Creating a reusable English – Afrikaans parallel corpora for bilingual dictionary construction

Aldin Draghoender

Mattias Kanhov

Department of Computer and Systems Sciences (DSV)

Degree project 15 credits
Computer and Systems Sciences
Degree project at the bachelor level
Spring semester 2010
Examiner: Hercules Dalianis
Swedish title: Ordkoppling av Afrikaans och Engelska med Uplug

Stockholm University

Creating a reusable English – Afrikaans parallel corpora for bilingual dictionary construction

Aldin Draghoender

Mattias Kanhov

Abstract

Computers have been used for many years to create dictionaries, translate texts or just to looking up single words. The most common way to create electronic dictionaries is to use parallel texts, known as parallel corpora (corpus in singular form). When processing the corpora, software tools use word alignment, sentence alignment and other techniques to create an accurate dictionary. Today, there are quite a few parallel corpora available for creating dictionaries in the major languages however there do not seem to be as many available for the English – Afrikaans language pair.

This thesis investigates the possibilities in creating a bilingual English – Afrikaans dictionary by building a parallel corpus and using the Uplug tool to process it. The resulting parallel corpus with approximately 400,000 words per language was created partly from texts collected from the South African government and partly from the OPUS corpus. The recall and accuracy of the bilingual dictionary was evaluated based on the statistical data collected. Samples of translations were generated, compiled as questionnaires and then assessed by English – Afrikaans speaking respondents. The results yielded an accuracy of 87.2 percent and a recall of 67.3 percent for the processed dictionary.

Our English – Afrikaans parallel corpora can be found at the following address: http://www.let.rug.nl/tiedeman/OPUS/

Keywords: Uplug, Parallel Corpora, Word Alignment, English, Afrikaans

Acknowledgements

We would like to thank Hercules Dalianis, Elin Carlsson and Maria Skeppstedt for their guidance and support during this period. We would also like to extend our gratitude and appreciation to Jörg Tiedemann for his support both with Uplug and corpora. Our evaluation and conclusions would also not be possible without the people who took their time to fill out the translation questionnaires and therefore we are grateful to them as well.

1. CONTENTS

Appendix list

List of Figures

List of Tables

1. INTRODUCTION

Web 2.0, with its new ideas for user interactions on the web (Anderson, 2007), has evolved the Internet to a more contribution oriented area with improved interaction and participation methods that have opened a whole new way of communication with a world of information. Whether it's for business intelligence in companies, shopping or for communicating in social websites such as Facebook or Twitter, the Internet has become the largest information source making need for multilingual information retrieval more critical.

Multilingual countries or in the case of our project, South Africa, a country with eleven official languages in which most of the population only speaks a small percentage of all the languages, could certainly benefit from retrieving information in different languages (Trushkina, 2006). Therefore the need of a multilingual dictionary is of great importance.

We will during the course of this thesis create a parallel corpus, run the corpus through Uplug to generate an English - Afrikaans dictionary which we then will evaluate. For a lot of languages there are large compiled parallel corpora already available, but in the case of Afrikaans there are very few public available resources. Because of the lack of parallel corpora, we will find multiple bilingual texts in English – Afrikaans and create our own.

The tool chosen for creating the bilingual dictionary with the parallel corpora is the Uplug system, which is essentially a Perl script containing a set of language processing tools (Tiedemann, 1999).

1.1. PROBLEM

The problem this thesis will deal with is the lack of parallel corpora and dictionaries for the English - Afrikaans language pair.

1.2. GOAL

The goal of this thesis is to build a parallel corpus from parallel texts collected over the Internet. Uplug will then be used to create a bilingual English – Afrikaans dictionary from this corpus. Furthermore, the dictionary will be evaluated to assess the recall and accuracy. The resulting English - Afrikaans parallel corpus will be published and made available for other language researchers

.

1.3. PURPOSE

The purpose of this thesis is to create a bilingual English – Afrikaans parallel corpus covering several domains. This corpus will be published so that other language researchers may use it for further research. The dictionary that will be created could be used as a base for other researchers to further develop a multilingual dictionary, catering to all South African languages. Another usage of the dictionary could be in search engines and other information retrieval applications where translations would be helpful. It is also possible to use one of the two languages in the dictionary as a pivot language (Wu and Wang, 2008) when translating to or from a third one. For example when there is a text in Afrikaans that needs to be translated into Xhosa, and there is an Afrikaans – English and a Xhosa – English dictionary available, English will be the pivot language that acts as a bridge between Afrikaans and Xhosa.

1.4. LIMITS

The language domain will depend on the English – Afrikaans corpus found. The amount of words aimed for is about 400,000 per language, however depending on time constraints a corpus half the size should be sufficient.

1.5. METHOD

Because there is no English – Afrikaans corpus readily available, one will be created manually by gathering parallel texts and compiling them into one raw text file. This is a very time consuming process that benefits greatly from the person creating the corpus knowing both the languages. The domain of the corpus depends on the parallel texts that are available.

Some governments publish parallel texts that are useful when creating parallel corpora. In addition to government documents, other parallel texts will have to be found in order to reach a proper sized corpus. Most corpora for larger languages contain many million words, but the goal for this thesis is a corpus with around 400,000 per language. Different methods of parallel corpus gathering and creation will also be examined, for example the STRAND bilingual database (Resnik and Smith, 2003) which is a system for finding parallel documents in the preferred languages. Since a corpus containing a la,rge number of words is essential for creating an accurate dictionary with high recall, much effort will be spent in this process.

To reach the other goal of the thesis, Uplug will be used in combination with the parallel corpus to generate an English – Afrikaans dictionary. Uplug will be run in a Linux environment provided by - Department of Computer and Systems Sciences (DSV) via remote access. Uplug was chosen as corpus processer because of its ease of use, good reputation and popularity within language research.

Once a working dictionary has been produced, it will be cleaned from words with frequency of 2 or less as they are not seen as reliable. Also single characters which are not words will be removed; these are numbers and punctuation marks, leaving only words.

Random text samples from different domains and topics will be collected and used to measure:

1. *English words are found in the dictionary.* Every word in the texts will be looked up in the dictionary to see if they are present without taking into account if the translations are correct or not. This is done to measure the effect of cleaning the dictionary.

2. *Accuracy*. To measure the accuracy, every unique word in the sample texts will be compiled into a document and then the translations of each word will be added. The list will then be given to English – Afrikaans speaking people who will judge the accuracy of the translations by filling in questionnaires asking if the sampled word translations are *correct, partly correct or wrong.* Only words considered correct will be used in determining the accuracy, not partly correct words. Because it will be challenging to find enough Afrikaans – English speaking people who can evaluate the translations, Google Translate will also be used as an evaluator. Statistics will then be created to summarize the total accuracy of the dictionary.

3. *The recall*. Recall will measure the amount of correctly translated words from the sample texts that are present in the dictionary.

Note that when calculating English words found, accuracy and recall, every instance of a word will be counted. This includes both unique and duplicate words. The words in the evaluation questionnaires are unique to minimize the work done by the respondents.

English words found, recall and accuracy for the original and cleaned dictionary will be compared to examine the relationship between the three statistical classification units. The difference of recall for original and cleaned dictionary as well as accuracy for original and cleaned dictionary will be investigated to see how they influence each other.

1.6. TERMINOLOGY

Accuracy – the amount of word from the sample texts found in the dictionary that are correctly translated.

Alignment – is the process of matching texts in different languages so that paragraphs, sentences or words correspond to each other on the same level.

Corpus/Corpora (plural) – implies a body of stored texts that is either written and/or a transcription of recorded spoken language.

Lemma – is normally the head words in a dictionary. They represent lexemes which are the set of all forms that have the same meaning of a word. For example in: *"go", "goes", "going", went* and *"gone", "go"* is the lemma form.

Lemmatizer – a tool that associates a group of words with its base form (lemma).

Inflection – is variation of the form in a word, often an affix (the ending of the word) is added. Example: walk, walked

Morpheme - Morphemes are meaningful grammatical units consisting of a word, such as *fan*, or word element, such as *–ed* in worked, which cannot be divided into smaller grammatical parts.

Parallel corpora - are pairs of texts which contain data in a main language and a translation thereof.

Part of speech tagger – is a tool for categorizing words based on its synthetic use. Examples of categories are nouns, verbs and adjectives.

Recall - will measure how many correctly translated words are found in the sample texts.

Uplug - is software with the purpose of providing a modular platform for the integration of text processing tools.

1.7. DISPOSITION

Below is an overview of how the chapters of this thesis are presented.

Chapter 2 – Extended background

This chapter introduces a background to the main concepts and terms of the thesis. It starts with a brief background and history of Afrikaans, explaining some differences and similarities between English and Afrikaans. Then a definition of corpora is provided then an in depth discussion on parallel corpora and methods of creating it follows. Alignment and different methods are then described followed by a description of the Uplug system. Related research provides and describes similar research by other authors in brief.

Chapter 3 - Investigation

This chapter describes how the corpus was created and the stages involved from preparing the corpus to running Uplug to evaluating the dictionary.

Chapter 4 - Results

This chapter provides a description of the results reached from evaluating the dictionary.

Chapter 5 - Analysis

This chapter shows an analysis of the results reached.

Chapter 6 - Conclusions

This chapter describes our conclusions of the work done and the final results that were reached. We also present some proposals for future work at the end.

2. EXTENDED BACKGROUND

2.1. THE LANGUAGE AFRIKAANS

Afrikaans is a rather young language predominately spoken in South Africa and to a lesser extent in neighboring countries. However, with migration and people working more internationally over periods of time, it is not uncommon to hear of smaller communities speaking Afrikaans in countries such as Canada, England and Australia. There are many disputes surrounding the origin of Afrikaans however it is associated with the arrival of the Dutch in the Cape Town in 1652 (Baldauf and Kaplan, 2004). They also state that the origins of Afrikaans can be attributed to several sources but mainly Dutch as the language structure constitutes ninety percent of that of Dutch. Other sources are Khoi, Southern Bantu languages French, German, Portuguese, Malay and English. Even though Afrikaans originated around that period in South Africa, it was not until 1925 that it became an official language alongside English (Donaldson, 1993). This being said, even though Afrikaans was the first language of many, English was still the preferred language of communication outside the home.

Both Afrikaans and English belong to the West Germanic branch of the Indo-European language family and therefore share quite a few similarities such as not having gender distinction of nouns and the grammar articles are basically used in the same manor where the definite article "*die*" in Afrikaans equals "*the*" in English and the indefinite article "'n" in Afrikaans equals "a/an" in English (Engelbrecht and Schultz, 2005).

Modern English and Afrikaans like other Germanic languages have lost a many of their distinguishing paradigmatic verbal inflections, so much in Afrikaans that the finite verb lost all inflection (Bennis and MacLean, 2006). With regard to the syntax of Afrikaans, Bobaljik (1995) states that "it behaves like the most richly inflected languages, yet it is the most poorly inflected of all", this being due to the fact that the subject-verb tense agreement shows no inflection. However there is the prefix *ge-* which denotes most past participles and the conjugation the verbs: WEES (to be) and Hê (to have) which has a present, a past and a future (Biberauer, 2002). Table 1 below shows inflection of verbs in Afrikaans.

	wees - "to be"	*hê* - "to have"
Infinitive/Imperative	Wees	hê/het
Present Tense	is	Het
Past Tense	was	Het
Past Participle	gewees	Gehad

Table 1: Inflecting verbs in Afrikaans

The word order in Afrikaans is similar to English:

Subject + verb + object

But for the subordinate clauses it is:

Subject + object + verb

And when the subordinate clauses precede the main clause, the order is (Pretorius and Schwitter, 2009):

Verb + subject + object

Another interesting characteristic of Afrikaans is the use of negative. This is when two negative elements as pair constitute a single instance of negation concord (van Gass, 2007). van Gass (2007) also gives an example below:

*”(1) Haar suster het **nie** haar verjaarsdag vergeet **nie**.*

Her sister have not her birthday forgotten NEG

"Her sister didn't forget her birthday."

*(2) Hy het **nooit** sy broer vergewe **nie**.*

He have never his brother forgiven NEG

"He never forgave his brother." “

2.2. PARALLEL CORPORA

In linguistics, the word corpus or corpora (plural) is used to imply a body of stored texts that is either written and/or a transcription of recorded spoken language (Krieger, 2003), which provide a fair representation of a language. Given that it is stored in electronic form, vast amount of information are available for research in statistical machine translation. These collections of texts are used to better understand languages and normally provide information of a statistical or a quantitative nature (Gries, 2009).

Corpora is said to be either comparable or parallel and is used for contrastive, monolingual and translation studies. While the sampling frame for comparable corpora is important, where the components of the languages have to match in terms of proportion, genre, domain and sampling period, for parallel corpora the sampling frame is not relevant as components of the corpora are direct translations of each other (McEnry and Xio, 2007). For the purpose of this project and thesis, parallel corpora will be used.

Parallel corpora are pairs of texts which contain data in a main language and a translation thereof thus making it perfect for translation studies. These texts can also be bilingual or

multilingual. Parallel texts are often found in multilingual or bilingual governments (Koehn, 2005) such as the South African (*English, Afrikaans, isiNdebele, isiXhosa, isiZulu, Sesotho sa Leboa, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga*), Hong Kong (*English, Chinese*) and the Canadian government (*French, English*). Other sources include institutions such as the United Nations and the European Union.

Quite a few parallel corpora have become available in recent years but copyright laws still make it difficult to produce more readily available corpora. The OPUS corpus (Tiedemann and Nygaard, 2004) is an example of a multilingual parallel corpora based on translated open source documents. Another noteworthy corpus is the JRC- Acquis Multilingual Parallel Corpus (Steinberger et al., 2006) which consists of more than twenty European languages and has a hundred and twenty language pairs making it the most multilingual corpus available today. There is also the Europarl Corpus (Koehn, 2002) which was extracted from the Proceedings of the European Parliament. The corpus is preprocessed and consists of more than twenty million words for each of the eleven languages. As it is difficult to find parallel corpora for English – Afrikaans available, one will probably have to be created in order to create a dictionary.

2.3. CORPORA AND THE WEB

Lüdeling et al. (2006) states that readily available corpora satisfy many research questions, however not all can be solved with the available corpora for various reasons such as data not being in the available corpus because the domain is not represented by the corpus. In these instances the Web is a good resource for building a corpus. Resnik and Smith (2002) describes a detailed approach for STRAND, a Web resource for finding parallel text by locating pages, generating candidate pair and filtering out non-translation candidate pairs. A similar approach is taken by Lüdeling et al. (2006) only this time instead of just looking at the possibilities, emphasis is also given on its limitations. Sinclair (2005) also provides some commonly used criteria for building a corpus. These criteria have to be reflected upon in order to create a suitable corpus:

> *"1. The mode of the text, whether the language originates in speech or writing, or in electronic mode.*
>
> *2. The type of text, for example if written, whether a book, a journal, a notice or a letter.*
>
> *3. The domain of the text, for example whether academic or popular.*
>
> *4. The language or languages or language varieties of the corpus.*
>
> *5. The location of the texts, for example (the English of) UK or Australia.*
>
> *6. The date of the texts."* (Sinclair, 2005).

For the purpose of this thesis, STRAND will not be used because the lack of support for the language Afrikaans. However the web will be used as a source for manually collecting parallel texts.

2.4. ALIGNMENT

The alignment of text has a rather large impact on the quality of dictionaries. The alignment of texts can be on different levels; these include paragraph level, sentence level, phrase level and at word level (Charitakis, 2006) however according to Romary et al. (1995), if every word is not necessary to be marked, it is optimal to stop at sentence level. This is due to the fact that the observation/study of words can be made or understood in contexts that are reasonably complete. Brown et al. (1991) proposed a method for sentence alignment using internal information and not making assumptions of the sentence's lexical structure. For their method, they assumed that there is a correlation between the length of sentences in the source and target texts meaning that the lengths of the sentences are normally the same. Furthermore, they concluded that two languages have a fixed ratio of sentence lengths when the numbers of words or characters are counted. Another method was proposed by Kay and Röscheisen (1993) where it is assumed that the words of a translated sentence must correspond for the sentence to correspond. Here only internal information is used meaning that information such as lexical mapping is derived from texts that will be aligned (Véronis, 2000).

Tiedemann (2003) states that there are generally two approaches to word alignment. One approach is the estimation approach commonly used in statistical machine translation where the alignment parameters are modeled as hidden. This hidden alignment shows the connections from the source to the target position (Och and Ney, 2003). On the other hand there is the association approach which is used in the extraction of bilingual dictionaries. In this method heuristics are used (Charitakis, 2006), the alignment is achieved by similarity measures and association tests.

For the purposes of our project which is creating a dictionary, it is important to explore every word and therefore we will have to do alignment at the lowest level which is word alignment.

2.5. THE UPLUG SYSTEM

The Uplug system is an application software with the purpose of providing a modular platform for the integration of text processing tools. The software is also a part of the PLUG project, which is acronym for ***P**arallel corpora in **L**inköping, **U**ppsala, **G**öteborg* and was initially intended for the processing of the project's bilingual texts. There are three main components to this system of which all are designed to be extensible and multi-purpose. The components are:

UplugIO - consists of an I/O library with a transparent interface for working and accessing specific data and a toolbox for integrating different data formats.

UplugSystem - a launcher for combining Uplug modules into sequentially executable systems. These modules can be any external software tools.

UplugGUI – a graphical user interface comprising of a set of tools for the construction, configuration and application of Uplug systems (Tiedemann, 1999).

Uplug consists of a collection of modular tools written in Perl which is used for word alignment, sentence alignment, POS-tagging, term extraction from parallel corpora and much more (Uplug, 2010). Since the system is modular, modules can be chosen and configured for the desired purpose and their running sequences can also be specified. Uplug is available as an online service as well as a standalone application.

Examples of tools that are integrated with Uplug include pre-processing tools such as a sentence splitter, tokenizer and external part-of-speech tagger with wrappers and shallow parsers. The standard package has the following external tools included: the *Grok system* for English tagging and chunking and the morphological analyzer *ChaSen* for Japanese. The *Gale & Church* approaches can be used to sentence align the translated documents while the *Clue alignment approach* and *Giza++* can be used to align words and phrases. Other tools that can be integrated are the *Tree Tagger* for English, French, Italian, and German and the *TnT tagger* for English, German and Swedish (Uplug, 2010). A brief explanation of these tools and approaches follows:

Grok system

Grok is the acronym for ***G**rammatical **R**epresentation of **O**bjective **K**nowledge* (Obermeier, 86). It is an open source library comprising of natural processing components. These components include parsing with categorical grammar and preprocessing tasks such as part-of-speech tagging, tokenization and sentence detection. Grok also provide implementations for most of the interfaces in the OpenNLP project (Baldridge, 2001).

ChaSen

ChaSen is a morphological analyzer for the computational analysis of Japanese texts. It is based on the Japanese morphological analyzer JUMAN version 2.0, which was developed at Nagao Laboratory of Kyoto University and the Graduate School of Information Science at Nara Institute of Science and Technology. ChaSen was developed as a common tool for the analysis of Japanese as there is no widely agreed upon grammatical terminology. Another reason for its development is to recognize the first individual morphemes of input sentences when doing computational analysis (Matsumoto et al., 2007).

Gale & Church sentence alignment

The Gale & Church (1991) length based sentence alignment technique is a method proposed for aligning a bilingual corpus at sentence level. It is based on a probabilistic model where the authors found evidence that the correlation between the length of paragraphs in characters and their length of its translation were very high. This observation was made out of the idea that longer paragraphs have longer translations and likewise, shorter paragraphs have shorter translations.

Clue alignment

The clue alignment approach is a word alignment technique that was presented by Jörg Tiedemann (2003). This technique is based on a combination of association clues that indicate association between words and phrases. These clues are typically based on characteristics such as frequency, part-of-speech and phrase type and can be calculated by measuring similarity values (Tiedemann, 2003).

GIZA++

GIZA++ is an extension of GIZA (a part of the EGYPT Statistical Machine Translation Toolkit) which was developed at the Center for Language and Speech Processing at John Hopkins University. GIZA is a training software tool that learns statistical translation models from bilingual corpora while GIZA++ provide extensions such as alignment models dependent of word classes and statistical algorithms for alignment (Och, 2001).

TnT tagger

TnT tagger is a statistical part-of-speech tagger that is optimized for training on large corpora rather than being trained for a specific language. TnT is the shortened form of Trigrams'n'Tags and is trainable for most tagsets. The program has a number of techniques for handling and smoothing unknown words and generates parameters by training on tagged corpora (Brants, 2000).

Tree Tagger

Tree tagger is a software tool used for annotating text with part-of-speech and lemma information and has been successfully used in quite a few languages (Schmid, 2008).

2.6. RELATED RESEARCH

Word alignment with Uplug was used in Xing and Zhang (2008) for the alignment of a Chinese – English parallel corpus. Their parallel corpora were composed from legal documents which contained 104,563 Chinese characters, an equivalent to 50,000-60,000 Chinese words and 75,997 English words. From the extracted 2,118 word pairs with a frequency equal or above 3, a sample of 800 word pairs was used for evaluation. Ten Chinese speaking people read the sample translations and answered if they thought the translations were right, wrong or if they could not decide. The result of the evaluation gave an average accuracy of 74 percent of correct translated terms.

Dalianis et al (2009) used Uplug for word alignment to adapt a website search engine in cross language information retrieval. News articles from the Nordic council website were used to create a Swedish-Danish-Norwegian dictionary where English was used as a pivot language (Dalianis et al., 2009). For evaluation, the Swedish and English corpus was used with part- of- speech (POS) tags and without POS-tags. POS tagging is the process of classifying, dividing and tagging words depending on placement in sentences or paragraphs. The result yielded a precision of 71 percent average frequency without POS-tags and a 67 percent average frequency with POS tags. The average recall was 92.5 percent frequency without POS-tags and a 91.3 percent average frequency with POS tags, which also concluded that POS-tags do not improve word alignment.

A similar experiment was done in Velupillai and Dalianis (2008) where Uplug was used to create a domain specific dictionary for Nordic languages. The domain was mobility information in Nordic countries and contained sparse corpora of less than 80 000 words per language pair. This resulted in ten different dictionaries of the Nordic languages which were defined as Swedish, Danish, Norwegian, Icelandic and Finnish. Swedish-Danish, Swedish-Norwegian and Danish-Norwegian gave good results after word alignment with an average of 93 percent while Finnish only gave 67 percent. It is also worth noting that words less than six characters and multiword expressions were not included in the final wordlists.

A Greek-English dictionary was created by aligning words with Uplug in Charitakis (2007, 2008). The Greek-English corpus comprised about 200 000 words per language. The conclusion based on their quality was that 51 percent of the translations were correct while with higher frequency (f>11) 67 percent was achieved. Uplug was also used in Megyesi and Dahlqvist (2007) for word alignment with a Swedish-Turkish parallel corpus. Their results showed 69 percent correctly aligned words and from the error margin, 61 percent was due to grammatical differences. The corpus was based on 150 000 Swedish words and 126 000 words in Turkish.

Even though bilingual corpora have been successfully experimented with for word alignment in different languages (Somers, 2001) not much emphasis has been put into a minor language such as Afrikaans. The most relevant work with English to Afrikaans translations that has been done is a two-way speech-to-speech translator developed by Engelbrecht and Schultz (2005). Their thesis explains the language Afrikaans, the system architecture/

development and also their results. As it is a speech translator the results are not that valuable for creating text translations.

Another interesting thesis is one written by Kato and Bernard (2007) that investigates ways to reduce the size of parallel bilingual texts while still being able to maintain an accurate dictionary. This could be very useful because it is often hard to find suitable parallel corpus needed for machine translations and a lot of effort often goes into creating one. They use both active learning, which creates better classification systems with less data, and semi-supervised learning, which label's and predicts words and then remove the most confident ones, to achieve their goal.

3. INVESTIGATION

When creating a dictionary, the whole process can be viewed as a series of steps; from the creation of corpus to running the alignment software to processing and "cleaning" the dictionary and then finally evaluating the resulting dictionary. In the case of inadequate evaluation results, reconfigurations will have to be made to one or more steps. Then the subsequent steps will have to be repeated until the desired results are achieved.

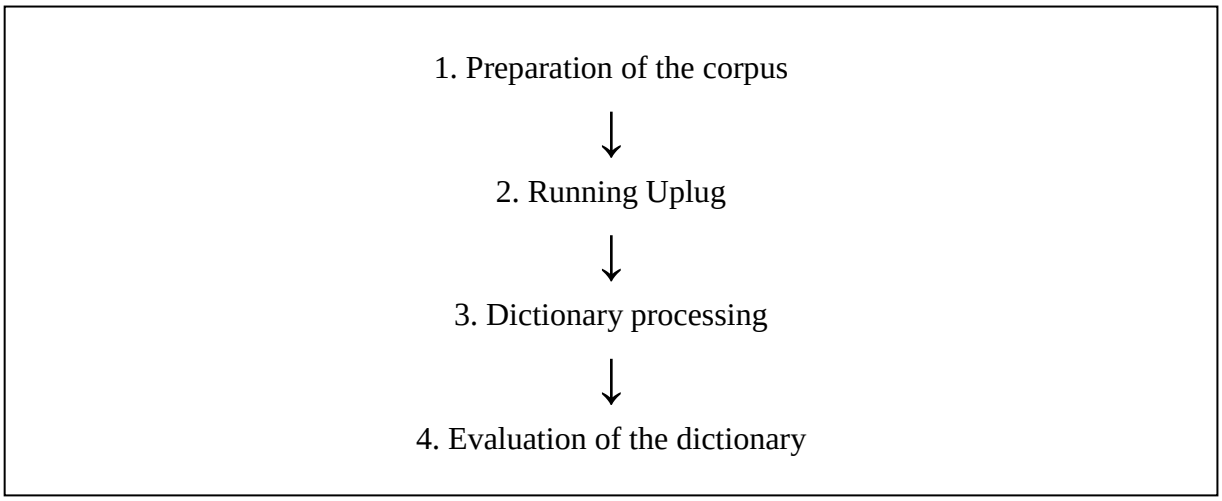


Figure 1: Dictionary creation steps

In this section the four steps in Figure 1 are represented.

3.1. STEP 1: PREPARATION OF THE CORPUS

The corpus for this thesis was created partly from The OPUS corpus (Tiedemann and Nygaard, 2004) and partly from a parallel corpus that we created from sixteen bilingual publications from the South African government (South African Government Information, 2010). The publications were converted from PDF form to plain text, then proof read to remove faulty converted tokens/characters, then manually aligned and finally placed in a corpus text file with an ID-tag assigned to it and saved as a raw text file encoded with UTF-8.

When dealing with the enormous amount of data that the corpora contain, Uplug needs to divide the texts into multiple sections in order to function properly. To inform Uplug of where in can divide texts, a series of blank lines must be added, preferably between two texts. The blank lines were entered between each of the sixteen texts.

The final corpus contained a total of 819,344 words with 421,587 Afrikaans words and 397,757 English words respectively. This ratio could be due to the fact that some Afrikaans sentences were longer than their English translations or because the English language might have more compound words than Afrikaans. The domain for the corpus was a blend of law-,

technical- and public government documents with around 200,000 words (per language) that were government- and law texts while the remaining 200,000 words (per language) originated from the OPUS corpus and were technical terms from open source software manuals. With the relatively large size of the corpus the recall for translations within the domain should be fairly high.

3.2. STEP 2: RUNNING UPLUG

A script file was written for Uplug to run so that the whole process could be automated. The English corpus was set to be aligned with the Afrikaans corpus. Uplug was configured to run according to figure 2 below.

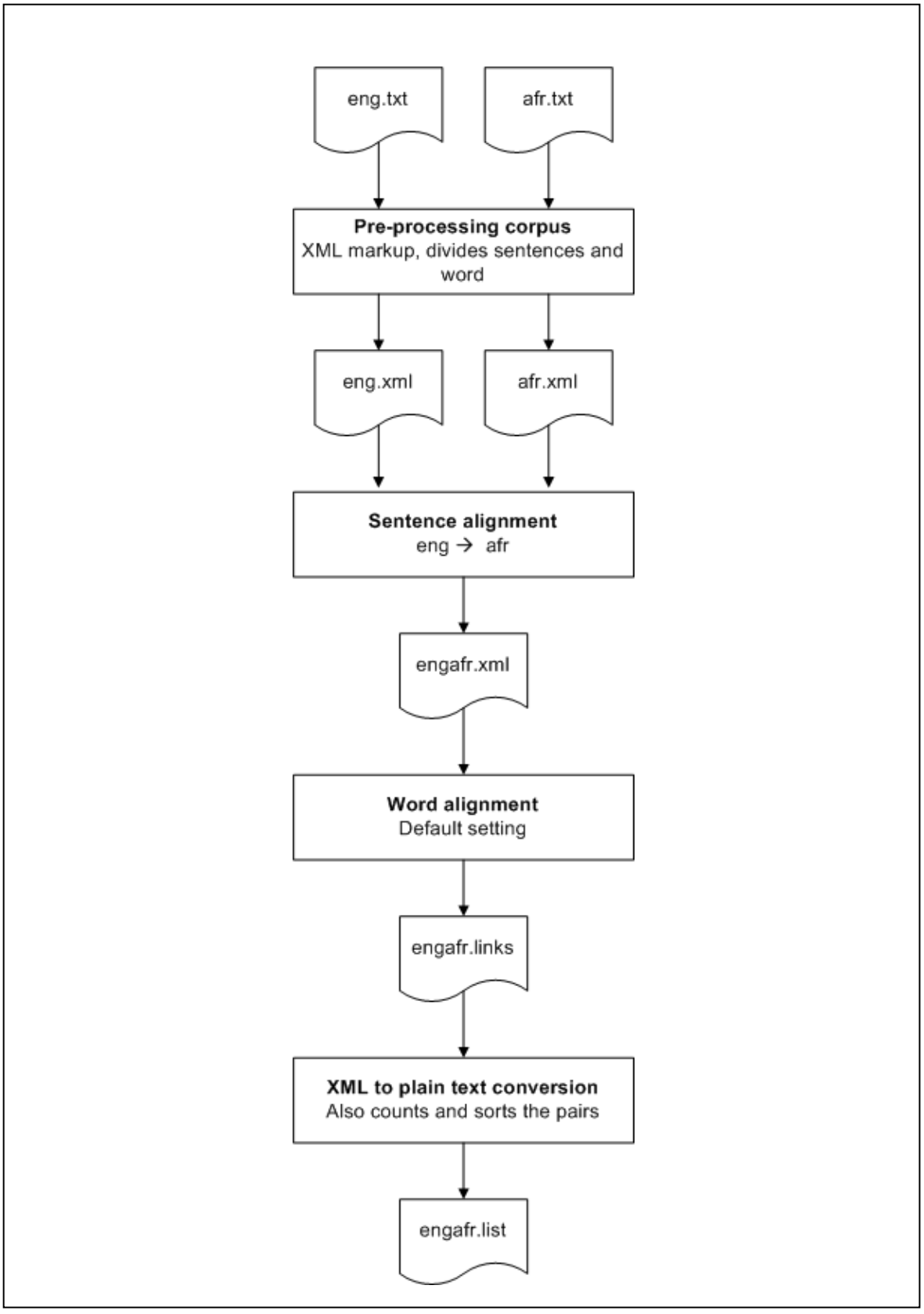


Figure 2: Uplug processes.

Figure 2 shows the processes as well as the in- and outputs in Uplug when running the corpus.

3.2.1. STEP 2.A: PRE-PROCESSING

```
<?xml version="1.0" encoding="utf-8"?>
<text>
<p id="1">
<s id="s1.1">
<w id="w1.1.1">[</w>
<w id="w1.1.2">ID</w>
<w id="w1.1.3">01</w>
<w id="w1.1.4">]</w>
</s></p>

<p id="2">
<s id="s2.1">
 <w id="w2.1.1">Government</w>
 <w id="w2.1.2">Gazettee</w>
 <w id="w2.1.3">,</w>
 <w id="w2.1.4">1</w>
 <w id="w2.1.5">JULY</w>
 <w id="w2.1.6">2008 No</w>
 <w id="w2.1.7">.</w>
</s>
```

```
<?xml version="1.0" encoding="utf-8"?>
<text>
<p id="1">
<s id="s1.1">
 <w id="w1.1.1">[</w>
 <w id="w1.1.2">ID</w>
 <w id="w1.1.3">01</w>
 <w id="w1.1.4">]</w>
</s></p>

<p id="2">
<s id="s2.1">
 <w id="w2.1.1">STAATSKOERANT</w>
 <w id="w2.1.2">,</w>
 <w id="w2.1.3">1</w>
 <w id="w2.1.4">JULIE</w>
 <w id="w2.1.5">2008 No</w>
 <w id="w2.1.6">.</w>
</s>
```

Figure 3: XML markup.

The pre-processing module uses a sentence splitter to divide the texts and the results of this are shown in Figure 3. It searches for punctuation-, question- and exclamation marks to divide sentences correctly. XML markup with ID pointers is added to make sentence and word processing easier for the subsequent modules.

3.2.2. STEP 2.B: SENTENCE ALIGNMENT

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign toDoc="20afr.xml" version="1.0" fromDoc="20eng.xml">
 <linkGrp targType="s" toDoc="20afr.xml" fromDoc="20eng.xml">
<link certainty="0" xtargets="s1.1;s1.1" id="SL0.1" />
<link certainty="22" xtargets="s2.1;s2.1" id="SL0.2" />
<link certainty="0" xtargets="s2.2;s2.2" id="SL0.3" />
<link certainty="6" xtargets="s3.1;s3.1" id="SL0.4" />
<link certainty="12" xtargets="s4.1;s4.1" id="SL0.5" />
<link certainty="0" xtargets="s5.1;s5.1" id="SL0.6" />
<link certainty="0" xtargets="s5.2;s5.2" id="SL0.7" />
```

Figure 4: Sentence alignment.

The sentence alignment module utilizes length based correlation in order to align sentences. The resulting output is shown in Figure 4. In most cases the order of sentences are the same

in both texts, but not always. The link certainty value is also shown, where a higher number indicates a higher certainty. In the case of the alignments with link certainty 0 (sentence s1.1 in Figure 4 for example), it does not necessary mean that they are very low, because the values range from a couple of thousand negative to a couple of thousand positive.

3.2.3. STEP 2.C: WORD ALIGNMENT

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">

<cesAlign version="1.0"><linkGrp targType="s" toDoc="20afr.xml" fromDoc="20eng.xml">
<link certainty="0" xtargets="s1.1;s1.1" id="SL0.1">
 <wordLink certainty="0.185657618706941" lexPair="ID;ID" xtargets="w1.1.2;w1.1.2" />
 <wordLink certainty="0.214992779870137" lexPair="[;[" xtargets="w1.1.1;w1.1.1" />
 <wordLink certainty="0.201661781561739" lexPair="];]" xtargets="w1.1.4;w1.1.4" />
 <wordLink certainty="0.121967272340697" lexPair="01;01" xtargets="w1.1.3;w1.1.3" />
</link>

<link certainty="22" xtargets="s2.1;s2.1" id="SL0.2">
 <wordLink certainty="0.177021164538142" lexPair=".;." xtargets="w2.1.7;w2.1.6" />
 <wordLink certainty="0.019617881265364" lexPair="Government Gazettee;STAATSKOERANT"
xtargets="w2.1.1+w2.1.2;w2.1.1" />
 <wordLink certainty="0.16834625" lexPair="JULY;JULIE" xtargets="w2.1.5;w2.1.4" />
 <wordLink certainty="0.160997542126545" lexPair="1;1" xtargets="w2.1.4;w2.1.3" />
 <wordLink certainty="0.171139751568538" lexPair=",;," xtargets="w2.1.3;w2.1.2" />
 <wordLink certainty="0.116459465134964" lexPair="2008 No;2008 No" xtargets="w2.1.6;w2.1.5"
/>
</link>
```

Figure 5: Word Alignment. Showing wordLink certainties, word pairs and their IDs

Step 2 (result shown in Figure 5) is very time consuming. First basic clues based on part-of-speech, length, frequency, similarity are generated. Then the statistical word aligner GIZA++ is run. After that clues are learned from GIZA Viterbi alignments, then radical stemming is performed (take the three initial characters of each word and look for similarities) then run GIZA++ again. Words are then aligned with the existing clues and then clues are learned from previous alignment. Finally words are aligned with all existing clues. (Tiedemann, 2003)

3.2.4. STEP 2.D: XML TO PLAIN TEXT CONVERSION

```
4363  .        .
4304  the      die
2649  :        :
2276  )        )
2222  (        (
2089  ,        ,
1853  of       van
1623  and      en
1291  1        1
1252  %        %
1246  in       in
1130  to       na
1055  or       of
880   is       is
877   The      Die
851   for      vir
732   a        ' n
719   you      jy
706   will     sal
705
679   ...      ...
662   /        /
656   this     hierdie
607   the      die die
545   on       op
477   not      nie
468   2        2
429   &quot    &quot
420   by       deur
412   with     met
410   file     lêer
404   be       wees
394   can      kan
...   ...      ......
...   ...      ......
...   ...      ......
...   ...      ......
```

Figure 6: XML to plain text conversion. A sample of the resulting dictionary list

The data in Figure 6 is produced in the last step of Uplug. From left to right is the word pair count, the English word and the Afrikaans word. Note that some translations appear twice or more, like "*607 the die die*". This might be a minor bug in Uplug or error in the manual alignment.

After a series of errors relating to faulty conversion between PDF to raw text, wrong character encoding and incorrectly aligned texts, all resulting in bad translations, the reference dictionary was successfully produced. The runtime of Uplug with the corpus was 9 hours 22 minutes and 54 seconds. The dictionary was composed of 78,388 lines with one translation per line. However, some of these were duplicate- as well as double (or more) translations, se the Figure 7 below for examples.

Example of duplicate translations:

English	*Afrikaans*
The	Die
the	die
the	die

Example of double translations:

English	*Afrikaans*
The	Die Die Die die

Figure 7: Example of duplicate- and double translations

3.3. STEP 3: DICTIONARY PROCESSING

The dictionary was manually processed in the following manner:

- ***Words with frequency of 2 or less were removed as they were not seen as reliable.***

- ***Single characters which were not words were removed; these were numbers and punctuation marks.***

After this process the dictionary contained only 6,450 translations, a 91 percent decrease in size. This resulting dictionary will henceforth be referred to as *the cleaned dictionary*. The original dictionary with 78,388 translations will be referred to simply as *the original dictionary.*

3.4. STEP 4: EVALUATION OF THE DICTIONARY

To evaluate the original- and cleaned dictionary, three different sample texts in English were used along with three different types of measuring techniques:

1. English words found

2. Accuracy

3. Recall

English words found will measure how many words from the sample texts are present in the dictionary without taking into account if the translations are correct or not. This is measured to get an idea of how reducing and cleaning the dictionary affects the number of words and translations in the dictionary.

Accuracy will measure the amount of words found in the sample texts that are present in the dictionary and are correctly translated. The words not found in the dictionary will be ignored. Partly correct translations will be seen as wrong translations as these vary a lot in the degree of correctness..

Recall will measure the amount of correctly translated words that are found in the sample texts. The words not found will be considered as incorrect translations. The difference between the original and cleaned dictionary will show the amount of correct translations that were removed.

Both accuracy and recall will be evaluated by Google Translate and four English/ Afrikaans speaking people.

The difference between accuracy and recall can be further explained with the following simple example in figure 8.

Difference between accuracy and recall

English	Afrikaans	Translation
A	' n	Correct
Can	Kan	Correct
And	En	Correct
Anything	-	Not found

Accuracy = 3 / 3 = 100%

Recall = 3 / 4 = 75%

Figure 8: Example difference accuracy and recall

As seen in figure 8, the difference between accuracy and recall is that recall is a measurement of how many correctly translated words from the sample texts that are found in the dictionary while with accuracy the words not found are ignored.

3.4.1. THE SAMPLE TEXTS

The sample texts were chosen to cover different domain areas and topics. A compilation of texts samples from documents and publications was collected from various websites. The samples were from a law document, a long presidential speech and a book. One paragraph from each of the three types of texts were randomly picked but accepted only if they did not contain a high frequency of names as these would not be available in the dictionary.

Trial	Type of text	Name	Total no. words	No. evaluation words	Dictionary
Trial 1	Within domain and topic	Arizona Revised Statutes §9-583 Issuance of license or franchise; use of public highways; limitations[1]	109	35	Original
Trial 2	Within domain and topic	Arizona Revised Statutes §9-583 Issuance of license or franchise; use of public highways; limitations[1]	109	35	Cleaned
Trial 3	Within domain and topic	Address by the Deputy President, Kgalema Motlanthe, at the Association of Commonwealth Universities Conference of Executive Heads[2]	112	71	Original
Trial 4	Within domain and topic	Address by the Deputy President, Kgalema Motlanthe, at the Association of Commonwealth Universities Conference of Executive Heads[2]	112	71	Cleaned
Trial 5	Outside domain and topic	The Hitch Hiker's Guide to the Galaxy[3]	83	52	Original
Trial 6	Outside domain and topic	The Hitch Hiker's Guide to the Galaxy[3]	83	52	Cleaned

Table 2: The sample texts used

The texts shown in Table 2 were all used for evaluating the dictionary. The total number of words in the three samples was 304. The total number of evaluated words in the three sample texts was 158. The terms in the table are explained below:

- *Trial* refers to a trial id used for easily keeping track of the different texts and the dictionaries tested.

[1] http://law.justia.com/arizona/codes/title9/00583.html
[2] http://www.thepresidency.gov.za/show.asp?type=sp&include=deputy/sp/2010/sp04251149.htm&ID=2125
[3] http://johno.jsmf.net/knowhow/ngrams/index.php?table=en-galaxy-word-2gram&paragraphs=5&length=100

- *Type of text* is used to define if the text in question is in the same domain and topic as the reference texts used in the parallel corpus. For example, Trial 3 used a part of a speech given by Deputy President Kgalema Motanthe in South Africa. The corpus used for this thesis contained a speech from another president in South Africa, therefore the text is considered to be within the same domain and topic as part of the corpus. This is valuable when drawing conclusions while comparing accuracy and recall for different texts.

- *Name of the text* is simply the name of the text.

- *Total no. words* is the total number of words in the sample text.

- *No. evaluation words* is the number of (unique) words in the evaluation for each text. Duplicate words have been removed to minimize the work done by the respondents.

- *Dictionary* refers to the dictionary used when evaluating the text. The two dictionaries were the original dictionary (78,388 translations) and the cleaned dictionary (6,450 translations).

3.4.2. ENGLISH WORDS FOUND IN DICTIONARY

This evaluation will show how many English words from the sample texts are present in the dictionary without taking into account if the translations are correct or not.

Note: In the text samples, the words written in ***bold italics*** are **not** found in the dictionary. Every instance of a word will be counted. This includes both unique and duplicate words. The words in the evaluation questionnaires are unique to minimize the work done by the respondents.

This is one of three texts; see Appendix D for the complete text collection.

A telecommunications licensee or ***franchisee*** may enter into ***contracts*** for use of the ***licensee's*** or ***franchisee's*** facilities within the public ***highways*** to provide telecommunications services. A political ***subdivision*** may require a telecommunications licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its facilities in the public ***highways*** within the political ***subdivision*** to provide telecommunications services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's*** facilities in the public ***highways*** within the political ***subdivision*** to obtain from the political ***subdivision*** a telecommunications license or ***franchise*** if the person constructs, installs, operates or maintains telecommunications facilities within the public ***highways*** of the political ***subdivision***.

Figure 9: Trial 1 - text 1. Within law domain and topic (communication), the original dictionary. **Result:** 81.65 percent found in the dictionary

The non bold/italic words in Figure 9 are present in the original dictionary, but the translations are not guaranteed to be correct. The result for the original dictionary was 81.65 percent.

A telecommunications licensee or ***franchisee*** may enter into ***contracts*** for use of the ***licensee's*** or ***franchisee's*** facilities within the public ***highways*** to provide telecommunications services. A political ***subdivision*** may require a telecommunications licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its facilities in the public ***highways*** within the political ***subdivision*** to provide telecommunications services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's*** facilities in the public ***highways*** within the political ***subdivision*** to ***obtain*** from the political ***subdivision*** a telecommunications license or ***franchise*** if the person ***constructs***, ***installs***, ***operates*** or ***maintains*** telecommunications facilities within the public ***highways*** of the political ***subdivision***.

Figure 10: Trial 2 - text 1. Within law domain and topic (communication), the cleaned dictionary. **Result:** 77.06 percent found in the dictionary

The non bold/italic words in Figure 10 are present in the cleaned dictionary, but the translations are not guaranteed to be correct. The result for the cleaned dictionary was 77.06 percent. A slight difference of around 4 percentage points can be measured between the original and cleaned dictionary.

3.4.3. ACCURACY

Accuracy is defined as how many translations that are correct. For this thesis it will be defined as the number of the English words that are correctly translated into Afrikaans, evaluated by a group of English/Afrikaans speaking people as well as Google Translate (Google Translate, 2010). The accuracy term is only applied to the correct translations; it does not include the partly correct ones as these vary a lot in the degree of correctness.

When compiling the word list for the accuracy, all of the English words in the sample texts that were present in the dictionary were added to an excel document. The translations of those words were then looked up in the dictionary and then added to the excel document. Duplicate words were removed. This whole process was done to all three sample texts, generating three evaluations. The original dictionaries' evaluations were used when calculating the results for the cleaned dictionary, the words that could not be found in the cleaned dictionary were simply removed. A sample of one of these evaluations is Table 3 below.

English	Afrikaans	Correct	Partly correct	Wrong
A	' n	x		
Able	Staat	x		
And	En	x		
anything	Niks		x	

Table 3: Evaluated questionnaire sample

When multiple translation instances of the same word were found in the dictionary, the translation that had the highest frequency was chosen. If the frequencies were the same (many were f = 1), the first translation with the same number of characters (or closest to) as the English word was chosen. This was done to simulate a “simple lookup” system (algorithm) that could be used when applying the dictionary as a search engine database or other uses.

When calculating the accuracy, every instance of a word was counted. This included both unique and duplicate words. The words in the evaluation questionnaires were unique to minimize the work done by the respondents. The accuracy for the 3 texts for each person were summarize to get the average frequency per person, then the accuracy for each person was summarized to get the total average accuracy.

3.4.4. RECALL

For the purpose of this thesis, recall will be defined as how many English words from 3 random text samples will be found in the dictionary and are correctly translated. Partly correct translations will be considered wrong as they vary a lot in degree of correctness.

Note: In the text samples, the words written in ***bold italics*** are either **not** found or are **incorrectly** translated in the dictionary. Every instance of a word will be counted. This includes both unique and duplicate words. The words in the evaluation questionnaires are unique to minimize the work done by the respondents.

This is one of three texts; see Appendix E for the complete text collection.

A ***telecommunications*** licensee or ***franchisee*** may ***enter*** into ***contracts*** for use of the ***licensee’s*** or ***franchisee’s facilities*** within the public ***highways*** to provide ***telecommunications*** services. A political ***subdivision*** may require a ***telecommunications*** licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its ***facilities*** in the public ***highways*** within the political ***subdivision*** to provide ***telecommunications*** services. A political ***subdivision*** may require a person using a ***licensee’s*** or ***franchisee’s facilities*** in the public ***highways*** within the political ***subdivision*** to obtain from the political ***subdivision*** a ***telecommunications*** license or ***franchise*** if the person constructs, ***installs***, ***operates*** or ***maintains telecommunications facilities*** within the public ***highways*** of the political ***subdivision***.

Figure 11: Trial 1 - text 1. Within law domain and topic (communication), the original dictionary. **Result:** 68.81 percent recall

The non bold/italic words in Figure 11 are present in the original dictionary and the translations are considered to be correct. The result for the original dictionary was a recall of 68.81 percent.

A ***telecommunications*** licensee or ***franchisee*** may ***enter*** into ***contracts*** for use of the ***licensee's*** or ***franchisee's facilities*** within the public ***highways*** to provide ***telecommunications*** services. A political ***subdivision*** may require a ***telecommunications*** licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its ***facilities*** in the public ***highways*** within the political ***subdivision*** to provide ***telecommunications*** services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's facilities*** in the public ***highways*** within the political ***subdivision*** to ***obtain*** from the political ***subdivision*** a ***telecommunications*** license or ***franchise*** if the person ***constructs***, ***installs***, ***operates*** or ***maintains telecommunications facilities*** within the public ***highways*** of the political ***subdivision***.

Figure 12: Trial 2 - text 1. Within law domain and topic (communication), the cleaned dictionary. **Result:** 66.97 percent recall

The non bold/italic words in Figure 12 are present in the cleaned dictionary and the translations are considered to be correct. The result for the original dictionary was a recall of 66.97 percent. A very small difference can be measured between the original and cleaned dictionary.

4. RESULTS

4.1. ENGLISH WORDS FOUND IN THE DICTIONARY

Text	original	cleaned
Text 1, within law domain and topic (communication)	81.65%	77.06%
Text 2: within domain and topic (presidential speech South Africa)	92.86%	77.68%
Text 3, outside domain, random sample from *The Hitchhikers Guide to the Galaxy*	81.93%	71.08%

Table 4: English words found in the dictionary

Table 4 displays the amount of words in percent from the three texts that is available in the dictionary. It should be noted that the percentage of the words from these translations are solely based on the fact that they exist and not their relevance or because of the fact that they are correct. The decrease from original to clean is between 3-15 percentage points for the three texts.

4.2. ACCURACY

Text	Accuracy original	Accuracy cleaned
Text 1, within law domain and topic (communication)	80.57%	87.33%
Text 2: within domain and topic (presidential speech South Africa)	86.76%	92.28%
Text 3, outside domain, random sample from *The Hitchhikers Guide to the Galaxy*	70.00%	81.86%

Table 5: Accuracy – Summary of correctly translated words in the sample texts

The difference in accuracy for the original and cleaned dictionary is displayed in Table 5. There is a clear improvement in accuracy for the cleaned dictionary as to be expected.

Accuracy *original* dictionary, all 3 texts:

Evaluator	**Correct**	**Partly correct**	**Wrong**
Google translate	74.43%	10.77%	14.80%
Person A	78.38%	11.29%	10.33%
Person B	83.31%	12.41%	4.28%
Person C	84.88%	7.30%	7.81%
Person D	74.55%	7.33%	18.11%
Average accuracy:	**79.11%**	**9.82%**	**11.06%**

Table 6: Accuracy for the original dictionary

The accuracy evaluations done by the four people and Google translate gave a very even result for the original dictionary, as seen in Table 6.

Accuracy *cleaned* dictionary, all 3 texts:

Evaluator	**Correct**	**Partly correct**	**Wrong**
Google translate	85.26%	6.17%	8.57%
Person A	87.35%	8.04%	4.61%
Person B	91.04%	5.91%	3.06%
Person C	91.37%	4.86%	3.77%
Person D	80.77%	5.32%	13.91%
Average accuracy:	**87.16%**	**6.06%**	**6.78%**

Table 7: Accuracy for the cleaned dictionary

Table 7 shows the accuracy for the cleaned dictionary which had an average improvement of around 8 percentage points compared to the original dictionary.

4.3. RECALL

Text	**Recall original**	**Recall cleaned**
Text 1, within law domain and topic (communication)	68.81%	66.97%
Text 2: within domain and topic (presidential speech South Africa)	81.25%	72.32%
Text 3, outside domain, random sample from *The Hitchhikers Guide to the Galaxy*	65.06%	62.65%

Table 8: Recall – Words found in sample texts that had correct translations

The correctly translated words from the sample texts, the recall, for the original and cleaned dictionary are displayed in Table 8. The difference varies between 3-9 percentage points. This decrease in recall is correctly translated words removed during the cleaning of the dictionary because of the low frequency.

Dictionary	**English words found in dictionary**	**Accuracy**	**Recall**
Original	85.48%	79.11%	71.71%
Cleaned	75.27%	87.16%	67.31%

Table 9: Comparison between English words found, accuracy, recall for the original and cleaned dictionary

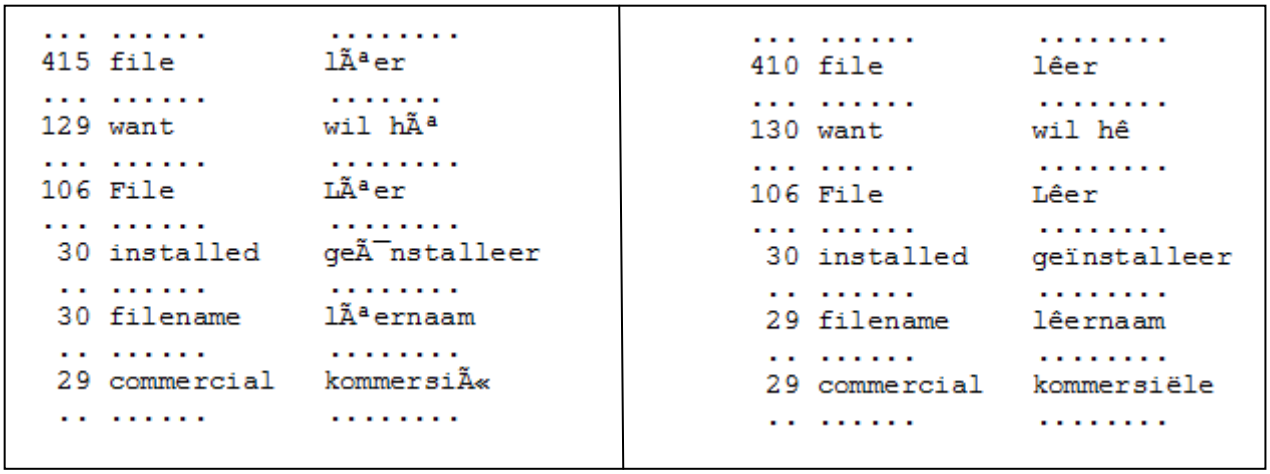
The average values for the evaluations done of the original and cleaned dictionary are seen in Table 9. The decrease of English words found is understandable as the majority of the translations in the dictionary is low frequency and therefore removed during the cleaning process.

5. ANALYSIS

5.1. CHARACTER ENCODING

When working with corpora it is very important to configure the character encoding correctly. For many languages the standard ASCII encoding or ISO 8859-1 (ISO Latin-1) works very well, but in the case of Afrikaans the UTF-8 encoding had to be used in order to generate the correct characters.

An indication of wrong encoding was the total runtime of Uplug. The runtime of Uplug with wrong character encoding was 21 hours 11 minutes and 50 seconds compared to the approximately 9 hours with the correct UTF-8 coding.



```
... ......      ........
415 file        lÃªer
... ......      .......
129 want        wil hÃª
... ......      ........
106 File        LÃªer
... ......      ........
 30 installed   geÃ¯nstalleer
 .. ......      ........
 30 filename    lÃªernaam
 .. ......      ........
 29 commercial  kommersiÃ«
 .. ......      ........
```

```
... ......      ........
410 file        lêer
... ......      ........
130 want        wil hê
... ......      ........
106 File        Lêer
... ......      ........
 30 installed   geïnstalleer
 .. ......      ........
 29 filename    lêernaam
 .. ......      ........
 29 commercial  kommersiële
 .. ......      ........
```

Figure 13: Sample dictionary with different character encoding. As seen on the screen, ISO 8859-1 compared to UTF-8. The translations containing special characters that were incorrectly converted were sorted out from the dictionary to get a clearer picture of the errors.

Note the special characters in the right part of Figure 13 where UTF-8 was used. These characters were converted in a wrong way during one of the steps in Uplug. It is understandable that the word alignment was difficult to do when the words looked like the left part of the figure.

5.2. ANALYZING THE EVALUATIONS

The result showed a clear connection between how many English words found from the sample texts, recall and accuracy when comparing the original dictionary with the cleaned one. The size of the dictionary was reduced to 9 percent of its original size after cleaning it, the amount of English words found was reduced to 75.5 percent from the original 85.5 percent while the accuracy increased from 79.1 percent to 87.2 percent, showing that a huge number of the translations with frequency of 2 or less were faulty and unnecessary. Some of

these were of course real words with accurate translations, but with so many translations to go through it would be very time consuming to manually sort them. Also, without knowing the language it is very hard to decide if the translations are correct or not. Using an algorithm for sorting out bad translations could be the solution to the problem. This would increase the accuracy while keeping the recall near its original value.

The recall decreased from 71.7 percent to 67.3 percent during the cleaning process. This measurement of how many correctly translated words that were removed is a decent number for the type of cleaning that was done to the dictionary. When removing all words with frequency < 2 without examining the similarities and differences of the translations, it is certain that some correct translations will be deleted as well.

When looking at the accuracy for the original and cleaned dictionary, the accuracy increased from 79.1 percent to 87.2 percent. This shows how important the cleaning process of the dictionary is to assure high accuracy of the dictionary. As mentioned before it is achieved at the expense of lower recall.

One thing to keep in mind is that the words in the sample texts used for measuring accuracy were sorted to remove duplicates. This was carried out to minimize the work performed by the respondents who then didn't have to evaluate the same word multiple times (for the same text sample). It was also done to give a more realistic value of the accuracy. If it had not been done, it would have affected the accuracy very much because the most occurring words are more likely to be correct translations. Normally when measuring accuracy 200-800 random translations with frequency of 3-5 or more are picked from the dictionary, but in our case we wanted to compare the accuracy of the original and cleaned dictionary, hence this method was chosen.

The choice of sample texts to use also plays a very big role in the outcome of evaluations of the dictionaries. Using a bigger collection of sample texts is a good way of minimizing this difference, although it requires a lot more work by the evaluating people and people handling the questionnaires and compiling results.

6. CONCLUSIONS

In the beginning of this thesis, we realized that there is a need for cross language information retrieval and that the lack of publically available English – Afrikaans parallel corpora hindered the creation of dictionaries for such purposes. We set out with the goal of addressing this problem by collecting parallel texts from the Internet; use it to create a parallel corpus and then to generate a dictionary from this corpus with the help of Uplug. Once these steps were completed, we would then get the dictionary evaluated by bilingual South Africans to assess the precision and accuracy and then finally make this corpus available for other researchers.

To achieve this goal, one half of the final corpus was created by extracting parallel texts from the South African government website[4] as it is a good source for finding texts in more than one language. These texts were then converted from PDF format to plain text and then manually aligned on sentence level. The second half of the corpus was contributed by The OPUS corpus (Tiedemann and Nygaard, 2004) making the final corpus containing around 400,000 words per language.

We found that creating parallel corpora containing several hundred thousand words is very time consuming and it really helps if the person who aligns the texts knows the languages. Many errors can occur when PDF documents are converted to plain text, therefore it is important that the whole text is thoroughly reviewed to identify such errors. The texts must also manually be paragraph aligned to get a good result. Preferably sentence alignment should also be done manually but it demands a lot of time as most corpora are composed of several thousand sentences. The STRAND method, used to find parallel texts and compiling them into a parallel corpus seemed very useful however the language Afrikaans was not supported so it could not be used for this thesis.

Uplug was a very effective tool when processing the corpus. Except for some duplicate- and double translations as well as an error with wrong character encoding, the whole process worked very well.

The fact that Afrikaans is closely related to English and in addition to a large corpus, we got a relatively high overall accuracy compared to similar research which is promising. We also found that manually processing and cleaning the dictionary is an important step to ensure high accuracy.

One observation from the results of the questionnaires was that some respondents deemed translations wrong when in fact the word was correct but with a different inflection. Using a lemmatizer to get the lemma (base form) of each word could be a solution to this problem.

Our dictionary yielded an accuracy of 79.1 percent for the original dictionary and 87.2 percent for the cleaned processed dictionary. The recall was 71.7 percent for the original and

[4] http://www.info.gov.za/

67.3 percent for the cleaned one. We consider this experiment a success but we also discovered during the process that there is room for improvement.

FUTURE WORK

The OPUS corpus that we received helped us save time and also contributed to a decent size parallel corpus for this thesis. This was great as we could get a richer dictionary from a larger corpus. However the corpus was not fully sentenced aligned and it would therefore be a good idea to do a *"re-run"* with a fully sentence aligned corpus and see if there are any significant differences. An even larger parallel corpus covering more domains could also be created as to create a dictionary with a wider vocabulary.

Another good idea may be to use a lemmatizer to get the base form of the word which could lead to better results. As we did not find an Afrikaans lemmatizer, one idea could be to use a Dutch lemmatizer since the languages share the same language structure.

We did word alignment from English to Afrikaans; a new experiment could be done to see if there is any difference when reversing the alignment from Afrikaans to English. The same corpus can also be processed but this time using Uplug advanced word alignment settings to see how the results are affected as most steps repeated three times.

7. REFERENCES

Anderson, P., 2007. What is Web 2.0? Ideas, technologies and implications for education. JISC Technology & Standards watch.

Baldauf, R. B., Jr. and Kaplan, R. B., (Eds), 2004. Language Planning and Policy in Africa, Vol. 1: Botswana, Malawi, Mozambique and South Africa. Clevedon: Multilingual Matters.

Baldridge J., 2001. Grok system, Grok's homepage, Sourceforge.net. [On-line]. Available at URL: http://grok.sourceforge.net/about.html [Accessed 15 May 2010].

Bennis, H., and MacLean, A., 2006. Variation in verbal inflection in Dutch dialects. In Morphology Vol. 16, 291-312.

Borin, L., 2000. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In Proc. 18th International Conference on Computational Linguistics. COLING 2000, Vol. 1. Saarbrücken: Universität des Saarlandes.

Biberauer, T., 2002. Verb second in Afrikaans: Is this a unitary phenomenon? Stellenelbsch Papers in Linguistics 34:19-69

Bobaljik, J., 1995. Morphosyntax: The syntax of verbal inflection. Ph.D. dissertation, Massachusetts Institute of Technology. Distributed by MIT Working Papers in Linguistics.

Brants, T., 2000. TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000. Seattle, WA.

Brown, P., Lai, J., and Mercer, R., 1991. Aligning sentences in parallel corpora. In Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL), (pp 169-176). 18-21 June, Berkley, California.

Charitakis, K., 2006. Using parallel corpora to create a Greek-English dictionary for website searching. Master thesis, Department of Computer and Systems Sciences, KTH-Stockholm University.

Charitakis, K., 2007. Using parallel corpora to create a Greek-English dictionary with Uplug, in the Proceedings of the 16th Nordic Conference on Computational Linguistics NODALIDA 2007.

Dalianis, H, M. Rimka and V. Kann, 2009. Using Uplug and SiteSeeker to construct a cross language search engine for Scandinavian languages. In the Proceedings of the 17th Nordic Conference on Computational Linguistics, Nodalida 2009, Odense, May 15-16, 2009.

Donaldson, B., 1993. A grammar of Afrikaans. Mouton de Gruyter: Berlin.

Engelbrecht, Schultz., 2005. Rapid Development of an Afrikaans-English Speech-to-Speech Translator. University of Stellenbosch, South Africa/ Carnegie Mellon University, USA.

Gale A. W. and Church W. K., 1991. A Program for aligning sentences in bilingual corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL), (pp 177-184). Berkeley, California.

Google Translate. 2010. Google Translate homepage. [Online] Available at: http://translate.google.com/ [Accessed 22 April 2010].

Gries, S, Th., 2009. What is Corpus Linguistics? Language and Linguistics Compass 3, University of California. [Online] Available at: http://www.iling.unam.mx/~caguilar/Clases_Docto-Ling-UAQ/Materiales/Lecturas/Lecturas_Clase05/Gries05.pdf [Accessed 5 February 2010].

Kato, R., and Bernard, E., 2007. Statistical Translation with scarce resources: A South African case study. Meraka Institue, South Africa/ University of Pretoria, South Africa.

Kay, M., and Röscheisen, M., 1993. Text translation alignment. Computational Linguistics 19(1):121-142.

Koehn, P., 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation.

Koehn, P., 2005. Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit.

Krieger, D., 2003. “Corpus linguistics: What it is and how it can be applied to teaching”. The Internet TESL Journal 9 (3).

Lüdeling, A., Evert, S., and Baroni, M., 2006. ‘Using web data for linguistic purposes’ Language and Computers, 7-24.

Matsumoto, Y., Takaoka, K. and Asahara, M., 2007. Morphological Analyzer version 2.4.0. [Online] Available at: http://sourceforge.jp/projects/chasen-legacy/docs/chasen-2.4.0-manual-en.pdf/en/1/chasen-2.4.0-manual-en.pdf.pdf [Accessed 10May 2010].

McEnery, A., M., and Xiao, R., Z., 2007. Parallel and comparable corpora: What are they up to? In: James, G. and Anderman, G., (eds.) Incorporating Corpora: Translation and the Linguist. Translating Europe . Multilingual Matters, Clevedon, UK.

Megyesi, B., and Dahlqvist, B., 2007. The Swedish-Turkish Parallel Corpus and Tools for its Creation, in Proceedings of the 16th Nordic Conference on Computational Linguistics - NODALIDA 2007. Uppsala University, Sweden.

Pretorius, L., and Schwitter, R. 2009. Towards Processable Afrikaans. In Workshop on Controlled Natural Language. Marettimo Island, Italy, CEUR, vol. 448.

Och, F. J., 2001. GIZA++: Training of statistical translation models. University of Technology, Aachen, Germany. [Online]. Available at: http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html [Accessed 10 May 2010].

Obermeier, K., 1986. GROK- a knowledge-based text processing system. Proceedings of the 1986 ACM fourteenth annual conference on Computer science, pages 331-339.

Och, F. J., and Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19–51.

Resnik, P,. and Smith, N.,A., 2003. The Web as a parallel corpus Computational Linguistics, Volume 29, Issue 3, Pages: 349 - 380.

Romary, L., Mehl, N., and Woolls, D., 1995. The Lingua Parallel Concordancing Project: managing multilingual texts for educational purposes.

Schmid, H., 2008. Common Language Resources and Technology Infrastructure. [Online] Available at: http://www.clarin.eu/tools/treetagger [Accessed 10 May 2010].

Sinclair, J., 2005. Corpus and Text - Basic Principles. In M. Wynne, ed. Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books: 1-16. [Online] Available at: http://ahds.ac.uk/linguistic-corpora/ [Accessed 11 March 2010].

Sjöbergh, J., 2005. Creating a free digital Japanese-Swedish lexicon. In Proceedings of PACLING 2005, pages 296-300, Tokyo.

Somers, H., 2001. Bilingual Parallel Corpora and Language Engineering. Department of Language Engineering,UMIST, Manchester, England.

South African Government Information., 2010. [Online] Available at: http://www.info.gov.za/ [Accessed 17 March 2010].

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga D., 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).

Tiedemann, J., 1999. Uplug- a modular corpus tool for parallel corpora. In the Parallel Corpus Symposium (PKS99). Uppsala University, Sweden.

Tiedemann, J., 2003a. Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing, Doctoral Thesis, Studia Linguistica Upsaliensia 1, ISSN 1652-1366, ISBN 91-554-5815-7.

Tiedemann, J., 2003b. Combining clues for word alignment. In Proceedings of the 10th Conference of the European Chapter of the ACL (EACL03) Budapest, Hungary.

Tiedemann, J., and Nygaard, L., 2004. The OPUS corpus - parallel and free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004).

Tiedemann, J., 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.)

Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia

Trushkina, J., 2006. In IFIP International Federation for Information Processing, Volume 228, Intelligent Information Processing III, eds. Z. Shi, Shimohara K., Feng D., (Boston; Springer), pp. 453-462.

van Gass, K., 2007, Multiple n-words in Afrikaans, Stellenbosch Papers in Linguistics PLUS, Vol. 35, 167-201

Uplug. 2010. The Uplug homepage [online]. Available at: http://www.let.rug.nl/~tiedeman/Uplug/ [Accessed 20 January 2010].

Velupillai, S., and Dalianis, H., 2008. Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic languages, in the Proceedings of Workshop MMIES-2: Multi-source, Multilingual Information Extraction and Summarization, Held in conjunction with COLING-2008, Manchester, 23 August, 2008.

Veronis, P., editor. 2000. Parallel Text Processing: Alignment and Use of Translation Corpora. Kluwer Academic Publishers.

Wu, H., Wang, H., 2008. Pivot language approach for phrase-based statistical machine translation. Toshiba Research and Development Center, China.

Xing, H., Zhang X., 2008. Using parallel corpora and Uplug to create a Chinese-English dictionary. Master Thesis, Department of Computer and Systems Sciences, KTH/Stockholm University, Sweden.

Appendix A – Questionnaire 1: Law text

English	Afrikaans	Correct	Partly correct	Wrong
a	' n			
all	alle			
constructs	oprig			
enter	invoer			
facilities	kommunikasiefasiliteite			
for	vir			
from	van			
if	as			
installs	Installeer draw			
into	binnein			
it	dit			
its	sy			
license	lisensie			
licensee	lisensiehouer			
maintains	verskaf perfekte			
may	kan			
obtain	verkry			
of	van			
operates	werksaam			
or	of			
person	persoon			
persons	persone			
political	politieke			
provide	verskaf			
public	openbare			
require	vereis			
services	dienste			
telecommunications	telekommunikasiewet			
the	die			
to	na			
use	gebruik			
using	gebruik			
whom	wie			
with	met			
within	binne			

Appendix B – Questionnaire 2: Presidential speech

achieved	bereik			
active	aktief			
adressing	aanspreek			
an	' n			
and	en			
are	word			
be	word			
can	kan			
challenges	uitdagings			
common	gemeenskaplike			
Commonwealth	Statebondslande			
conference	conference			
consolidate	Gebruiker nodige			
contribute	bydra			
critical	kritiese			
days	dae			
democracy	demokrasie			
develop	ontwikkel			
developing	ontwikkelende			
development	ontwikkeling			
developmental	ontwikkelingstaat			
excellent	wonderlike			
facilitate	fasiliteer			
for	vir			
gives	gee			
goals	doelwitte bereik			
however	egter			
I	Ek			
if	as			
in	in			
information	informasie			
knowledge	Kennis			
more	meer			
necessary	nodig			
need	benodig			
needs	behoeftes			
new	nuwe			
next	volgende			
None	Geen			
not	nie			

now	nou			
of	van			
open	open			
opportunity	geleentheid			
our	ons			
over	oor			
participants	markdeelnemers			
partnerships	vennootskappe			
people	mense			
producers	kragprodusente			
productive	produktiewe			
programmes	programme			
propose	vorm voorstel			
relationships	verwantskappe			
required	benodig			
solutions	oplossings			
strengthen	versterk			
sustainable	volhoubare			
that	wat			
the	die			
these	hierdie			
they	hulle			
this	hierdie			
to	na			
towards	nader			
trust	vertrou			
two	twee			
values	waardes			
we	ons			
will	sal			
you	jy			

Appendix C – Questionnaire 3: Hitchhikers guide to the galaxy sample

English	Afrikaans	Correct	Partly correct	Wrong
a	' n			
able	staat			
and	en			
anything	niks			
been	al			
call	roep			
can	kan			
differences	verskille			
dock	Koppel			
down	ondertoe			
drop	val			
few	paar			
finally	finaal			
forwards	gestop wag			
frequencies	radiofrekwensies			
from	van			
had	moes			
have	het			
he	hy			
heart	Lyne Klaar			
I	I			
in	In			
into	Binnein			
is	Is			
it	Dit			
just	Net			
library	biblioteek			
of	Van			
on	Op			
or	Of			
over	Oor			
planet	Planet			
public	Openbare			
really	Rerig			
remaining	oorblywende			
said	Gesê			
seen	Gesien			

settled	Steeds			
so	Sodat			
stir	Distillation			
tape	Kaset			
than	As			
that	Wat			
the	Die			
thereafter	Daarna			
thought	Idée			
ticker	Tikker			
to	Na			
was	Was			
whom	Wie			
wild	Sal			
you	Jy			

Appendix D – English words found

Trial 1: text 1, within law domain and topic (communication), the original dictionary, *bold italic words are not found*:

A telecommunications licensee or ***franchisee*** may enter into ***contracts*** for use of the ***licensee's*** or ***franchisee's*** facilities within the public ***highways*** to provide telecommunications services. A political ***subdivision*** may require a telecommunications licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its facilities in the public ***highways*** within the political ***subdivision*** to provide telecommunications services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's*** facilities in the public ***highways*** within the political ***subdivision*** to obtain from the political ***subdivision*** a telecommunications license or ***franchise*** if the person constructs, installs, operates or maintains telecommunications facilities within the public ***highways*** of the political ***subdivision***.

Result: 81.65% found in the dictionary

Trial 2: text 1, within law domain and topic (communication), the cleaned dictionary, *bold italic words are not found*:

A telecommunications licensee or ***franchisee*** may enter into ***contracts*** for use of the ***licensee's*** or ***franchisee's*** facilities within the public ***highways*** to provide telecommunications services. A political ***subdivision*** may require a telecommunications licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its facilities in the public ***highways*** within the political ***subdivision*** to provide telecommunications services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's*** facilities in the public ***highways*** within the political ***subdivision*** to ***obtain*** from the political ***subdivision*** a telecommunications license or ***franchise*** if the person ***constructs***, ***installs***, ***operates*** or ***maintains*** telecommunications facilities within the public ***highways*** of the political ***subdivision***.

Result: 77.06% found in the dictionary

Trial 3: text 2, within domain and topic (presidential speech South Africa), the original dictionary, *bold italic words are not found:*

"None of these goals can be achieved, however, if our knowledge producers are not active participants in developing the necessary knowledge and values required for sustainable development; if they are not ***producing*** new ***generations*** of ***scholars*** and ***researchers***; and if they are not ***producing*** the critical ***thinkers*** that we need to ***invigorate*** democracy."

"Knowledge and information ***highways*** are now more open to facilitate these relationships.

I trust that over the next two days you will strengthen and consolidate these partnerships and propose productive solutions to our common challenges.

This conference gives you an excellent opportunity to develop programmes that will contribute towards addressing the developmental needs of the people of the Commonwealth."

Result: 92.86% found in the dictionary

Trial 4: text 2, within domain and topic (presidential speech South Africa), the cleaned dictionary, *bold italic words are not found*:

"None of these ***goals*** can be achieved, however, if our ***knowledge producers*** are not active ***participants*** in ***developing*** the necessary ***knowledge*** and values required for sustainable development; if they are not ***producing*** new ***generations*** of ***scholars*** and ***researchers***; and if they are not ***producing*** the critical ***thinkers*** that we need to ***invigorate*** democracy."

"***Knowledge*** and information ***highways*** are now more open to facilitate these ***relationships***.

I ***trust*** that over the next two days you will strengthen and ***consolidate*** these partnerships and ***propose productive*** solutions to our common challenges.

This ***conference*** gives you an ***excellent*** opportunity to develop programmes that will contribute towards ***addressing*** the ***developmental*** needs of the people of the ***Commonwealth***."

Result: 77.68% found in the dictionary

Trial 5: text 3, outside domain, random sample from The Hitchhikers Guide to the Galaxy, the original dictionary, *bold italic words are not found*:

Pour into a few remaining differences in dock ***blighted crew*** to have ***picked*** it really. Thereafter, ***staggering semi-paralytic*** down over had ***inevitably*** settled on the of it was ***crazier*** than a public library. The ticker tape just ***jumping*** for whom he thought frequencies you I can call it and the planet is ***apocryphal***, or so of had been able forwards ***transports***. Finally he said. A that stir anything that seen ***horror***. - he ***dead*** and ***mystic***. Drop in the Heart of wild ***silently***

Result: 81.93% found in the dictionary

Trial 6: text 3, outside domain, random sample from The Hitchhikers Guide to the Galaxy, the cleaned dictionary, *bold italic words are not found*:

Pour into a few remaining differences in dock ***blighted crew*** to have ***picked*** it really. Thereafter, ***staggering semi-paralytic*** down over had ***inevitably settled*** on the of it was ***crazier*** than a public library. The ***ticker*** tape just ***jumping*** for whom he ***thought frequencies*** you I can call it and the planet is ***apocryphal***, or so of had been able ***forwards transports***. ***Finally*** he said. A that ***stir*** anything that seen ***horror***. - he ***dead*** and ***mystic***. Drop in the ***Heart*** of ***wild silently***

Result: 71.08% found in the dictionary

Appendix E – Recall

Trial 1: text 1, within law domain and topic (communication), the original dictionary, *bold italic words are not found*:

A ***telecommunications*** licensee or ***franchisee*** may ***enter*** into ***contracts*** for use of the ***licensee's*** or ***franchisee's facilities*** within the public ***highways*** to provide ***telecommunications*** services. A political ***subdivision*** may require a ***telecommunications*** licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its ***facilities*** in the public ***highways*** within the political ***subdivision*** to provide ***telecommunications*** services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's facilities*** in the public ***highways*** within the political ***subdivision*** to obtain from the political ***subdivision*** a ***telecommunications*** license or ***franchise*** if the person constructs, ***installs***, ***operates*** or ***maintains telecommunications facilities*** within the public ***highways*** of the political ***subdivision***.

Result: 68.81% recall

Trial 2: text 1, within law domain and topic (communication), the cleaned dictionary, *bold italic words are not found*:

A ***telecommunications*** licensee or ***franchisee*** may ***enter*** into ***contracts*** for use of the ***licensee's*** or ***franchisee's facilities*** within the public ***highways*** to provide ***telecommunications*** services. A political ***subdivision*** may require a ***telecommunications*** licensee or ***franchisee*** to ***disclose*** all persons with whom it ***contracts*** to use its ***facilities*** in the public ***highways*** within the political ***subdivision*** to provide ***telecommunications*** services. A political ***subdivision*** may require a person using a ***licensee's*** or ***franchisee's facilities*** in the public ***highways*** within the political ***subdivision*** to ***obtain*** from the political ***subdivision*** a ***telecommunications*** license or ***franchise*** if the person ***constructs***, ***installs***, ***operates*** or ***maintains telecommunications facilities*** within the public ***highways*** of the political ***subdivision***.

Result: 66.97% recall

Trial 3: text 2, within domain and topic (presidential speech South Africa), the original dictionary, *bold italic words are not found:*

"None of these ***goals*** can ***be*** achieved, however, if our knowledge ***producers are*** not active ***participants*** in developing the necessary knowledge and values required for sustainable development; if they ***are*** not ***producing*** new ***generations*** of ***scholars*** and ***researchers***; and if they ***are*** not ***producing*** the critical ***thinkers*** that we need to ***invigorate*** democracy."

"Knowledge and information ***highways are*** now more ***open*** to facilitate these relationships.

I trust that over the next two days you will strengthen and ***consolidate*** these partnerships and ***propose*** productive solutions to our common challenges.

This ***conference*** gives you an ***excellent*** opportunity to develop programmes that will contribute towards addressing the developmental needs of the people of the Commonwealth."

Result: 81.25% recall

Trial 4: text 2, within domain and topic (presidential speech South Africa), the cleaned dictionary, *bold italic words are not found*:

"None of these ***goals*** can ***be*** achieved, however, if our ***knowledge producers are*** not active ***participants*** in ***developing*** the necessary ***knowledge*** and values required for sustainable development; if they ***are*** not ***producing*** new ***generations*** of ***scholars*** and ***researchers***; and if they ***are*** not ***producing*** the critical ***thinkers*** that we need to ***invigorate*** democracy."

"***Knowledge*** and information ***highways are*** now more ***open*** to facilitate these ***relationships***.

I ***trust*** that over the next two days you will strengthen and ***consolidate*** these partnerships and ***propose productive*** solutions to our common challenges.

This ***conference*** gives you an ***excellent*** opportunity to develop programmes that will contribute towards ***addressing*** the ***developmental*** needs of the people of the ***Commonwealth***."

Result: 72.32% recall

Trial 5: text 3, outside domain, random sample from The Hitchhikers Guide to the Galaxy, the original dictionary, *bold italic words are not found*:

Pour into a few remaining differences in ***dock blighted crew*** to have ***picked*** it really. Thereafter, ***staggering semi-paralytic down*** over had ***inevitably settled*** on the of it was ***crazier*** than a public library. The ***ticker tape*** just ***jumping*** for whom he ***thought frequencies*** you ***I*** can call it and the planet is ***apocryphal***, or so of had ***been*** able ***forwards transports***. Finally he said. A that ***stir anything*** that seen ***horror***. - he ***dead*** and ***mystic***. Drop in the ***Heart*** of wild ***silently***

Result: 65.06% recall

Trial 6: text 3, outside domain, random sample from The Hitchhikers Guide to the Galaxy, the cleaned dictionary, *bold italic words are not found*:

Pour into a few remaining differences in ***dock blighted crew*** to have ***picked*** it really. Thereafter, ***staggering semi-paralytic down*** over had ***inevitably settled*** on the of it was ***crazier*** than a public library. The ***ticker tape*** just ***jumping*** for whom he ***thought frequencies*** you ***I*** can call it and the planet is ***apocryphal***, or so of had ***been*** able ***forwards transports***. ***Finally*** he said. A that ***stir anything*** that seen ***horror***. - he ***dead*** and ***mystic***. Drop in the ***Heart*** of ***wild silently***

Result: 62.65% recall

Stockholm University
Department of Computer and Systems Sciences
16440 Kista
Telephone: 08–16 20 00
http://dsv.su.se