus linguists assembling large machine-readable corpora have made intuitively-guided decisions about what to include and in what proportions (Sebba 1991).

The decision about what types of text to include or exclude is also difficult. The Brown corpus, for example, contains samples of text which were drawn from sources such as newspaper reports, government documents and popular fiction, but excludes poetry. It is, therefore, not possible to generalize results without difficulty.

Bilingual corpora only exist in restricted fields. Although corpora may be viewed as translations of each other, the human translators do not usually translate sentence by sentence. The translations are strictly suited to the context in which the individual source language sentence occurs. Therefore, there are several problems regarding the utilization of automated methods to translate new sentences on the basis of existing translations, as one source language sentence can translate to zero, one or two target language sentences. Thus, it seems that the short-term value of these systems, if they go beyond the experimental stage, would be as (possibly interactive) aids to human translators.

4. The Alignment of Bilingual Corpora

In section 2 it was mentioned that a bilingual corpus is best utilized if the texts are aligned in such a way that they are translations of each other. The production of alignments by hand is extremely time-consuming and requires the skill of individuals with knowledge of both languages. Recently researchers in bilingual lexicography as well as machine translation have shown interest in the study of parallel texts and done some work on the alignment of sentences (Simard *et al.* 1992, Brown *et al.* 1991, and Gale and Church 1991). Alignment does not have to stop at sentence level and research has also been conducted to find alignments between syntactic structures, noun phrases, collocations and words.

The most common form of alignment takes the sentence to be the organizing unit and techniques exist for performing this alignment of bitext automatically with a high level of accuracy (Arnold *et al.* 1994). The alignment of paragraphs and sentences is only the first step towards the identification of word-correspondences, the construction of a probabilistic dictionary for the utilization in the alignment of words in machine translation and for the construction of a bilingual concordance for use in lexicography. Alignment can be defined as follows (Simard *et al.* 1992):

> Given a text and its translation, an alignment is a segmentation of the two texts such that the *n*th segment of one text is the translation of the *n*th of the other (empty segments are allowed as the result of additions or omissions).

The extraction of pairs of sentences from corpora that are translations of one another remains a problem as the alignment algorithm must cater for several scenarios, namely (Simard *et al.* 1992):

1. A single sentence in one language translates to one sentence in the other language.
2. A single sentence in one language may give rise to two or more translated sentences in the other language.
3. Two sentences can translate into one.
4. Two sentences in one language translate to two sentences in the other language.
5. A sentence may not be translated at all.
6. A new sentence may have no equivalent in the source text.

The possibility also exists that sentences, paragraphs or even passages can be missing from the corpora. These obstacles prevent many potential users from taking advantage of the many benefits of bilingual corpora as the solutions to these problems are computationally prohibitive, and/or unreliable (Simard *et* *al.* 1992). An aligned bilingual corpus provides several advantages. Some of these advantages are (Simard *et al.* 1992):

1. A valuable source of information is given.
2. A text and its translation can be viewed side by side, with explicit connections between individual components.
3. An alignment may form the basis of deeper automatic analysis of translation. For example, it could be utilized to indicate possible omissions in a translation. It could also be utilized for the detection of errors, for example to identify common translation mistakes.

Until recently, the South African parliamentary debates were by law published in both Afrikaans and English and is known as the South African Hansards. A subset of the Hansards was used to construct an Afrikaans-English corpus for the purpose of the project that will be described in detail in section 6. A method was developed for the alignment of the bilingual corpus at sentence level and the objective of the next section is to provide a description of the steps followed to align the English-Afrikaans corpus.

5. Alignment of the English-Afrikaans Hansard Corpus

The Hansards of the 1990 parliamentary sessions were used for the project and it was necessary that the corpus was in an aligned format. For each parliamentary session, there were two files: one English and one Afrikaans. The text was in XYWrite format and had to be converted to ASCII format. The next step was to combine all the English files into one large English corpus and all the Afrikaans files into one large Afrikaans corpus.

The corpora were not exact translations, since sentences, paragraphs or even passages were missing from the corpora. The duplication of sentences and paragraphs was also a phenomenon that had to be taken into account. It was, therefore, necessary to compare the texts and to manually remove the unmatched sentences, paragraphs and passages.

In total, between 10% and 15% of the data in each corpus were rejected. After the completion of this exercise the number of files in each corpus were the same, no unmatched passages existed, but about 10% unmatched paragraphs and about 20% unmatched sentences were still present. The process of removing the unmatched paragraphs and sentences is described in the next section.

The sizes of the two corpora made it impractical to obtain a complete set of alignments by hand. Therefore, some method had to be used for the alignment process to be done automatically. The method employed is outlined in the next section.

5.1 The Process of Alignment

The alignment of the South African Hansard corpora was done by implementing a simple algorithm. The alignment was done first at paragraph level and then the sentences within the paragraphs were aligned. The correctness of the method was checked by hand. A program was written to identify the start and the end of the proceedings.

The next step of the alignment process was to mark the end of each paragraph. It was found that the paragraphs in the South African Hansards were already within certain regions. Therefore, the process of marking the paragraphs was not too complicated. The paragraphs were aligned automatically by applying the alignment algorithm to the corpora and each paragraph was assigned a number. About 90% of the paragraphs were aligned correctly. The unmatched paragraphs occurred as a result of missing paragraphs or where the paragraphs did not correspond one to one. The texts were inspected for possible errors, and mismatches still present were removed. The alignment algorithm was applied again and at this stage the paragraphs corresponded 100%.

The paragraphs were aligned automatically and each paragraph was assigned a number. The texts were again inspected for possible errors, and mismatches still present were removed. The next step in the process was the implementation of the sentence-alignment program. Determining sentence boundaries was a problem. It was found that most of the sentences ended with ".", "]", "!", "?", or ">", and by using these symbols as beacons, each sentence was moved to a new line. The program then automatically numbered each sentence and a success rate of about 80% was achieved in aligning the sentences. The texts were inspected for possible errors and mismatched sentences still present were removed. The process of sentence alignment and checking was repeated and this time the success rate was higher. This process was repeated until all the sentences corresponded. Although the alignment of the corpus was a rather difficult and at times a tedious process and the correction of errors time-consuming, the lack of quality data and a roughly aligned corpus would have had a negative impact on the usefulness and accuracy of the system for this project. Tables 1 and 2 illustrate the format utilized for the numbering of the files, paragraphs and sentences in the Afrikaans and English corpora.

This alignment process forms the basis of the research project and the aligned corpora will serve as input to the program for the extension of a bilingual lexicon.

```
1 «SOF AFR1.TXT»
1.1
1.1.1
WOENSDAG, 2 MEI 1990«EOP»
1.2
1.2.1
VERRIGTINGS VAN DIE UITGEBREIDE OPENBARE KOMITEE-VOLKSRAAD«EOP»
1.3
```

```
1.3.1
Lede van die Uitgebreide Openbare Komitee kom om 15:30 in die Raadsaal van die Volksraad byeen.«EOP» ...
1.9
1.9.1
Die MINISTER VAN LANDBOU: Mnr die Voorsitter, terwyl ons vanmiddag die gebed gedoen het, het dit my deur die gemoed gegaan hoe gepas dit is dat ons die teenwoordigheid van die Almagtige Vader ook afbid op die dinge wat vandag begin.
1.9.2
Die gesprek oor gesprekke is 'n begin en ons kan bid dat dit goed sal verloop. [Tussenwerpsels.]
1.9.3
Ek wil netnou daarby uitkom en bietjie oor onderhandelinge gesels, maar ek wil dit sterk afets dat dit nie dieselfde is as wat vandag aan die gang is nie.«EOP» ...
1.302
1.302.1
«EOF AFR1.TXT»
```

Table 1: Alignment of paragraphs and sentences — Afrikaans section

```
1 «SOF ENG1.TXT»
1.1
1.1.1
WEDNESDAY, 2 MAY 1990«EOP»
1.2
1.2.1
PROCEEDINGS OF EXTENDED PUBLIC COMMITTEE - ASSEMBLY«EOP»
1.3
1.3.1
Members of the Extended Public Committee met in the Chamber of the House of Assembly at 15:30.«EOP» ...
1.9
1.9.1
The MINISTER OF AGRICULTURE: Mr Chairman, while prayers were being read this afternoon, I was struck by how fitting it is that we call upon the Almighty to bless what we are starting to do today.
1.9.2
The talks about talks are a start, and we can only pray that they will go well. [Interjections.]
1.9.3
In a moment I want to come back to that and say a few words about negotiations, but I clearly want to outline the fact that this is not the same as what is going on today.«EOP» ...
1.302
1.302.1
«EOF ENG1.TXT»
```

Table 2: Alignment of paragraphs and sentences — English section

6. The Implementation of a System for the Extension of a Bilingual Lexicon

As mentioned before, the purpose of this research project was to examine the viability of utilizing aligned parallel corpora to extend existing bilingual lexicons. To reach this goal the system that is developed relies on the existence of a sentence-aligned corpus as well as a bilingual lexicon. The English-Afrikaans corpus was constructed from a subset of the South African Hansards. The Machine Translation Research Group at the University of Pretoria constructed an English-Afrikaans lexicon as well as an Afrikaans-English lexicon. In the English-Afrikaans lexicon the source words are in English and the target words in Afrikaans and vice versa in the Afrikaans-English lexicon. The availability of both lexicons made it possible to use either the English text or the Afrikaans text as the source language text depending on which lexicon the user wants to extend.

Following the advice of a computer linguist it was decided to start with a simple domain and gradually move towards a more complex domain. The first experiments were conducted by only using one source language sentence and one target language sentence. These sentences are sentence-pairs that have been extracted from the aligned corpus. The other experiments used samples consisting of several sentences extracted from the aligned corpora. The program had to take the markers added by the alignment process into account when reading the source language file and the target language file. The markers indicated the start and end of each file, the end of paragraphs as well as the file numbers, paragraph numbers and sentence numbers.

The aim of the next section is to give some background information of the system. An overview of the methodology that was used to implement and test the system is also provided. Various experiments were conducted and the results obtained from these experiments will be discussed.

6.1 Background

This system provides a mechanism for the automatic identification of word-pairs that are present in a sample consisting of bilingual texts, but are absent from the bilingual lexicon (Pienaar 1996). It is also possible that a word in the source language text does have an entry in the lexicon, but that the target word entry could not be found in the target language text and that another translation has been used. Depending on which lexicon is chosen to be updated, the Afrikaans-English lexicon or the English-Afrikaans lexicon, the English corpus will be used as the source language text and the Afrikaans corpus will be used as the target language text and vice versa. One source and one target language sentence are read at a time. The source sentence is parsed and for each source language word the lexicon is checked if an entry exists. If the source language word is present in the lexicon, the target language sentence is parsed to determine if the target language word as specified in the lexicon appears in the target language sentence. An indicator is used to show if the target language word has been found or not.

By taking the positions of the unmatched source and the target language words in the sentences as well as the sentence-lengths into account, the likelihood that certain words are translations of each other, is calculated. The formulas used to determine the likelihood factor are:

$$\mathit{difference} = \mathit{abs}(\mathit{pos}_{\mathit{source}} - \mathit{pos}_{\mathit{target}})$$

$$\mathit{sentence\text{-}length} = (\mathit{sentence\text{-}length}_{\mathit{source}} + \mathit{sentence\text{-}length}_{\mathit{target}}) / 2$$

$$\mathit{likelihood\ factor} = (\mathit{sentence\text{-}length} - \mathit{difference}) / \mathit{sentence\text{-}length}$$

The difference value is determined by calculating the absolute difference in position of each unmatched source and target word in the source and target language sentence. Unfortunately it is not possible to utilize the positions of the words and the sentence-lengths only in calculating the likelihood factor. Some reasons are:

- The lengths of the target language sentence and the source language sentence often differ.
- Two or more words in the source language sentence translate to zero or one word in the target language sentence.
- One word in the source language sentence translates to zero, one, two or more words in the target language sentence.
- The positions of the source words in the source language sentence do not correspond with the positions of the target words in the target language sentence.

To determine more accurately if (a) source and target word(s) are translations of each other, the incorporation of linguistic knowledge was inevitable. For each entry in the lexicon, the grammatical category (type) of the word is specified. By using this syntactic information, it is possible to determine the grammatical types of the neighbouring word(s) of the unmatched word. A few basic grammar rules were implemented to determine the possible type of each word-pair. At present, the rules can give an indication whether the grammatical category of the word-pair is a noun, verb, adverb, adjective or a preposition.

The grammatical category of the preceding word as well as of the following word is queried, and the simple grammar rules are applied by taking these known types into consideration. In the case of insufficient information, for example, if the type of the preceding or following word is also unknown, or if the grammatical information does not conform to the rules, the grammatical category cannot be ascertained and is indicated as *unknown*. If more than one rule is satisfied, the possible types will be provided and it will be the responsibility of the user to select the correct type of the word-pair. The possibility also exists that two words following each other can translate to one, two or more words or that one word can translate to two or more words. The system caters for these scenarios in so far that it can identify a translation pair consisting of a compound noun written as two words (such as *water tariff*) and a compound noun written as one word (such as *watertarief*) as well as a verb-preposition pair (such as *gaan oor*) and a verb (such as *crosses*). Although the simplified grammar rules currently implemented give an idea of what the possible types of the unmatched words are, more intensive investigation is required for the refinement, improvement and extension of these rules.

The output of the system is a list of unmatched source words and possible translations. A likelihood factor indicating the numeric possibility that the word-pairs are translations of each other, as well as the possible grammatical type(s) are provided. The human translator or the user has to inspect the results and identify the correct translation pairs. The grammatical type of each word-pair must be verified. In the case of the type not being known the translator has to rely on his/her linguistic knowledge to identify the correct grammatical type of the word-pair. The word-pair and the grammatical type can then be added to the lexicon as a new entry. The objective of the next section is to give a brief overview of the methodology that the system follows to extend a bilingual lexicon.

6.2 Methodology

The process of identifying word-pairs, ascertaining the grammatical category and calculating the likelihood factor can be divided into the following stages:

- **Input**

The source and target words as well as the grammatical types are extracted from the lexicon and recorded. The source and target language sentences are read from the corpus.

- **Lexicon lookup**

The words from the source sentence are checked against the extracted lexicon entries. The target language sentence is searched for the translation specified in the lexicon. If the translations are located, an indicator for each target word is set to show that the searching process was successful. The source words not found are recorded.

- **Calculation of likelihood factor**

For each unmatched word-pair the likelihood factor is determined by taking the positions of the source and target words as well as the sentence lengths into account.

- **Assignment of grammatical categories**

For each unmatched word-pair, grammatical rules are applied to attempt to recognize the grammatical type. For example, the grammar rule constructed for the identification of the adjective <*tall*> in the sentence *The very tall trees* takes into account that the adjective is preceded by the adverb <*very*> and followed by the noun <*trees*>. Thus the rule which will satisfy this specific example has the following format:

> < *(det) the (adverb) very (?) tall (noun) trees* > →
> < *(det) the (adverb) very (adj) tall (noun) trees* >

If the recognition process is successful, the grammatical type is assigned to the word-pair. In the case of an unsuccessful recognition process the assigned type is *unknown*.

- **Identification of compounds**

Compound nouns (such as *election manifesto* and *verkiesingsmanifes*) and verb-preposition pairs (such as *oorneem* and *take over*) are identified.

- **Output**

The result of the program is a list containing an entry for each identified word-pair, the possible grammatical type(s) and the likelihood factor.

- **Inspection**

The list must be inspected by the human translator or the user for correctness. The correct word-pairs and their grammatical type must be extracted and added to the lexicon. The purpose of the next section is to describe some of the experiments that were conducted.

6.3 Experiment

Purpose

The aim of the experiment was to evaluate the performance of the system by utilizing several sentences differing in terms of:

- complexity (for example, the tense of the sentence, the number of noun and verb phrases, positions of the verbs and presence of compound prepositions),
- sentence length and
- number of unmatched words.

Method

Twenty sentences were randomly chosen from the Afrikaans-English corpus and used as input to the system. The sentences satisfied the requirements as stated in the purpose of this experiment, since they differed in complexity, sentence length and the number of unmatched words. The sample extracted from the Afrikaans corpus containing the source language sentences is shown in table 3.

1. Ek sê dit, want ek was betrokke by 'n verkiesing.
2. Die skrapping van hierdie betrokke wet het my volle ondersteuning.
3. Geen deel van die samelewing bly onaangeraak deur hierdie probleme nie.
4. As ons praat van spesifieke behoeftes, beteken dit nie 'n swak standaard van dienslewering nie.
5. Ek dink ons moet 'n duidelike onderskeid tref tussen die vlakke van dienslewering.
6. Ek bedank agb lede vir hul steun aan hierdie wetgewing.
7. Die begrotingswetsontwerp is aanvaar.
8. Een van die probleme ten opsigte van die solvensie van die pesioenfonds was die geweldige las wat die stelsel op die fonds geplaas het.
9. Ek wil net vir die agb Minister vra of sy departement betrokke is by die opstel van sulke programme en of hy enige onderwysdepartemente genader het.
10. Daar is 'n behoefte aan formele en informele voorligtingsprogramme.
11. Die natuurlewe dra by tot hierdie gehalte.
12. Na 'n baie deeglike ondersoek is 'n omvattende gewysigde skema daargestel.
13. Die beweging van mense in die wêreld stimuleer groei.
14. 'n Private monopolie tree dikwels op teen die belang van die verbruiker.
15. Kruissubsidiëring word dus 'n werklikheid.
16. Ek het begrip vir die feit dat ons die bronne moet beskerm en dat weersomstandighede, seisoene en al daardie dinge 'n invloed het op ons bronne aan die kus.
17. Ek dink dat ons 'n tydperk binnegaan waar verwagtinge geskep gaan word.
18. Die departement behou die bevoegdheid oor die toewysing en toedeling van water uit die hoofbron, en hierdie water sal ook aan die raad teen 'n tarief beskikbaar gestel word.
19. Ek wil graag begin deur 'n woord van hulde te spreek soos dit al die gebruik is wanneer 'n senior amptenaar aftree na lang jare van diens.
20. Ons weet dat beperkte toegang tot die rekeninge en beleggings van vermeende handelaars dikwels ondersoeke kortwiek.

Table 3: Sample of the Afrikaans corpus

The sample extracted from the English corpus containing the target language sentences is shown in table 4.

1. I say this, because I was involved in an election.
2. The abolition of this specific act has my full support.
3. No segment of the society stays untouched by this problem.
4. When we talk about specific requirements, this does not entail a poor standard of service.
5. I think we should draw a clear distinction between the levels of service.
6. I thank the hon members for their support for this legislation.
7. The appropriation bill was adopted.
8. One of the problems in respect of the solvency of the pension fund was the tremendous strain which the system had placed on the fund.
9. I simply want to ask the hon Minister whether his department is involved in the installation of such programmes and whether he has approached any education departments.
10. There is a need for formal and informal education programmes.
11. The wildlife adds to this quality.
12. After a very exhaustive investigation, a comprehensive amended scheme was established.
13. The movement of people around the world stimulates growth.
14. A private monopoly frequently acts against the interests of the consumers.
15. Thus cross-subsidisation becomes a reality.
16. I have understanding for the fact that we have to protect all the resources and that weather conditions, seasons and those things have an influence on our coastal resources.
17. I think that we are entering a period where expectations will be created again.
18. The department retains its powers in regard to the allocation and appointment of water from the main source, and this water will also be made available to the board at a specific tariff.
19. I should like to begin by paying tribute, as is customary, to a senior official who has retired after many years of service.
20. We know that limited access to the accounts and investments of suspected dealers frequently hampers investigations.

Table 4: Sample of the English corpus

The performance of the system was measured in terms of the number of the translation pairs correctly identified. For each word-pair the likelihood factor was calculated and the grammar rules were applied to assign a grammatical type to each translation pair.

Results and Discussion

The performance of the system was dependent on the sentence length, the number of unmatched words present in the sentence as well as the complexity of the sentence. The results varied according to the circumstances. The various performance results are:

- The system could identify the translation pairs and the possible grammatical categories to which each of them belonged, if the sentences contained only a few (three or less words out of ten words per sentence) unmatched words, irrespective of the sentence length. For example, for the sentence pair

> *Ek wil net vir die agb Minister vra of sy departement betrokke is by die opstel van sulke programme en of hy enige onderwysdepartemente genader het.*

and

> *I simply want to ask the hon Minister whether his department is involved in the installation of such programmes and whether he has approached any education departments.*

five words were unmatched and the target sentence length was twenty-seven words. Thus 18% of the words were unmatched. The word-pair <*net, simply*> was identified as a translation pair purely on the value of the likelihood factor since the grammatical type of the pair could not be determined by the grammar rules. The following word-pairs as well as their grammatical type were correctly identified:

- <*agb, hon*> and grammatical type of ***adjective,***
- <*opstel, installation*> and grammatical type of ***noun,***
- <*programme, programmes*> and grammatical type of ***noun*** and
- <*onderwysdepartemente, education departments*> and grammatical type of ***noun + noun.***

Thus, a success rate of 80% was achieved for this specific example.

- As was to be expected, the system did not perform very well when the source and target sentences were very long and contained many unmatched words. For example, for the sentence pair

> *Ek wil graag begin deur 'n woord van hulde te spreek soos dit al die gebruik is wanneer 'n senior amptenaar aftree na lang jare van diens.*

and

> *I should like to begin by paying tribute, as is customary, to a senior official who has retired after many years of service.*

fourteen words were unmatched and the target sentence length was twenty-three words. Thus 57% of the words were unmatched. The system was able only to identify five (23%) translation pairs and their grammatical types correctly. The grammar rules were not very successful in determining the grammatical types, since most of the grammatical types of the neighbouring word(s) of the unmatched word were unknown.

- The system performed rather well if the grammatical types of the neighbouring words of the unmatched word were known. Unfortunately the grammar rules are less successful if the neighbouring words are also unmatched. For example, for the sentence pair

Ons weet dat beperkte toegang tot die rekeninge en beleggings van vermeende handelaars dikwels ondersoeke kortwiek.

and

We know that limited access to the accounts and investments of suspected dealers frequently hampers investigations.

the grammatical types of all of the neighbouring words of the unmatched words were known. The unmatched word-pairs for this specific example were: <*beperkte, limited*>, <*rekeninge, accounts*>, <*vermeende, suspected*> and <*kortwiek, hampers*>. The system was successful in identifying the correct grammatical types for the four word-pairs.

For the sentence pair

Die skrapping van hierdie betrokke wet het my volle ondersteuning.

and

The abolition of this specific act has my full support.

the grammatical types of all of the neighbouring words of the unmatched words were not known. The unmatched word-pairs for this specific example were: <*skrapping, abolition*>, <*betrokke, specific*>, <*wet, act*> and <*volle, full*>. The grammar rules were unable to determine the grammatical types of the unmatched word-pairs <*betrokke, specific*> and <*wet, act*> which follow each other. However, the system could identify the correct grammatical types for the other two word-pairs which were positioned elsewhere in the sentence and were surrounded by words of known grammatical types.

Table 5 is a graphical representation of the results obtained from the experiment as described above. The x-axis (horizontal axis) shows the percentage success rate that was achieved. The formula for the x-axis (horizontal axis) is as follows:

$$\% \textit{ correct} = (x / y) * 100$$

where x = *number of word-pairs correctly identified* and y = *number of unmatched words.*

The y-axis (vertical axis) indicates the percentage unmatched words per sentence. The formula for the y-axis (vertical axis) is as follows:

$$\% \; unknown/length = (y / l) * 100$$

where *y = number of unmatched words* and *l = length of sentence.*

The graph confirmed the expectation that the system is less successful in handling very long or very short sentences containing many unmatched words. The graph also implies a higher success rate if few unmatched words were present irrespective of the sentence length.

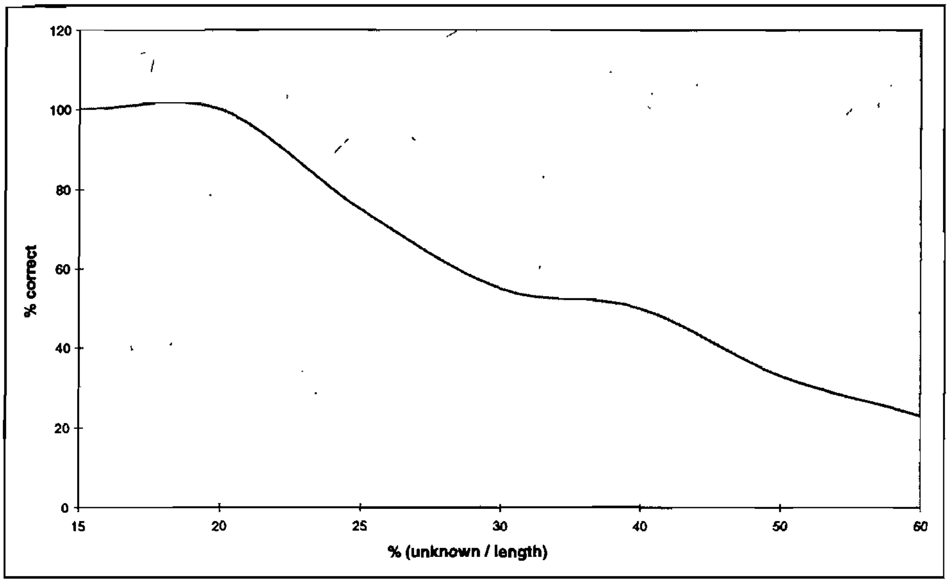

Table 5: Statistics for identification of word pairs

Conclusions

The results were encouraging and it would seem that the system is more successful in identifying unmatched words and determining their grammatical type if there are not too many unmatched words per sentence. Since the purpose of the system is to extend an existing lexicon and not to produce a new lexicon from scratch, the system will perform better if most of the words in the sentences already have entries in the lexicon and the system only has to identify a few new words per sentence.

7. Future Research and Concluding Remarks

The object of this project was to examine the hypothesis that parallel corpora, in this case the English-Afrikaans corpus, can be utilized to extend an existing bilingual lexicon, in this case either the English-Afrikaans or the Afrikaans-English lexicon. To investigate and demonstrate the validity of this argument, a simple system was developed which was applied to sentence-aligned samples of parallel texts extracted from the South African Hansards.

Encouraging results were obtained from the experiments conducted. The results confirmed the expectation that sentences, irrespective of length, containing many unmatched words, would yield a lower success rate than sentences containing only a few unmatched words. Several problems were identified, but further research is necessary to clarify these issues and to propose possible solutions. Some of these issues as well as other possible future research topics are discussed in the next section.

7.1 Future Research

The alignment of the English and Afrikaans corpora was done at sentence level. The alignment process does not have to stop at this level and it is possible to determine the most probable word-pair alignments. Phrasal alignment is another possibility and can be achieved by phrasal parsing and the utilization of phrasal information. Although the success rate of these probabilistic techniques is dependent on the size and the quality of the corpus, the English-Afrikaans corpus satisfies these dependencies. It is difficult to align translations on the basis of words and to achieve this goal, a tagged corpus will be a valuable resource. Some specific issues identified that require further investigation, are:

- The simple grammar rules that were implemented for the identification of the grammatical categories can be refined in order to achieve more accurate results.
- Problems exist in determining the grammatical category for the unmatched word-pair if the neighbouring source and target words are also unmatched and the grammatical categories are not related. For example, for the sentence *Daar is 'n behoefte aan informele programme* the simple grammar rules were unable to assign the grammatical type of ***adjective*** to the unmatched word, *informele,* and ***noun*** to the unmatched word, *programme.*
- The treatment of verbs in terms of the difference in position in the sentence for different languages, for example in Afrikaans the verb often occurs to the end of the sentence while this is not usually the case in English.
- Difficulties with split verbs, for example *vat ... saam.*
- The treatment of compound prepositions, for example *out of.*
- The treatment of complex prepositions, for example *in place of.*

A tagged corpus would also add value to the system, since the linguistic information provided by the tags can be used to refine the process of determining the likelihood that word-pairs in parallel sentences are translations of each other and to ascertain the grammatical category of the word-pair. A tagged corpus can be utilized for deriving linguistic rules which can be incorporated into this system or into an existing machine translation system and thereby contribute to the improvement of the quality of the product. As the tags provide part-of-speech information, the tagged corpus can be used as a resource to study actual language-usage in English and Afrikaans.

7.2 Concluding Remarks

Although the system is very simple and some problems still exist, the results were encouraging. The refinement of the grammar rules and the incorporation of more linguistic knowledge should improve the results, while simultaneously reducing the cost of human inspection of the newly identified translation pairs. It would also be interesting to investigate the use of tagging and the effect it will have on the results. The system proved valuable in that it supported the extension of an existing bilingual lexicon with reduced human effort.

Bibliography

Armstrong, S. (Ed.). 1994. *Using Large Corpora.* Cambridge, Massachusetts: MIT Press.

Arnold, D., L. Balkan, R.L. Humphreys, S. Meijer and L. Sadler. 1994. *Machine Translation — An Introductory Guide.* Manchester: Blackwell Publishers.

Brown, P.F., J.C. Lai and R.L. Mercer. 1991. Aligning Sentences in Parallel Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (Berkeley).* Berkeley: University of California.

Brown, P.F., J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer and P.S. Roosin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, J.D. Lafferty and R.L. Mercer. 1992. Analysis, Statistical Transfer and Synthesis in Machine Translation. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal.*

Carbonell, J.G., T. Mitamura and E.H. Nyberg. 1992. The KANT Perspective: A Critique of Pure Transfer. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal.*

Catizone, R., G. Russel and S. Warwick. 1989. Deriving Translation Data from Bilingual Texts. *Proceedings of the First International Acquisition Workshop, Detroit.*

Chang, J-C. and H. Chen. 1995. Using Partially Aligned Parallel Text and Part-of-Speech Information in Word Alignment. *Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia.*

Chomsky, N. 1962. A Transformational Approach to Syntax. Hill, A.A. (Ed.). *Proceedings of the 1958 Conference on Problems of Linguistic Analysis in English.* Austin: University of Texas.

Gale, W.A. and K.W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (Berkeley).* Berkeley: University of California.

Gale, W.A., K.W. Church and D. Yarowsky. 1992. Using Bilingual Materials to Develop Word Sense Disambiguation Methods. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal.*

Gale, W.A., K.W. Church and D. Yarowsky. 1993. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and Humanities* 26.

Garside, R., G. Leech and G. Sampson. 1987. *The Computational Analysis of English — A Corpus-based Approach.* London: Longman.

Hutchins, W.J. 1986. *Machine Translation: Past, Present, Future.* Chichester: Ellis Horwood.

Hutchins, W.J. and H.L. Somers. 1992. *An Introduction to Machine Translation.* London: Academic Press.

Klavans, J. and E. Tzoukermann. 1996. Combining Corpus and Machine-Readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation* 10.

Maclovitch, E. 1992. Where the Tagger Falters. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal.*

Mitamura, T. and E.H. Nyberg. 1995. Controlled English for Knowledge-Based MT: Experience with the KANT System. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven.*

Pienaar, J. 1996. The Utilization of Parallel Corpora for the Extension of Machine Translation Lexicons. Unpublished M.Sc. thesis. Pretoria: University of Pretoria.

Renouf, A. 1987. Corpus Development. Sinclair, J.M. (Ed.). *Looking Up: An Account of the COBUILD Project in Lexical Computing.* London: Collins.

Sato, S. and M. Nagao. 1990. Toward Memory-Based Translation. *COLING-90: Papers presented to the 13th International Conference on Computational Linguistics (Helsinki).* Helsinki: Helsinki University.

Sebba, M. 1991. The Adequacy of Corpora in Machine Translation. *Applied Computer Translation* 1.

Simard, M., G.F. Foster and P. Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal.*

Smadja, F. 1992. How to Compile a Bilingual Collocation Lexicon Automatically. *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques, San Jose, CA.* San Jose, California: AAAI Press.

Somers, H.L. 1993. Current Research in Machine Translation. *Machine Translation* 7.

Su, K. and J. Chang. 1990. Some Key Issues in Designing MT Systems. *Machine Translation* 5.

Su, K. and J. Chang. 1992. Why Corpus-Based Statistics-Oriented Machine Translation? *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montréal.*

Sumita, E., H. Iida and H. Kohyama. 1990. Translating with Examples: A New Approach to Machine Translation. *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (Austin TX).* Austin: University of Texas.

Weaver, W. 1955. Translation. Booth, A.D. and W.D. Locke (Eds.). *Machine Translation of Languages.* Cambridge, Massachusetts: MIT Press.