Developing and improving a statistical machine translation system for English to Setswana: a linguistically-motivated approach

Ilana Wilken
North West University
Potchefstroom, South Africa
ilanawilken@gmail.com

Marissa Griesel and Cindy McKellar
CTexT©
North West University
Potchefstroom, South Africa
{Marissa.Griesel; Cindy.McKellar}@nwu.ac.za

***Abstract* — This paper describes the methods that were followed in the development and improvement of a statistical machine translation system for translation from English to Setswana. Setswana is regarded as a resource scarce language and therefore an adequate amount of parallel data is not freely available. The methods created attempt to improve the quality of a machine translation by manipulating the data during processing. The methods include the creation of sentence reordering, term deletion and term replacement rules. The rules were applied to training and testing data in the pre- and post-processing stages of development. The systems were compared to one another to detect whether the quality of the machine translation improved.**

***Keywords*—**statistical machine translation, pre-processing, post-processing, sentence reordering, English, Setswana, term replacement, term deletion

I. Introduction

South Africa is a diverse, multi-lingual country and has eleven official languages [1]. According to the South African Bill of Rights [2], "everyone has the right to use the language…of their choice" as well as "the right of access to any information held by the state."

The South African government strives to provide information in all of the languages, but according to Prinsloo & De Schryver [3], corpora (and even more so parallel corpora) for all eleven official languages of South Africa is not always obtainable. Statistical machine translation (SMT) systems could serve as an additional tool for human translators to simplify, standardise, and expedite the translation process in the South African context.

For this research, it was decided to develop a SMT system for the translation of English to Setswana. Setswana falls into the Southeastern Bantu language group [4] and this research will contribute to the advancement of other closely related languages. These languages include Sesotho and languages in the Nguni, Tsivenda, and Xitsonga groups.

The development of the SMT system was done in two stages: first, a baseline system was developed. A text was translated and results were obtained. For the second stage, six adapted systems were developed. The adapted systems are an extended, a reordering, a replacement, a deletion, a deletion-replacement, and a deletion-reordering system. During the development of the adapted systems, linguistically motived rules were written and applied to the data. A text for each system was translated and individual results were obtained. The results of all the systems were compared to establish if an improvement of the quality of the translation took place.

The rest of this paper is organised as follows: Section 2 describes related work and Section 3 describes the development of the rules as well as the training of the systems. Section 4 gives details on the development of the adapted systems as a whole. Section 5 explains the evaluation of the systems as well as how the quality of the output of the systems improved. The conclusion and an overview of future work can be found at the end of the paper in Section 6.

II. Related Work

Truly automated machine translation of complex text cannot deliver output of the quality human translators would achieve. This project aims at improving the workflow and quality of language services in the government sector. Machine-aided human translation was therefore recognized as a means to achieve this aim. Machine-aided human translation can be explained as a draft translation initially done by a computer, but a human translator still remains responsible for correcting any errors. Such systems have already been developed for numerous international studies as well as for South African language pairs.

A machine translation (MT) system employing a pre-processing step is the English to Swahili, Swahili to English machine translation system [5]. The SAWA Corpus Project developed an English-Swahili parallel corpus and then built a SMT system for the application of the corpus. Swahili is a

strong agglutinative language and so words were first morphologically deconstructed to facilitate the connection between the morphemes and their corresponding English words. This improved the automatic word alignments drawn during the training phase. The very basic SMT system's results were compared to those of the Google Translate [6] system's results for Swahili. The results showed that the SAWA system disappointed in comparison to the Google Translate System for the translation from English to Swahili. The BLEU [7] and NIST [8] scores declined by 0.06 and 1.04 respectively. However, for the Swahili to English system, the SAWA system fared much better and showed improvements in the BLEU and NIST scores, increasing by 0.06 and 0.38 respectively.

A recent high scale machine translation project for South African languages was undertaken in the Autshumato project [9]. Smaller (research) experiments were previously conducted for SA languages, but this was the first project to develop an integrated strategy for the government domain. The project concentrated on translation from English to Afrikaans, Sesotho sa Leboa and isiZulu and a pre-processing method of syntactic reordering of the source language was used to improve on the results of the baseline systems [10]. The experiments showed positive results, resulting in improvements of the BLEU and NIST scores. The English-Afrikaans system's NIST score improved by 0.0274, whereas the English-Sesotho sa Leboa system's BLEU and NIST scores improved by 0.0406 and 0.5321, respectively.

SMT systems require great amounts of data, but large English-Setswana parallel data does not exist, because Setswana is considered a low resource language[1]. Simply gathering more bilingual data is not a practical option when developing SMT systems and so other methods to improve the quality of machine translation output is essential. The Autshumato project set a benchmark for machine translation for South African languages. Accordingly, the purpose of this research is to serve as an extension of the Autshumato project. For the English to Setswana SMT system developed in this research, it was decided to attempt improving the translation quality of the baseline system by applying pre- and post-processing steps. The steps include sentence reordering, as well as linguistically motivated deletion and replacement rules.

III. Rule Development

The data sets used for the development of the linguistic rules consist of 200 randomly selected sentences of each language and was taken from the training data mentioned above. The English data set was first translated with the baseline system to identify areas suitable for potential improvement. By comparing the English and Setswana data sets, it was noted that certain words exist only in one language and not in the other. The word order of the sentences did not align either. The original English data was then annotated with part-of-speech tags and by applying extensive linguistic knowledge [11], the rules were developed.

The numbers of core technologies to draw from are limited for Setswana and therefore limit the amount of processing that we are able to perform on the target language. However, numerous core technologies exist for English and it was decided that merely a part-of-speech tagger for English would be adequate for this project. The Stanford Log-linear Part-of-Speech Tagger (Stanford PoS Tagger) [12] was used to annotate the English training data and the development data set. This tagger was chosen because of the output data's usable quality. All of the reordering, deletion and replacement of words was done based on these tags. The rules were created and implemented using Perl [13] regular expressions.

The reordering and deletion rules are similar to those used by the Autshumato system for English-Sesotho sa Leboa. This is possible because both Setswana and Sesotho sa Leboa belong to the syntactically similar Sotho language family group [4]. However, a different approach was followed in the implementation thereof.

For this project, the rules were implemented individually, as well as in groups of rule sets. The deletion, replacement and reordering rules were implemented each on their own and so formed three of the six systems. These three systems are the deletion, replacement, and reordering systems. The deletion and replacement rules were grouped together, forming the deletion-replacement system; and the deletion and reordering rules were grouped together to form the deletion-reordering system. All the rules were then grouped together to form the extended system.

1. Deletion Rules

The deletion rules remove English words for which no Setswana equivalent exists. There are only three deletion rules and all three rules affect specific determiners in English. The determiners affected are *the*, *an* and *a*.

2. Replacement Rules

The purpose of the replacement rules is to ensure that the English conjunction word is translated with the correct Setswana conjunction word. In Setswana, the conjunction of nouns and the conjunction of verbs differ. When nouns are joined, *and* is translated as *le*, but when verbs are joined, *and* is translated as *mme*. Other conjunctions that are translated with the correct Setswana word is *or*, *but* and *because*. They are respectively replaced with *kgotsa*, *mme* and *ka gore*.

3. Reordering Rules

The reordering rules address the differences in the word order between English and Setswana. In Setswana, nouns are written first, followed by adjectives, and/or pronouns, and/or cardinal

[1] In 2005 the Pretoria Setswana Corpus consisted of 6 130 557 words, whereas the Pretoria English Corpus consisted of 12 799 623 words [3].

numbers, and/or specific determiners. The reordering rules change the order of the English words.

When the replacement and reordering rules are implemented on their own, determiners must be taken into consideration and the rules must be able to detect determiners when they are not deleted by the deletion rule. The rules can be explained as follows: the sequences of certain words are written in square brackets. *Possible:* means that an adjective, an adverb, or a determiner will be detected, but it will not matter if no adjective, adverb or determiner is present. When a word in a rule is written in bold in square brackets (for example [**or**]), it means that the word must be present for the rule to be applied.

The three basic rule groups were each applied separately and are set out below. An example sentence of the implementation of the rule is also given. The first sentence is the original sentence, as found in the baseline system, followed by the adapted sentence for that particular system.

A. Deletion System

- [determiner: **the** or **an** or **a**] → delete [determiner: **the** or **an** or **a**]

Example:

the status of a person as an only member of a state

→ **status of person as only member of state**

B. Replacement System

- [noun] [**and**] [possible: the or an or a] [possible: adjective] [noun] → translate *and* with *le*
- [verb] [**and**] [possible: the or an or a] [possible: adverb] [verb] → translate *and* with *mme*
- [conjunction **or**] → translate *or* with *kgotsa*
- [conjunction **but**] → translate *but* with *mme*
- [conjunction **because**] → translate *because* with *ka gore*

Example:

we have limited opportunities because we have limited resources and help from volunteers

→ **we have limited opportunities ka gore we have limited resources le help from volunteers**

C. Reordering System

- [specific determiner] [possible: *the* or *an* or *a*] [possible: adjective] [noun] → [possible: *the* or *an* or *a*] [possible: adjective] [noun] [specific determiner]
- [cardinal number] [**to**] [cardinal number] [possible: adjective] [noun] → [possible: adjective] [noun] [cardinal number] [**to**] [cardinal number]
- [cardinal number] [possible: adjective] [noun] → [possible: adjective] [noun] [cardinal number]
- [pronoun] [possible: adjective] [noun] → [possible: adjective] [noun] [pronoun]
- [adjective] [**and**] [adjective] [noun] → [noun] [adjective] [**and**] [adjective]
- [adjective] [noun] → [noun] [adjective]

Example:

the unacceptable misapplication of government power

→ **the misapplication unacceptable of government power**

The basic rules were also combined in three different systems to optimize the rule ordering. The rules for the deletion-replacement and deletion-reordering systems are similar to the separate rules explained above, but for these rules the determiners *the*, *an* and *a* do not need to be detected, since they are deleted before the next steps are reached. The combination of the rules and the changes affecting the rules are listed below:

D. Deletion-Replacement System

- [determiner: **the** or **an** or **a**] →delete [determiner: **the** or **an** or **a**]
- [noun] [**and**] [possible: the or an or a] [possible: adjective] [noun] →translate *and* with *le*
- [verb] [**and**] [possible: the or an or a] [possible: adverb] [verb] →translate *and* with *mme*
- [conjunction **or**] →translate *or* with *kgotsa*
- [conjunction **but**] →translate *but* with *mme*
- [conjunction **because**] →translate *because* with *ka gore*

Example:

having to report and explain to a higher authority

→ **having to report mme explain to higher authority**

E. Deletion-Reordering System

- [determiner: **the** or **an** or **a**] →delete [determiner: **the** or **an** or **a**]
- [specific determiner] [possible: *the* or *an* or *a*] [possible: adjective] [noun] → [possible: *the* or *an* or *a*] [possible: adjective] [noun] [specific determiner]
- [cardinal number] [**to**] [cardinal number] [possible: adjective] [noun] → [possible: adjective] [noun] [cardinal number] [**to**] [cardinal number]
- [cardinal number] [possible: adjective] [noun] → [possible: adjective] [noun] [cardinal number]

- [pronoun] [possible: *the* or *an* or *a*] [possible: adjective] [noun] ⟶ [possible: *the* or *an* or *a*] [possible: adjective] [noun] [pronoun]
- [adjective] [**and**] [adjective] [noun] ⟶ [noun] [adjective] [**and**] [adjective]
- [adjective] [noun] ⟶ [noun] [adjective]

Example:
compulsory enlistment in the armed forces
→ **enlistment compulsory in forces armed**

F. Extended System

All of the abovementioned rules were applied during the training of the extended system. The order of the rules is as follows:

- Deletion rules
- Replacement rules
- Reordering rules

Example:
a branch or subdivision of the public service and the relationship between the state and its citizens
→ **branch kgotsa subdivision of service public mme relationship between state le citizens its**

IV. Training of the Systems

The training data used in this research project consist of a parallel corpus[2] of English-Setswana sentence pairs and a monolingual Setswana corpus for language modelling. The corpora contain data from the South African government domain. The parallel corpus was automatically aligned with an algorithm developed by Robert Moore [14]. The open source statistical machine translation toolkit, Moses [15], was used for the training of both the baseline and adapted systems and the SRILM toolkit [16] was used to train the language model. Table 1 indicates the quantity of data used.

TABLE I. Data Quantity

Corpus	**Number of sentences / -pairs**
Parallel Corpus	34 321 English-Setswana sentence pairs
Monolingual Corpus	50 923 Setswana sentences

[2] For more information on these corpora, please contact CTexT® [12]

V. The Development of the Adapted Systems

The different systems were created to evaluate which linguistic rule – whether on its own or grouped together – provided the best translation quality of a translated text. An example of how the adapted systems were developed can be seen in Fig. 1. This example shows the development of the extended system, where all the rules are grouped together. For the other systems, the applied rules are adapted to suit each system.

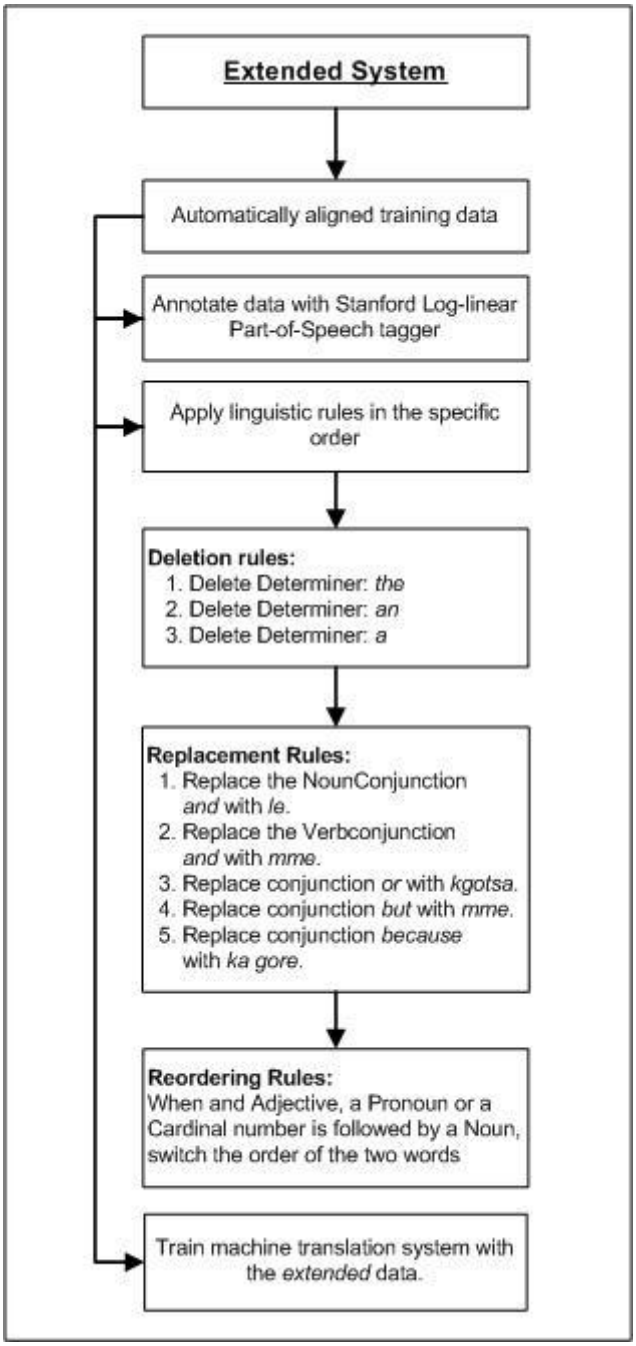

Fig. 1. Development of an adapted system

All the linguistic rules were applied during pre-processing. However, post-processing of the *the*-deletion rule was necessary for the deletion, deletion-reordering, deletion-replacement and extended systems. The reason for this is that the Setswana data contain English words and phrases. Since we have such a small amount of data and because we do not have the means to cleanse the data manually, we decided not to dispose of the sentences containing English words and phrases.

The data containing English words poses a big problem, because when the language model is trained, the English words in the Setswana data are seen as Setswana words. They are therefore included in the Setswana language model and when an English text is then translated, an English word will be translated with a 'Setswana' word, when the word is in fact also an English word.

When the testing data was translated, the *the*-determiner was detected in the translated text. Post-processing was the preferred method of choice because it is quick and effective in the removal of wrongly inserted/translated English words. The removal of these words could also have a positive impact on the word-level evaluations done later.

VI. Evaluation

Testing data consisted of 500 manually aligned English-Setswana sentence pairs. The alignments were done using the CTexT® Alignment Interface [9]. This data set is from the same domain as the training data; however, none of these 500 sentence pairs appear in the training data.

A baseline machine translation system was trained and the original testing data was translated. All six adapted systems received the same testing data set, but the manner in which the linguistic rules were applied differed, as explained in section V. The adapted testing data sets were translated and results were obtained. The results of all six adapted systems were compared to the results of the baseline system.

The output quality of a machine translation system can be evaluated in two ways: human evaluation or automatic evaluation [17]. Only the automatic evaluation was used for this project, because it is a sufficient way of obtaining reliable results immediately. These results determine whether or not the quality of the translation generated by the SMT system improved.

The BLEU and NIST scores were calculated for each machine translation system. A BLEU score measures the closeness between a machine translation and a reference translation. The quality of the machine translation is determined by how identically similar it is to the professional human translation. The BLEU evaluation is done according to a numerical metric, ranging from 0 to 1. When a score of 1 is reached, it means that the translated text is as identically similar to the reference translation as possible.

The machine translation output and the reference translations are compared in terms of the statistics of short sequences of words, also known as word n-grams. The NIST score calculates how informative a particular n-gram is. The translation quality is judged to be at its best when a translation shares as many n-grams as possible with the reference translation. Table 1 indicates the BLEU and NIST scores for the baseline and extended machine translation systems. The BLEU scores are also represented as percentages, as well as the difference between the baseline system and the other systems indicated in brackets.

For each of the adapted systems, the results showed gains in both the BLEU and NIST scores when compared to the results of the baseline system. The biggest gain was for the extended system, where both the BLEU and the NIST scores improved by 0.0136 (1.36%) and 0.0896 respectively.

The system that showed the least improvement of both the BLEU and NIST scores is the replacement system. The reason is that no word reordering or replacement took place – one word is merely translated with another, which might be correctly translated by the system from the start.

A translation is evaluated on different levels when evaluation takes place. They are: an overall scoring, an individual n-gram scoring, and a cumulative n-gram scoring level. The overall scoring level presents the BLEU and NIST scores used to determine if an improvement in a system's translation quality took place. The individual n-gram level does comparisons of the translated and reference data on an isolated n-gram level. The levels range from 1-gram to 9-gram, meaning one word is compared to one word, two words to two words, and so forth until nine words are compared to nine words. The cumulative n-gram level does comparisons of groups of words that occur in the translated and reference data. These levels also range from 1-gram to 9-gram, but now the first word is compared to the first word of the translated text, the first two words of both texts are compared, and so it continues until a group of nine words are compared to a group of nine words of both texts.

When the NIST scores are compared for the individual n-gram scoring level, the Extended system fares the best of all the systems evaluated, showing improvements from 1-gram through to 8-gram. On the other hand, the replacement and deletion-replacement systems indicated the second and third highest improvements of the NIST scores of the individual n-gram scoring level. Both their improvements range from 1-gram through to 7-gram. It would therefore be safe to say that when results seem to be unimportant at first, there could be prospective improvements on a smaller scale. These small-scale improvements might be equally as useful as the overall results.

TABLE II. EVALUATION RESULTS

System	BLEU	BLEU %	NIST
Baseline	0.2744	27.44%	6.0911
Extended	0.2880	28.80% (+1.36%)	6.1807
Reordering	0.2781	27.81% (+0.37%)	6.1155
Replacement	0.2751	27.51% (+0.07%)	6.1049
Deletion	0.2813	28.13% (+0.69%)	6.1488
Deletion-Reordering	0.2861	28.61% (+1.17%)	6.1444
Deletion-Replacement	0.2817	28.17% (+0.73%)	6.1495

To determine whether the difference in the overall BLEU score results is statistically significant for the baseline and extended systems, a two-sample t-test between proportions was performed with a statistical calculator [19]. The following hypothesis was made: the null hypothesis states that the difference between the results of the baseline system and the extended system is statistically significant. For this hypothesis, the p-value calculated must be bigger than the significance level alpha (α), so that the null hypothesis is not rejected. The p-value and α-value were calculated as follows, using the test set of 500 sentences as the samples:

$$p = 0.31625$$
$$\alpha = 0.05$$

Thus:

$$0.31625 > 0.05$$
$$p > \alpha \quad (1)$$

The p-value is bigger than the α-value, therefore the null hypothesis is not rejected, and the difference between the BLEU scores of the baseline and extended systems is statistically significant.

VII. CONCLUSION AND FUTURE WORK

Although the results showed marginal improvements, it indicates that there is potential for a machine translation system for English to Setswana using these pre- and post-processing methods.

This is an initial experiment and only one reference translation was used. However, for an automatic evaluation to be truly successful, a number of reference translations are needed. The need for these extra reference data is because two separate translations of the same text done by the same (or different) translator are seldom identical. Synonyms play a big part in translations and a SMT system does translations based on its language model, which might not always contain all possible synonyms of a target language. A machine translation system will only be an effective tool to human translators if the translator does not spend more time adjusting the output than doing a translation from scratch. A human evaluation will certainly give information as to how good the quality of the machine translation really is and how useful it would be in an everyday working environment. For future experiments, more reference translations and human evaluations will be used.

Also included in future work, is the assessment of the linguistic rules in isolation. This will determine the effectiveness of the rules' application. The rearrangement of the rules before implementation might have a positive effect on the success of other rules, by ensuring that one rule doesn't overwrite another in the processing stage. The rules can also be extended to include the correct translation of the time forms of the verbs as well as the direct relative verb construction.

As the results obtained indicate, SMT systems with these particular pre- and post-processing methods show that by developing SMT systems for a resource scarce language like Setswana, improvements in the translation quality can be achieved. However, because the development of machine translation systems is never-ending and because there is still room for improvement of the system as a whole, continuous effort will be made to achieve the highest translation quality possible.

REFERENCES

[1] South African Government information: South Africa at a glance, http://www.info.gov.za/aboutsa/glance.htm, 2012.

[2] Republic of South Africa, "Constitution of the Republic of South Africa: Chapter 2 - Bill of Rights", http://www.info.gov.za/documents/constitution/1996/96cons2.htm, 1996.

[3] D.J. Prinsloo and G-M. de Schryver, "Managing eleven parallel corpora and the extraction of data in all official South African languages," In Multilingualism and Electronic Language management. Proceedings of the 4th international MIDP Colloquium, 22-23 September 2003, Bloemfontein, South Africa. W. Daelemans, T. du Plessis, C. Snyman and L. Teck (eds.), pp. 100-122, Pretoria, Van Schaik Publishers, 2005.

[4] D. Joffe, "African Languages: Setswana (Tswana)," http://africanlanguages.com/setswana/, 2012.

[5] G. De Pauw, P.W. Wagacha and G.M. De Schryver, "Towards English-Swahili machine translation," Proceedings of Machine translation and morphologically rich languages: Research workshop of the Israel Science Foundation, pp. 1-2, University of Haifa, Israel, 2011.

[6] See http://translate.google.com/ for more information on Google Translate.

[7] K. Papineni, S. Roukos, T. Ward and W.J. Zhu, "BLEU: A method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting of the Association for

Computational Linguistics, pp. 311-318, Philadelphia, USA, 2002.

[8] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," Proceedings of the 2nd International Conference on Human Language Technology Research, pp. 138-145, San Diego, California, 2002.

[9] H.J. Groenewald and L. Du Plooy, "Processing parallel text corpora for three South African language pairs in the Autshumato project," Proceedings of the 2nd Workshop on African Language Technology, pp.27-30, Valetta, Malta, 2012.

[10] M. Griesel, C.A. McKellar and D. Prinsloo, "Syntactic reordering as preprocessing step in statistical machine translation from English to Sesotho sa Leboa and Afrikaans," Proceedings of the 21st annual Symposium of the Pattern Recognition Association of South Africa (PRASA), pp. 205-110, Stellenbosch, South African, 2010.

[11] A.S. Berg and R.S. Pretorius, "Tswana: taalkunde, leeswerk en stelwerk," Study guide ATSW 114 A & 124 A, North-West University, Potchefstroom Campus, South Africa, 2009.

[12] K. Toutanova, D. Klein, C.D. Manning and Y. Singer, "Feature-rich part-of-speech tagging with a Cyclic Dependency Network, Proceedings of HLT-NAACL, pp. 252-259, Edmonton, Canada, 2003.

[13] See http://www.perl.org/ for more information about the programming language.

[14] R. Moore, "Fast and accurate sentence alignment of a bilingual corpus," Proceedings of the 5th conference of machine Translation in the Americas, pp. 135-144, Tiburon, CA, 2002.

[15] P. Koehn et al., "Moses: Open source toolkit for statistical machine translation," Proceedings of the ACL demo an poster sessions, pp. 177-180, Prague, Czech Republic, 2007.

[16] A. Stolcke, "SRILM – an extensible language modeling toolkit," Proceedings of the 7th International Conference on Spoken Language Processing, pp. 901-904, Denver, Colorado, 2002.

[17] D. Jurafsky and J.H. Martin, "Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition," 2nd ed. pp. 897-931, New Jersey: Pearson Education, 2009.

[18] See http://www.nwu.ac.za/ctext for more information on the Centre for Text Technology (CTexT®).

[19] StatPac, "The Statistics Calculator: statistical analysis at your fingertips," http://www.statpac.com/statistics-calculator/percents.htm, 2012.