An overview of the EtsaTrans machine translation system: compilation of an administrative domain

L. Ehlers & G. v.d. M. Hanekom
University of the Free State
Department of Afroasiatic Studies
Sign Language and Language Practice
BLOEMFONTEIN
Email: ehlersl.hum@ufs.ac.za
hanekomg.hum@ufs.ac.za

Abstract

An overview of the EtsaTrans machine translation system: compilation of an administrative domain

The EtsaTrans machine translation system has been in development at the University of the Free State for the last four years and is currently the only machine translation system being developed in South Africa for specialised and non-general translation needs. The purpose of this exposition is to present the program through its phases of development, and to report on current levels of performance. We analyse the output, the size of the database, and then propose the future implementation of a part of speech tagger and word stemmer into the program to improve its linguistic performance. Our goal with the system is not to translate all types of document, but to work in a specialised domain that will allow the system to translate documents that are repetitive in nature. This will enable translators to spend more time on non-repetitive subject matter. By capturing the nature of the language of such repetitive documents in the database, we are able to create a standardised language usage for the specialised domain.

Opsomming

'n Oorsig van die EtsaTrans-masjienvertalingstelsel: die samestelling van 'n administratiewe domein

Die EtsaTrans-masjienvertalingstelsel word die afgelope vier jaar reeds aan die Universiteit van die Vrystaat ontwikkel. Dit is tans die enigste masjienvertalingstelsel in Suid-Afrika wat vir gespesialiseerde (nie-algemene) vertalingsdoeleindes ontwikkel word. In hierdie uiteensetting word die program na gelang van sy ontwikkelingsfases beskryf en word daar oor die huidige verrigtingsvlakke verslag gegee. Ons kyk na die uitsette, databasisgrootte en die toekomstige inkorporering van 'n woordsoortetiketteerder en woordstamherkenner om die program se linguistiese werkverrigting te verbeter. Ons doel is nie om alle tipes tekste te kan vertaal nie, maar wel om in 'n gespesialiseerde domein te werk wat die stelsel in staat sal stel om dokumente van 'n repeterende aard te vertaal. Dit sal vertalers vrystel om tyd aan minder repeterende tekste te wy. Deur die aard van die taalgebruik in sulke repeterende dokumente in die databasis vas te vang, is ons in staat om 'n gestandaardiseerde taalgebruik vir die gespesialiseerde domein te skep.

1. Historical background

The University of the Free State took over the rights of the LEXICA system from the company EPI-USE Systems in 2000. LEXICA was a transfer system that was used to do morphological, syntactic, semantic and contextual analyses and could be used for the following language pairs: Afrikaans, Setswana, Swahili and Portuguese to English; and English to isiXhosa, isiZulu and Afrikaans. The development of EPI-USE's LEXICA system began in 1990 and continued until the beginning of 2003. An evaluation done on the system showed that continuing with the development of a purely rule-based machine translation (RBMT) system would be futile in terms of the latest developments within machine translation (see Snyman & Naudé, 2003). Sumita and Iida (1999) state that conventional machine translation systems use rules as knowledge, and that it is difficult to build a practical system because of the problem of building such a large-scale rule-base. They also mention the difficulties involved in improving translation performance because the effect of adding a new rule is hard to anticipate, and because translating using a large-scale rule-based system is time-consuming. The dictionaries that were developed were too broad to be of any use in a domain-specific field. Tests showed that although LEXICA could translate a document well enough to convey meaning, the results were not syntactically satisfactory (Snyman & Naudé, 2003). Since

the addition of new rules and data to the system did not cause any significant improvement, new avenues of development needed to be explored within the latest developments in machine translation (MT).

The discontinuation of the RBMT system excluded pure RBMT from being considered as a possible MT paradigm. Therefore, it was decided to incorporate example-based machine translation (EBMT) and statistical-based machine translation (SBMT) with RBMT. Although other paradigms are available they will not be discussed in this article.

In a survey conducted by Dorr *et al*. in 1998, the research techniques are discussed in terms of three categories

- Those that propose to rely most heavily on linguistic knowledge
- Those that use more mathematical knowledge
- Those that use a combination of the two

The first category includes RBMT as one of the techniques that rely heavily on linguistic knowledge. “The RBMT [technique] is associated with systems that rely on different linguistic levels of rules for translation between the source and target language.” (Dorr *et al*., 1998:25.)

The second category contains SBMT which is derived from speech-processing techniques. SBMT is an application of Bayes’ rule and is used to show that the probability that a string of words is a translation of a source string is proportional to the product of the probability that a source string is a translation of target string (Dorr *et al*., 1998). The Bayes rule is the mathematical basis of the “noisy channel model”, which has been applied to problems like speech recognition, optical character recognition, MT and spelling correction (Fry, 2007:11).

The second category (mathematically-based techniques) also contains EBMT. Dorr *et al*. (1998:33) states that “the basic idea of EBMT assumes a database of parallel translations which is searched for the source language sentences and phrases closest matching a new source language sentence”. The accuracy and quality of the translation depends heavily on the size and coverage of the parallel database. The development of an EBMT system was determined to be most suitable for the first phase of development because parallel data was available.

The third category, is a hybrid between the first two techniques. Dorr *et al*. (1998:35) state that

> SBMT does not handle long range contextual dependencies and EBMT has difficulties with complex sentence structure. It was quickly recognised that these mathematical knowledge could be combined with linguistic knowledge to exploit the strengths of each.

Having done some EBMT development, the team soon realised that linguistic knowledge and its applications are essential. This led to the development of a hybrid system which incorporates both techniques – EBMT and RBMT.

2. EtsaTrans developmental aspects

In 2003 the development of EtsaTrans commenced. The name EtsaTrans is a combination of *Etsa* which is *do* in Sesotho and *Trans* which is an abbreviation of *translation*, and which literally means *Do translation*. It was decided to retain the LEXICA system's dictionaries which had been incorporated into the EBMT database, as much hard work had been put into their development and the body of collected information is useful and has proven its value although syntactically the LEXICA system did not rise to our expectations, the dictionaries were still usable. In addition, the development of new dictionaries from scratch would be extremely time-consuming. It was, however, deemed necessary to work through the existing LEXICA dictionaries in order to correct errors and tag the entries with the new parts of speech information. See below an example of the original LEXICA format for the language pair Afrikaans-English:

> ADD(source=aangevlieg,target=flown,type=verb,form=pastpart, irreg=yes,deel=aan,context=GE)
>
> ADD(source=liggame,target=bodies,type=noun,num=p,irreg= yes,context=GE)
>
> ADD(source=redelik,target=equitable,type=adj,context=GE)

A brief description of some of the meta information used in the LEXICA system:

- Source – refers to the source word token;
- target – refers to the target word token;

- type – shows the part of speech;
- form – indicates the tense, e.g. past tense, present tense, future tense; and
- irreg – is the word token irregular, yes or no.

In order to expand and test the dictionaries, a random selection of newspaper articles was used as test sets. First, the articles were translated by LEXICA. Next, the results were evaluated and missing entries were added to the database, as were multi-word expressions. Although the original LEXICA dictionaries were mono-directional (if a term was added to the Afrikaans-English (A-E) language pair in the dictionary it would not be available in the English-Afrikaans (E-A) dictionary), EtsaTrans is bidirectional. The present EtsaTrans database consists of multi-word units of two or more word tokens (henceforth multi-word units) as well as containing single word tokens (henceforth single-word units). There is currently information available for thirteen language pairs.

The EtsaTrans team works in the following way to develop a language pair: once a new language pair has been identified the first phase is the development of the mathematical knowledge (EBMT system). The second phase incorporates linguistic knowledge. These phases are applied systematically to each language pair.

2.1 Phase one (mathematical knowledge)

In this phase the following steps are followed:

- Determine in which domains to work;
- assess the texts already available;
- translate texts into target languages (where necessary);
- develop parallel multi-word and single-word units for the database;
- manually tag the wordlist for parts of speech with the appropriate word type(s); and
- check the existing LEXICA system dictionary for the language pair (if there is one).

Once these steps have been completed, EtsaTrans can begin translating documents. The translated text, however, is certainly not 100% correct at this stage, and requires some editing by the user.

Since word order differences and problems relating to homonyms (examples to follow later) in the source language are the largest hurdles on the road to accurate translation, linguistic knowledge is required in the form of taggers and stemmers for each language.

2.2 Phase two (application of linguistic knowledge)

This phase entails the development of a part of speech tagger (POS tagger). POS tagging, also called grammatical tagging, is a process in which syntactic categories are assigned to words. The two factors determining the syntactic category of a word are its lexical probability (for example, out of context, *man* is more probably a noun than a verb), and its contextual probability (for example, after a pronoun, *man* is more probably a verb than a noun, as in "they man the boats" (Daelemans & Zavrel, 1996:14)). The development of a POS tagger for EtsaTrans's A-E and E-A language pairs has been provisionally completed and will need to be expanded once testing starts.

The second part of phase two is the development and installation of a stemmer. A stemmer is a program/part of a program that identifies or extracts core roots from a word, removing prefixes and suffixes. For example, the words *run, runs, ran* and *running* all have *run* as the root. A stemmer is often used in matching processes to make it possible to recognise that documents are about the same topic even when they use variants of the same words.

The stemmer helps with information retrieval and works in conjunction with the POS tagger: if a word is unknown, but its stem is available in the database, then the POS tagger can determine its type and enable manipulation of the word. The purpose of including a stemmer and a POS tagger is to improve both the word selection and word-order capabilities of the system. Each language requires its own stemmer and POS tagger. No attempt has yet been made to design or develop a stemmer since all possible avenues are first being researched before a final decision on implementation is made.

3. EtsaTrans at work: translating in an administrative domain

In the previous section we saw that phase one (development of the A-E/E-A language pair, based on examples) in the development of the EtsaTrans system has already been undertaken. What follows is a description of how the EtsaTrans database creation phase was conducted and how the program functions at present, i.e. to what

degree of success it translates without incorporation of the stemmer and tagger.

3.1 Building a database

The EtsaTrans team decided to focus on an administrative domain, namely minutes of official meetings, from which texts were chosen to extract information. The team chose these texts for the following reasons:

- They contain a fair amount of repetition, which makes it easier for an EBMT system to attain a reasonably high standard in outputs.
- The language usage in such texts is mainly standard, with little fluctuation regarding formality, thus improving the quality of an EBMT translation.
- The translation of such texts is extremely time-consuming for translators and by allowing EtsaTrans to do most of the work (the translations merely require editing) these translators have more time available for more specialised texts.

The team obtained a body of official minutes already translated either from English to Afrikaans or vice versa, aligned the source and target texts and proceeded to build the bilingual database by selecting and pairing multi-word or single-word units and saving them in a database for future translations. These official minutes thus constituted the training data for the program's administrative domain. A percentage of these texts were kept on one side to serve as test set during the test phase discussed below, under 4.2.

This process is not without problems. A major problem associated with using these minutes to build a database to translate other minutes is that the standard tenses in which English and Afrikaans minutes are written differ. In English the past tense is usually used, while Afrikaans is generally in the present. For instance:

> Afrikaans – "sal aangepas word*"* ("*will* be adjusted")
>
> English – "*would be adjusted"*

This makes it difficult to correctly pair and edit equivalent phrases. After many discussions it was decided that, for this administrative domain, the specialised database would be built using past perfect tense on the English side and present tense on the Afrikaans side. Had the database been developed with corresponding tenses, the

results would have been incorrect according to the norms of human translations in this context.

The largest general problem, even in texts that contain a fair amount of repetition, is that multi-word units are rarely exactly repeated from one text to another. This makes it virtually impossible for the software to find exact target translations to match the source text. One way to try to solve this problem is to continue expanding the database as much as possible, thereby improving the software's chances of finding exact equivalents in the database to match translations in the source text.

Keeping this in mind, let us now look at the level of success of EtsaTrans's translations.

3.2 Testing EtsaTrans

In the first phase of testing exclusive use was made of human evaluation. In the future machine evaluation will also be included as part of the test phase. The program was formally tested (Test 1) by translating a total of three different texts of varying lengths from the year 2004. These make up approximately 5% of the total training data from that year. The gathered information is divided into year groups from 2003 to 2006. 2003 has no test set because all the data was added to the database as training data. With such a small database it was considered of no value to conduct a test phase for that year.

At the beginning of each new year the previous year's data is added: for instance, at the beginning of 2008 the data for 2007 will be added. The test data consists of a certain percentage of randomly selected texts from the training data before it is added to the database. The percentage selected for training data decreases with each passing year:

2003 – 100% training data

2004 – 90% training data

2005 – 70% training data

2006 – 50% training data

The percentage by which it is reduced is determined by the linguists on the project who analyse a few texts for each year to determine how much new data can be extracted and added to the database.

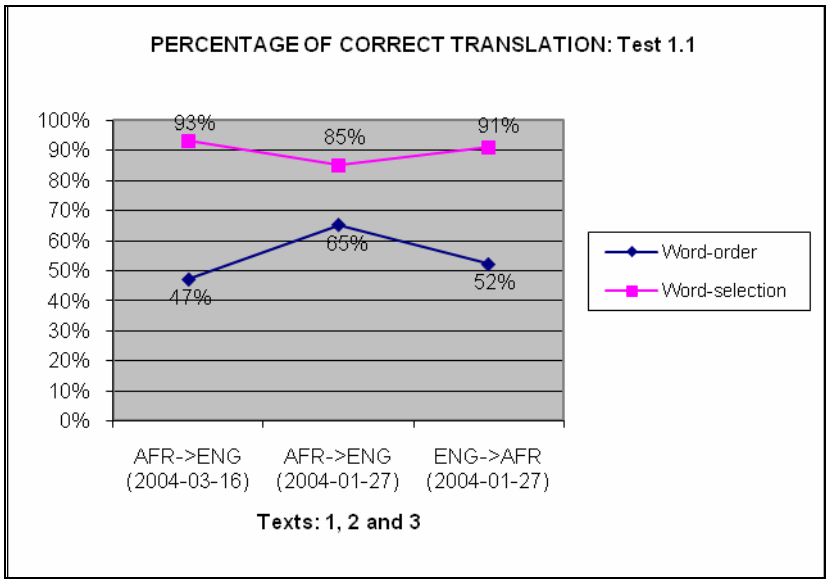
3.2.1 Test 1

For the first test the three texts (the first two translated from A-E, the third from E-A) were run through the program and the resulting target texts were analysed with regard to accuracy in word order and word selection.

The former was intended to measure accuracy on the level of sentence structure, while the latter was meant to measure accuracy on word level. Word order accuracy was determined by counting the number of sentences in the target text that contained word order errors and then processing this number as a percentage of the total number of sentences in the target text. Word selection accuracy was determined by counting the number of words incorrectly translated or omitted in the target text and processing this number as a percentage of the total number of words in the target text.

The results of the first test (Test 1.1) were as follows:

Fig. 1: Results of Test 1.1



(**Note**: Word order refers to the order of words within sentences as a whole, whether single words or a string of words, while word selection refers to the specific words/strings with which the system chooses to translate specific words/strings from the source text.)

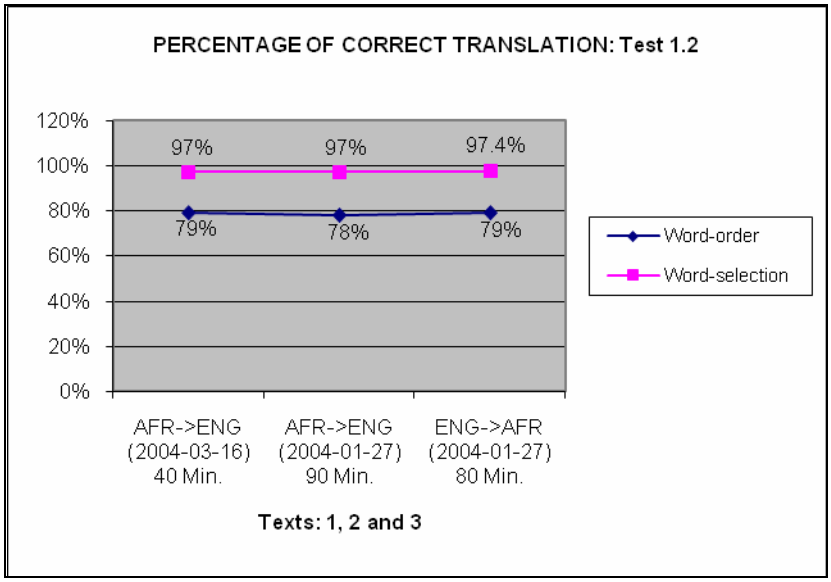
The following table is an analysis of the three texts used as the test set in Test 1. The statistics were compiled using WordSmith Tools.

Table 1: Analysis of the test sets used in Test 1

DESCRIPTION	TEXT 1	TEXT 2	TEXT 3
Total number of words in text	644,00	2 811,00	2 917,00
Total number of sentences	25,00	148,00	135,00
Mean (words per sentence)	25,76	18,99	21,61
Mean (word length in characters)	3,12	5,20	5,04
Total unique words in text	308,00	878,00	765,00
Total unique words that are not part of the vocabulary (counted manually)	1,00	5,00	2,00

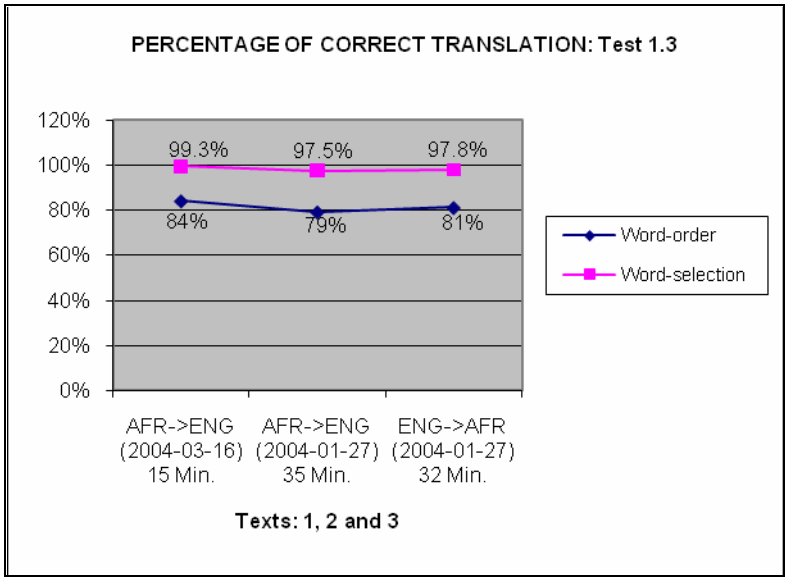
After the initial run, the resulting target texts were then automatically aligned with their corresponding source texts by EtsaTrans, and the database was expanded by pairing each faulty multi-word or single-word unit in the target text with its equivalent in the source text, correcting or replacing the data and then adding it to the database. The time taken to edit each of the texts in this way was written down as another statistic which could chart the program's possible future improvement. After editing, the same source texts used during the initial run were again run through EtsaTrans (Test 1.2) with the following results (editing time indicated):

Fig. 2: Results of Test 1.2



Following Test 1.2 the same test set was run for a third time (Test 1.3), following exactly the same process as in 1.2. These were the results:

Fig. 3: Results of Test 1.3



The statistics show a clear decrease in both word order and word selection errors from Tests 1.1 to 1.3, as well as a decrease in editing time from Tests 1.2 to 1.3. Even so, the results were still nowhere near human translation quality, especially regarding word order. Word order errors still occurred in many of the sentences that

were edited before rerunning the texts. A typical example of such a word order error is the translation of the sentence "Kennis word geneem dat 2 000 studente by die universiteit geregistreer is" as "Cognisance was taken that 2 000 students with the University were registered". The target sentence's word order is obviously faulty and should rather read "Cognisance was taken that 2 000 students were registered at the University". The reason for this is probably that the editor of the target text cannot fully anticipate the exact multi-word data that the software chooses to compare against the database. Therefore, while still expanding the database, he/she cannot save exact matches in the database for those multi-word units that the software chooses to translate.

The improvement of word selection was almost satisfactory at this stage, but the main problem was related to homonyms in the source language. In cases where a source word has a number of homonyms with different respective translations in the database, the software has to choose one word with which to translate the source word. Context makes this extremely problematic. Consider, for instance, the word *Kaap* in the following sentence: "Professor X het geen terugvoer oor die Noord-Kaap inisiatief nie". An acceptable English translation of this sentence would be: "Professor X had no feedback on the Northern *Cape* initiative". As the software cannot recognise the context of the given sentence it may translate the source sentence as follows: "Professor X had no feedback on the Northern *hijack* initiative". The Afrikaans word *Kaap* could be translated in English as either *Cape* or *hijack*. In order to improve on this selection, the EtsaTrans team decided to install a tool with which to prioritise word selection before starting Test 2. This "prioritiser" could instruct the software to favour one translation of a certain homonym above all others in future. In the light of the above example, the prioritiser would enable the software to prefer *Cape* as translation to *Kaap*. The results of Test 2 would not be significantly different from Test 1 only as a result of information added to the database, because the texts are all in the same domain and contain the same general terminology. Any further significant improvement would therefore be the result of including the prioritiser.

Note that although only two test sets are discussed in this article more than two tests were conducted with these and other texts. Test 2 was conducted using a new test set, because using the same test set would yield misleading results. The reason for this is that the database update would render the first text perfectly the second time round, because the missing information has now been added to the

database after the text 1 editing. The purpose of the Test 2 was to determine the program's performance with a new text.

3.2.2 Test 2

Having installed the prioritiser the team decided to test EtsaTrans with a different set of texts, which are numbered Texts 4, 5 and 6 (all translated from A-E), not only to test the effectiveness of the prioritiser, but also to ascertain whether the program, along with its database, had improved in general. In other words, the team wanted to see if the program now performed better with new texts, compared to its performance with the texts in Test 1.1, because what happened during the test phase described thus far was that, essentially, the test set had also become training data, which meant that EtsaTrans's performance could no longer be accurately assessed using these particular texts. Exactly the same process was followed as with Test 1. These are the results:

Fig. 4: Results of Test 2.1

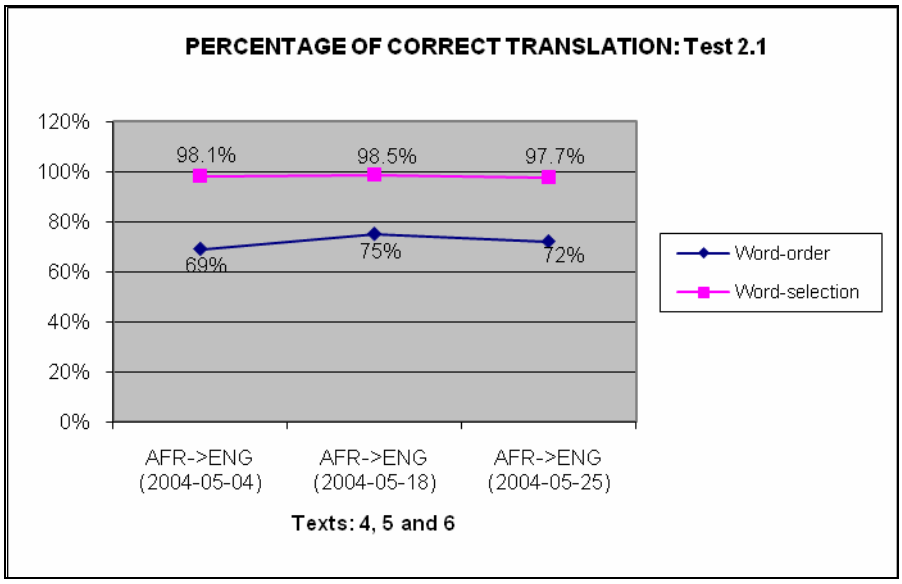
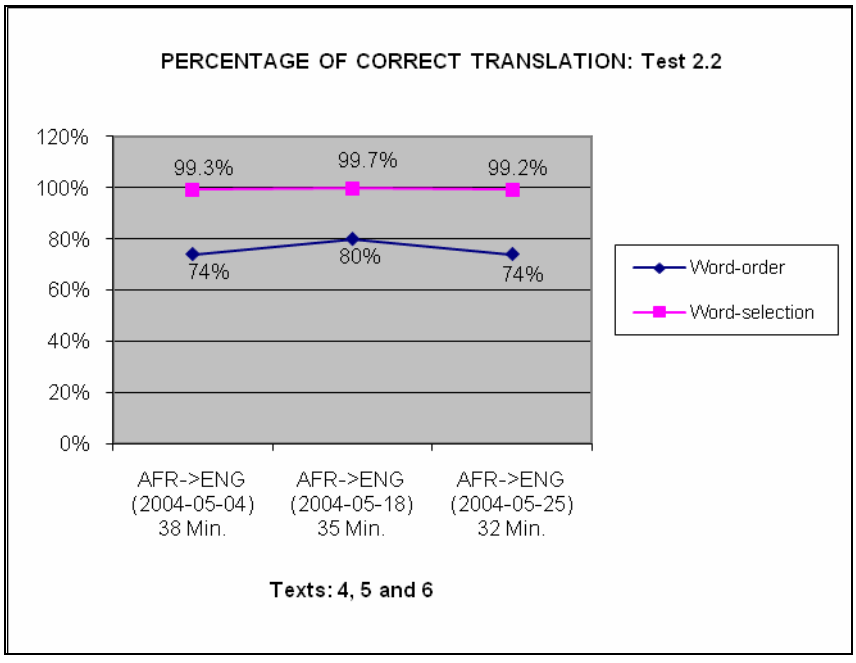


Table 2: Analysis of the test set used in Test 2

DESCRIPTION	TEXT 4	TEXT 5	TEXT 6
Total number of words in text	1 455,00	1 374,00	949,00
Total number of sentences	53,00	57,00	36,00
Mean (words per sentence)	27,45	24,11	26,36
Mean (word length in characters)	5,08	5,06	5,31
Total unique words in text	472,00	525,00	370,00
Total unique words that are not part of the vocabulary (counted manually)	12,00	6,00	13,00

Fig. 5: Results of Test 2.2 (after editing)



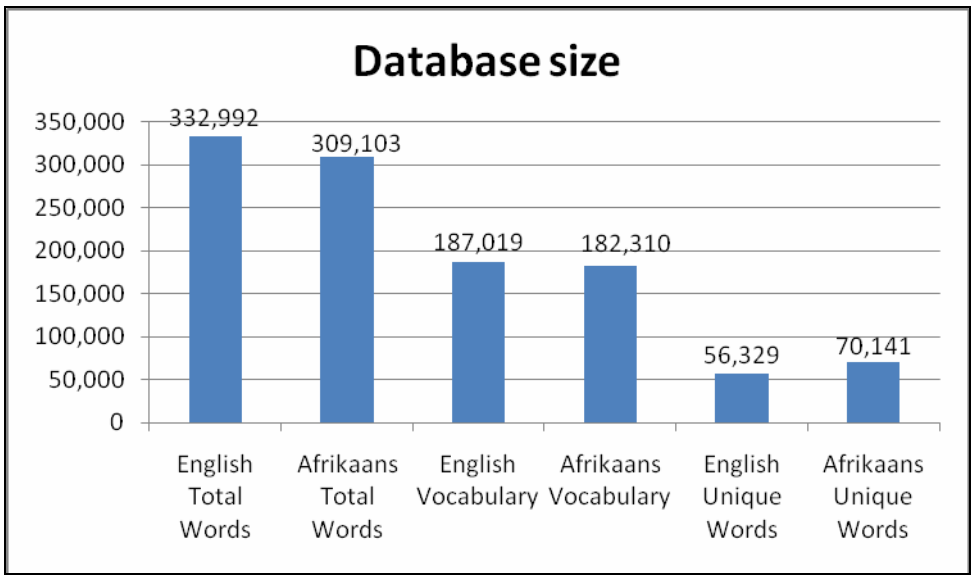
It was decided not to do a third run, as the results regarding word selection already showed much improvement compared to the first test, thanks to the prioritising tool. What these statistics show is that the program improved considerably between Tests 1.1 and 2.1 regarding both word order and word selection. Editing the resulting

texts of Test 2.1 also took considerably less time than editing the resulting texts of Test 1.1. The percentages of word selection errors visible in Test 2.2 were generally lower than those visible in Tests 1.2 and 1.3, presumably on account of the prioritising tool. For example, during Test 1.1 the word *raamwerk* in the sentence "Die *raamwerk* word goedgekeur" was translated as "fuselage": "The *fuselage* was approved". In the context of the source text, however, *raamwerk* refers to "framework." After setting the priority for translating *raamwerk* as "framework" rather than "fuselage", the former was subsequently chosen by the software to translate *raamwerk* in any future translated text. The result is that in Test 2.2 the sentence "Die *raamwerk* word deur die raad bespreek" was translated as "The *framework* was discussed by the council" and not as "The *fuselage* was discussed by the council".

Regarding word order errors, consider again the example cited under heading 4.2.1, namely "Kennis word geneem dat 2 000 studente by die universiteit geregistreer is". During the last test using the specific source text, this sentence was translated as "Cognisance was taken that 2 000 students were registered with the University". It represents the improvement mentioned above, although there were still sentences that contained word order errors, for example "Met waardering kennis geneem van ..." was translated incorrectly as "By means of appreciation cognisance was taken of ...", which could rather be translated as "Cognisance was, with appreciation, taken of ..."

4. Database size

Fig. 6: Database size for the language pair English-Afrikaans



For the purpose of this article we visually represent and focus on the size of the E-A language pair. The full spectrum of indigenous language pairs are Afrikaans, IsiXhosa, IsiZulu, Sepedi, Sesotho, Setswana and Swahili to English. All these languages translate to and from English only, and are not interchangeable. In Figure 6 we have the following data:

- English Total Words – total English word tokens in database
- Afrikaans Total Words – total Afrikaans word tokens in database
- English Vocabulary – each word token that appears in the database one or more times (each counted only once)
- Afrikaans Vocabulary – each word token that appears in the database one or more times (each counted only once)
- English Unique Words – word tokens that appear in the database only once, and
- Afrikaans Unique Words – word tokens that appear in the database only once

Although the database size is not large (see Figure 6), the results of the tests show that a small domain-specific database can yield good results when translating domain-specific texts.

5. Conclusions and summary

From what we have seen during the test phase it is clear that the EtsaTrans machine translation system, after the first phase of development, shows potential in producing satisfactory target texts in a specialised domain. Its improvement throughout this stage is largely ascribed to the expansion of its multi-word and single-word database by editing target texts. Regarding word selection improvements, the prioritising tool has proved a valuable instrument in improving the quality of translations. The further installation and development of the tagger and stemmer will hopefully improve the quality of translations. Another conclusion reached is that, at this stage, the system is viable in a controlled environment rather than for general language translation, and could be utilised to lighten the burden of translators by removing repetitive and tedious translations from their everyday work.

List of references

DAELEMANS, W. & ZAVREL, J. 1996. MBT: a memory-based part of speech tagger-generator. http://66.102.1.104/scholar?hl=en&lr=&q=cache:qfdu-FNDAIMkJ:www.folli.uva.nl/CD/1998/pdf/vandenbosch/vandenbosch_2.pdf Date of access: 25 Jun. 2007.

DORR, B., JORDAN, P.W. & BENOIT, J.W. 1998. A survey of current paradigms in machine translation. http://stinet.dtic.mil/oai/oai?&verb-=getRecord& metadataPrefix=html&identifier=ADA455393 Date of access: 25 Jun. 2007.

FRY, J. 2007. Basic probability theory for natural language processing. www.sjsu.edu/faculty/fry/165/prob-2x2.pdf Date of access: 30 Jul. 2007.

SNYMAN, F.P.J. & NAUDÉ, J.A. 2003. The assessment of translation accuracy of the lexica machine translation system. *Journal for Southern African linguistics and applied language studies*, 21(4):295-306.

SUMITA, E., & IIDA, H. 1999. Experiments and prospects of example-based machine translation. http://portal.acm.org/citation.cfm?id=981368 Date of access: 25 Jun. 2007.

Key concepts:

administrative translation
domain specific
machine translation

Kernbegrippe:

administratiewe vertaling
domeinspesifiek
masjienvertaling