Experiments with syllable-based Zulu-English machine translation

Friedel Wolff
Academy of African Languages and Science
University of South Africa
wolfff@unisa.ac.za

Gideon Kotzé
Academy of African Languages and Science
University of South Africa
kotzegj@unisa.ac.za

***Abstract*—Due to morphological complexity and scarce resources, machine translation from Zulu to English is challenging. We investigate the possibility of phrase-based statistical machine translation from Zulu to English using syllables as the tokens in the Zulu source text. Initial experiments on a relatively small but multi-domain data set suggest merit in our approach, with our best syllable-based model outperforming the best word-based model by 12,90% using the BLEU evaluation measure. Our syllabification approach is largely language independent, at least within the Bantu language family, and holds promise for similar efforts in related languages.**

I. Introduction

Statistical machine translation (SMT) requires vast textual resources as training data in order to deduce the information required for good quality output. This is required for good coverage of the source lexicon, accurate lexical transfer, and fluent output in the target language. For languages with less data available, this is a limiting factor.

Zulu has a complex morphology which can result in full clauses (and indeed even a full sentence) to be written as a single word. In terms of SMT this leads to very sparse language models and little data per unique word, since many words are possible—most being very unlikely. This compounds the problem of limited resources.

Proper morphological analysis would be the desired way to deal with this problem. This would reduce the data sparsity by separating morphemes, thereby reducing the size of the "lexicon" (number of unique tokens) and providing more training opportunities for each token—both for lexical coverage and lexical transfer. However, constructing a morphological analyser for a language is time consuming and the end result is language specific.

In this paper, we investigate the use of Zulu syllables as tokens in a Zulu-to-English SMT experiment, while keeping the English words unchanged. While processing language data on the syllable level is not unusual in the field of speech processing, this is definitely not as common in the field of text processing. At least in terms of phrase-based SMT, we believe this is a novel contribution.

We believe that our choice of Zulu-to-English translation has a few advantages:

- We can benefit from many available resources for building a target language model for English. We can therefore investigate the consequence of using syllables during the alignment and the decoding phase of the machine translation, without it also having an influence during language modelling. The quality of the target language model is therefore hopefully eliminated as a source of big concern.
- The output is normal English and does not require any post-processing. If we were translating in the other direction, the output would be syllabified Zulu text.
- A cursory inspection of the output is possible by anybody with knowledge of English—an advantage at this stage of our research.

II. Background

Statistical machine translation is based on the idea of viewing the text in the source language as a variant of the target language that was transmitted through a noisy channel. The search for the best English translation $\hat{E}$ is often formulated according to Bayes' rule as follows:

$$\hat{E} = \underset{E \in English}{argmax}\ P(Z|E)P(E) \tag{1}$$

where Z is the Zulu (input) text and where we attempt to choose or construct the optimal E from the training data to optimise the probabilities. The first factor $P(Z|E)$ refers to decoding using the translation model, and the second factor $P(E)$ to language modelling to ensure fluent output in the target language.

For phrase-based SMT, if the phrase table (translation model) does not contain a certain phrase, it would have to fall back to smaller and smaller constituents, until it reaches the word level. If a source language word is not present in the phrase table, it can not be translated, and normally might simply be duplicated into the target language, or dropped entirely.

Zulu morphology is characterised by agglutination, and the orthography by a conjunctive writing system. This means that many forms can be derived from each stem by combining with various prefixes and suffixes to form a single word. Up to thousands are possible for each verb, for example. This means that a verb stem might occur in many different surface forms in a training corpus, on face value a different word from occurrences of other forms. Although SMT engines such as

Moses allow for factored models [1] that can process each word annotated with a stem and/or part-of-speech information, this remains a problem when an accurate morphological analyser and/or part-of-speech tagger is not available—the case for many languages in the Bantu family. While bootstrapping of rule-based morphological analysers [2] shows promise, it remains a large undertaking to build a high-quality lexicon for a language.

Some languages in the Bantu family, such as those in the Sotho group, use a disjunctive writing system. Although there is still affixation, the problem of sparseness in a phrase table is not as pronounced as it is in Zulu and languages with a similar orthography.

While a morphological analyser for Zulu exists [3], that still leaves the matter of choosing a single analysis from the alternatives. Furthermore, we believe our approach is significantly simpler and far less language dependent.

The intuition for using syllables as cheap substitutes for morphemes, stems from the fact that many Zulu prefixes are indeed single syllables and often have obvious alignments with a parallel English text. For example, verb prefixes indicating subject (*-ngi-*) and negation (*a-*) are single syllables in *a+ngi+hamb+i* (English: I don't walk; literally: not+I+walk+not). Although multi-syllabic stems will be oversegmented, our hope is that they will be transferred to the target language semantically in tact due to frequent co-occurrence.

Zulu and most languages in the Bantu family have a preference for open syllables [4]. This means that syllables occur in a very regular way, making syllabification easy to perform, even though this is only a rough substitute for proper morphological analysis.

Using syllables in text processing has been attempted with success for information retrieval in European languages [5] and word alignment between English and Zulu [6]. The early results from the latter work are promising, considering the small dataset used in that experiment.

A small note appears in [7] about a machine translation group at IBM that "were inclined to test the hypothesis that the syllable is the carrier of the morpheme as the minimal semantic unit". Whether any real work in this direction was carried out, is not clear. Considering the state of the art in 1960, we can safely assume that a separate consideration here in terms of current-day phrase-based SMT is fully justified, especially for our language combination.

The automatic induction of a morphological analyser is possible with supervised, semi-supervised and unsupervised methods [8], [9]. Compared to our simpler syllable-based approach, we note some differences:

- It requires vast textual resources for training—not an ideal situation for the environment we are working in. Our investigation into the feasibility of using tokens below the word level for SMT is in part inspired by the resource scarceness of Zulu and the related languages on which we hope to apply our technology.
- Being based on machine learning, it would have some level of dependence on the domain and style of the training data. Our approach is inherently domain and genre independent.
- Depending on the exact approach of the induced morphological analyser, the matter of disambiguating between analyses might remain. Although all analyses can be added to a lattice in the SMT engine, it is not clear how ambiguous analyses for tokens in a sentence would be handled in token alignment. Disambiguation is not required with our syllable-based approach, since only a single output is produced.

In [10], such unsupervised morphological analysers were used to generate the training data for morpheme-based machine translation engines which resulted in slightly lower evaluation scores (according to the BLEU metric) within the context of the complex morphology of the Nordic languages. Morphological analysis on Swahili text has been shown to improve word alignment [11].

Previous work in SMT from Zulu to English is limited. It was attempted as one step in a pivoting approach to translate from Xhosa to English [12]. Google Translate was augmented with support for Zulu in 2013. The Autshumato project [13] also worked on these languages, but only on the English-to-Zulu direction.

In this work, we build on the results reported in [6] by performing a series of machine translation experiments based on syllabified Zulu text.

III. Data and preprocessing

For training an SMT system, one needs a large collection of preprocessed parallel data, also called *bitexts*. We extracted text from three different sources for this purpose: the Bible, the Autshumato English-Zulu corpus [13] and the South African constitution of 1996.[1]

The Zulu Bible is the version from 1959,[2] whereas the English text is from the World English Bible British English Edition[3]. The choice of this English Bible has a few advantages:

- Although it is based on the King James translation, the lexicon is significantly more modern in comparison. Words such as "thy", "thou" and "betwixt", were replaced as a policy for this translation.
- It is in the public domain, which removes copyright constraints.
- It is published in a clean XML format for easy processing.
- Being based on the King James translation, it uses a formal equivalence approach for translation, similar to the Zulu Bible [14]. Additionally, the use of punctuation (for example quotation marks) seems very similar. This raised our hopes for higher quality alignment.
- This specific variant of the World English Bible follows British spelling, which would be more appropriate for the

[1] http://www.polity.org.za/polity/govdocs/constitution/
[2] http://wordproject.org/
[3] http://ebible.org/

South African context. Admittedly, at this early stage in our research, this is not a big consideration.

Preprocessing of the parallel corpus consists of the following steps:

- removal of undesirable elements, such as markup and erroneously encoded segments
- tokenisation, which consists of separating words from punctuation, such that each type can be processed as a separate unit during the training and translation process
- sentence splitting, in order to limit the search space for algorithms to one sentence pair at a time. This is a standard practice in the SMT community.
- sentence alignment, in order to pair the correct parallel sentences so that their translations can be learned. This is important since none of our bitexts are perfectly aligned. The Bible and constitution texts have been extracted separately from their sources, whereas we also found some misaligned sentences in Autshumato. The tool that we used for this purpose is Hunalign [15].[4]
- segmentation of Zulu words into syllables. We implemented a script which regards each vowel as signifying the end of a syllable. This is based on the open syllable assumption mentioned in section II.
- removal of duplicate sentence pairs. This is partly since it seems that the Autshumato dataset already contains parts of the constitution.

For the tokenisation process we applied TreeTagger [16][5] for both the English and the Zulu text. We additionally ensured that the em-dash (Unicode: U+2014) was also correctly isolated as a token on its own.

For sentence splitting, we used the split-sentences.perl script[6] that is included with Moses [17], the SMT system that we adopted for our experimental work. We implemented an additional post-processing script in order to deal with phenomena that the Moses script failed to address, particularly embedded quotes.

Proper tokenisation usually requires a list of abbreviations so that periods are not mistaken for full stops, keeping the abbreviations together as single tokens and making them easier to learn as single units of meaning. An example would be "S.E." which refers to "south-east". The existence of "S.E." in the abbreviation list would ensure that the output is written as "S.E." instead of "S . E . ". Apart from the list of English abbreviations bundled with TreeTagger, we also supplied a small list of Zulu abbreviations.

An interesting property of both the Zulu and English Bible texts we use, is the somewhat frequent use of semicolons and quotations—even embedded quotations several levels deep. We therefore decided to additionally consider semicolons and colons as sentence boundaries. This reduces the average length of sentences, which we think should also improve word alignment.

[4]http://mokk.bme.hu/en/resources/hunalign/

[5]http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[6]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl

TABLE I
CORPUS STATISTICS AFTER SENTENCE ALIGNMENT

	Bible	Autshomato	Constitution	Total
Sentence pairs	39 916	36 292	2 788	78 996
Zulu words	380 432	415 976	36 190	832 598
Zulu syllables	1 116 900	1 428 983	109 974	2 655 857
English words	626 187	554 212	47 602	1 228 001

For an overview of the corpus composition, see Table I. In the composition of the SAWA corpus for English-Swahili [11], a similar majority of religious text was also reported, even though a greater attempt was made there to include texts of a wider variety.

The factor $P(E)$ from equation 1 refers to modelling for the target language, which contributes to a more fluent output. To achieve this, monolingual data in the target language is processed in a separate step during training.

For the target language modelling we used the target side of the bilingual data mentioned above (including unaligned segments, but excluding tuning and test data). We would have preferred to augment the model by mostly using South African English text. However, many American, British and European corpora are easily available and of substantial size. We therefore included a selection of three books from Project Gutenberg[7] relating to South Africa as well as a 1% random sample of the English side of the English-French EU bookshop corpus [18][8] and a 1% random sample from the English part of the WMT13 2012 news corpus[9]. We tokenised the text and removed duplicates, just as with the bitexts. The final size of this monolingual corpus is 7 886 613 tokens.

IV. EXPERIMENT

All Zulu source text in the training data was syllabified prior to training. Zulu input to the engine (for example tuning and testing data) is also syllabified before being given to the SMT engine proper. Training the SMT engine on syllabified Zulu text takes substantially longer than for the normal, unsyllabified text. For example, we found that the word alignment step for text where the Zulu has been syllabified takes almost five times as long as alignment for the normal word-based approach.

After having preprocessed all the texts, we randomly divided the parallel corpus as follows: 90% for training, 5% for tuning, 5% for testing. The test set was divided into two, one of which we regard as a development test set and the other as a final test set, only to be used once. However, we ended up only testing our approach on the development test set once, and therefore, we will report those results. In any case, as the data from the final test set comes from the same documents and was extracted in the same way, we do not expect any meaningful differences. In future experiments, we plan to use an out-of-domain test set that should properly test the robustness of our models.

[7]http://www.gutenberg.org/

[8]http://opus.lingfil.uu.se/EUbookshop.php

[9]http://www.statmt.org/wmt13/translation-task.html

As mentioned before, our SMT system of choice was Moses. Its Experiment Management System (EMS) was utilised, from which we used the default values of most parameters. We mention some specific details of our experimental setup:

- We used KenLM [19], [20] for target language modelling. Most default settings were unchanged. It uses Modified Kneser-Ney smoothing [21], [22] with no pruning.
- For word alignment, we used MGIZA++ [23], which is a multi-threaded implementation of GIZA++ [24] leading to faster performance.
- The maximum length of extracted phrases was 5, which is the default. In the future, we might experiment with longer lengths, as the Zulu tokens are finely segmented.

Finally, we also kept a data set with unsplit Zulu words as a baseline, in order to compare the effect of syllabified Zulu on SMT performance. For both of these sets, we ran Moses using eight different word alignment approaches that are implemented by Moses. MGIZA++ alignments are asymmetrical, meaning that a source-to-target alignment differs from the target-to-source alignment on the same bitext. Apart from the intersection and union of these alignments—leading to a high-precision and high-recall set, respectively—there is also a set of *heuristics*, each one of which takes the intersection and adds additional links to a specified set of neighbouring tokens. The idea behind this is that it is usually regarded as more likely that the neighbours of tokens with high alignment probabilities are also aligned.

V. Evaluation

In the evaluation stage, the Zulu part of the test set is translated into English by the decoder, using the model that has been trained. The English output is then compared to its equivalent in the test set. Quantitatively, the differences between the automatic output and the test set English can be expressed as a score using one of a number of measures.

For our evaluation metrics, we used BLEU [25] and TER [26], [27]. BLEU is especially regarded as the *de facto* standard today. Note that TER is a measure of the rate of error and therefore, lower scores are better.

On a reduced data set, we found that the BLEU scores of the syllabified texts outperformed those of the word-based approach across the board, apart from one of the heuristics (*grow-diag-final*). The reason for the latter is not yet clear, although so-called "blind heuristics" may of course introduce some noise. For TER, the word-based approach reversed the situation, leading to better scores. However, the best word-based TER is still just barely better than the TER of the best syllable-based model.

The results mostly agreed with the results in word alignment using the heuristics in [6]. Since our work in that paper reported on alignment from English to Zulu, those results are comparable to ours here since we model probabilities in the same direction for the estimation of $P(Z|E)$ (equation 1). In the future we hope to report in detail on the effects of alignment heuristics on our SMT results.

TABLE II
BLEU / TER scores for best translation systems based on truecased text.

Model	Syllables	Words
intersection	**30.98** / 0.60	21.07 / 0.64
target-to-source	29.47 / 0.62	27.44 / 0.58
Google Translate	N/A	29.12 / **0.56**

Based on these findings, we selected the alignment approaches resulting in our best models and applied them to the full data set for comparison. For reference, we also include results of Google Translate[10]. It is important to keep in mind that Google Translate is trained on different, and almost definitely more parallel data, and that it obviously benefits from a substantially larger English language model. It could even be trained on segments from the test set. On the other hand, our engine has been trained on segments from the same documents from which the test set was extracted.

Table II shows our evaluation results on truecased text. One may notice that the scores are relatively high. This can be easily explained by virtue of the fact that the test set has been extracted from the same documents. Although there are no overlapping sentence pairs between the training data and the test data, we did notice some which are similar.

Another interesting point is that the highest BLEU score is produced by intersection alignments. This is quite unusual since the stochastic nature of SMT usually benefits from a high-recall approach, hence the existence of the beforementioned heuristics. We hope that a closer inspection of the alignments in question may reveal the reasons for this.

The target-to-source alignments also lead to very good results. We are surprised that both the word-based and syllable-based approaches perform well with this alignment approach.

For these final models, the next logical step would be some kind of tuning process. Often, Minimum Error Rate Training (MERT) [28] is applied. Because this is quite a slow process, time constraints have forced us to postpone it for future work.

On a qualitative level, we inspect a few examples from the automatically translated test set. In each case we show the results of the best word-based model and the best syllable-based model. We share and discuss a few interesting examples with untranslated input indicated in italics.

The first noteworthy attribute of the output from our syllable-based model is far less out-of-vocabulary words, as anticipated. Here is a somewhat typical example where non-trivial morphology results in untranslated entries in the word-based model, while the syllable-based model produces a translation with perfect lexical coverage, and reasonable semantic transfer:

Source text	"Uvunyelwe ukuba uzikhulumele."
Word model (baseline)	"*uvunyelwe* to *uzikhulumele*."
Syllable model	"You shall be allowed to speak."
Reference	"You may speak for yourself."

In a few cases, lexical choices were also superior:

[10]Obtained manually through the web interface in September 2014.

Source text	Nengqikithi yentengo yokuqhutshwa kwebhizinisi, ...
Word model (baseline)	Content price of the transaction, ...
Syllable model	And the total price of the transaction, ...
Reference	and the total price of the transaction, ...

Proper names were more frequently translated correctly with syllabification (in the word-based model, they are often rendered untranslated with prefixes still attached, as in the input text):

Source text	njengezwi likaJehova ngesandla sikaSamuweli.
Word model (baseline)	according to the LORD's word by *sikaSamuweli*.
Syllable model	according to the LORD's word by Samuel.
Reference	according to the LORD's word by Samuel.

English words in the Zulu source text caused strange lexical choices with the syllable model, such as "burn-out" being translated with the proper name "Bezuidenhout" (note the similar word endings). While it is somewhat understandable that this happens within our model, the results are definitely not ideal. Borrowed words such as these are difficult to translate anyway, and might be best handled by preprocessing to avoid putting them through syllabification and the SMT engine.

The English output sometimes seemed like more of a literal translation. For example, the owner in possessive clauses were often in post-position with the preposition "of" instead of using the possessive clitic "'s" or a noun compound.

Source text	Imiphumela yokuhlaziywa kwamanzi ...
Word model (baseline)	the results of the *yokuhlaziywa* ...
Syllable model	the results of the analysis of water ...
Reference	Results of water analyses ...

Although this is a variation on natural English text, it does not detract in any way from comprehensibility or grammatical correctness.

The lack of gender information in Zulu pronouns and object and subject references are an expected shortcoming as seen in this example:

Source text	wambamba ngengubo yakhe, wathi:
Word model (baseline)	Took his cloak, said:
Syllable model	He took him by his garment, saying,
Reference	She caught him by his garment, saying,

Here we see the object and subject morphemes being translated with an unnecessary pronoun in English with both models. While this is not optimal English, it remains quite comprehensible:

Source text	U-Asaheli wamxosha u-Abineri;
Word model (baseline)	Asahel chased him Abner;
Syllable model	Asahel he pursued Abner;
Reference	Asahel pursued Abner;

The syllable-based model did not perform better than the word-based model in all cases. Here is an example of incorrect word order, probably due to insufficient long distance modelling:

Source text	USiba wayenamadodana ayishumi nanhlanu nezinceku ezingamashumi amabili.
Word model (baseline)	Ziba fifteen sons and twenty.
Syllable model	Ziba had sons and five and twenty servants fifty.
Reference	Now Ziba had fifteen sons and twenty servants.

The correct lexical transfer and required reordering of the noun and qualifying numeral is performed successfully by the word-based model, but not the syllable-based model. In the syllable-based model, the input would contain *wa ye na ma do da na a yi shu mi na nhla nu*, a phrase of length 14, thereby exceeding the maximum phrase length of the SMT engine, making lexical reordering on the level of real words impossible.

VI. Future work

Since the Zulu text is very finely segmented in our experiment, it would make sense to repeat this experiment with some level of morphological analysis on the English text. This should cause more fine-grained alignments on the word level. For example, the tokeniser we used did not isolate possessive clitics in English, which would have allowed better alignments with possessive prefixes in Zulu.

If a morphological analyser can be adapted to produce unique analyses, we believe better results would be possible than with syllabification. It would put the performance obtained here in context. However, because it is language independent and simple to implement, we believe this approach still has value for other languages. The comparison of the two approaches for Zulu would give an indication of the improvement to expect in related languages if a morphological analyser would become available.

Many more parameters of the SMT engine can be investigated, such as the alignment heuristics mentioned above. We chose to mostly keep to the default configuration, which still leaves room for possible improvement. Specifically, tuning the translation systems to a held-out set should still improve results.

The size and composition of our training data, both parallel and monolingual, could still be improved. A good test set of completely out-of-domain sentences might also give a different perspective on the results obtained here. Out of necessity, this experiment used only single reference translations. Using more reference translations, or a measure such as HTER [27], should give a more accurate measure for evaluation.

VII. Conclusion

Our approach to use Zulu syllables as the tokens in the source text when applying phrase-based SMT to English seems to have merit. The output contains less out-of-vocabulary words, and sometimes improved lexical choice over the word-based baseline. The results using a small parallel training corpus is also quite encouraging.

Acknowledgement

This research was supported in part by funding from the Academy of African Languages and Science Strategic Project of the University of South Africa.

REFERENCES

[1] P. Koehn and H. Hoang, "Factored translation models," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Prague, Czech Republic, June 2007.

[2] L. Pretorius and S. Bosch, "Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology," in *Proceedings of the First Workshop on Language Technologies for African Languages*, ser. AfLaT '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 96–103. [Online]. Available: http://dl.acm.org/citation.cfm?id=1564508.1564526

[3] L. Pretorius and S. E. Bosch, "Finite-state computational morphology: An analyzer prototype for Zulu," *Machine Translation*, vol. 18, no. 3, pp. 195–216, 2003.

[4] P. Spinner, "Review article: Second language acquisition of Bantu languages: A (mostly) untapped research opportunity," *Second Language Research*, vol. 27, no. 3, pp. 418–430, 2011. [Online]. Available: http://slr.sagepub.com/content/27/3/418.abstract

[5] K. Kettunen, P. McNamee, and F. Baskaya, "Using syllables as indexing terms in full-text information retrieval," in *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2010, pp. 225–232. [Online]. Available: http://dl.acm.org/citation.cfm?id=1860924.1860962

[6] G. Kotzé and F. Wolff, "Experiments with syllable-based English-Zulu alignment," in *Proceedings of the SaLTMiL Workshop on free/open-source language resources for the machine translation of less-resourced languages (at LREC 2014)*, Reykjavik, Iceland, 2014, pp. 7–11.

[7] M. Zarechnak, "The history of machine translation," in *Machine Translation*, B. Henisz-Dostert *et al.*, Eds. Mouton Publishers, 1979, pp. 1–87.

[8] S. Spiegler, B. Golénia, K. Shalonova, P. A. Flach, and R. C. F. Tucker, "Learning the morphology of Zulu with different degrees of supervision." in *SLT*, A. Das and S. Bangalore, Eds. IEEE, 2008, pp. 9–12. [Online]. Available: http://dblp.uni-trier.de/db/conf/slt/slt2008.html#SpieglerGSFT08

[9] U. Quasthoff, S. Bosch, and D. Goldhahn, "Morphological analysis for less-resourced languages: Maximum affix overlap applied to Zulu," in *Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era, LREC, Reykjavik*, 2014.

[10] S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi, "Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner," in *Proceedings of the MT Summit XI*, 2007, pp. 491–498.

[11] G. De Pauw, P. W. Wagacha, and G.-M. Schryver, "Exploring the SAWA corpus: collection and deployment of a parallel corpus English-Swahili," *Language Resources and Evaluation*, vol. 45, no. 3, pp. 331–344, 2011. [Online]. Available: http://dx.doi.org/10.1007/s10579-011-9159-7

[12] B. Sharwood, "Machine translation of under-resourced languages," University of Cape Town, Honours project report, 2013. [Online]. Available: http://people.cs.uct.ac.za/~bsharwood/downloads/BeeSharwood_Report.pdf

[13] C. A. McKellar and H. J. Groenewald, "Frequency-based data selection for statistical machine translation with scarce resources," in *Language Science and Language Technology in Africa: Festschrift for Justus C Roux*, H. S. Ndinga-Koumba-Binza and S. E. Bosch, Eds. Stellenbosch: Sun Media, 2012, pp. 271–290.

[14] E. A. Hermanson, "A brief overview of Bible translation in South Africa," *Acta Theologica, Supplementum 2*, vol. 22, no. 1, pp. 6–18, 2002. [Online]. Available: http://dx.doi.org/10.4314/actat.v22i1.5451

[15] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy, "Parallel corpora for medium density languages," in *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*, ser. 'Current Issues in Linguistic Theory', N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, Eds. John Benjamins, 2007, vol. 292, pp. 247–258.

[16] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

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

[19] K. Heafield, "KenLM: faster and smaller language model queries," in *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 187–197. [Online]. Available: http://kheafield.com/professional/avenue/kenlm.pdf

[20] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696. [Online]. Available: http://kheafield.com/professional/edinburgh/estimate_paper.pdf

[21] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, Detroit, Michigan, May 1995, pp. 181–184.

[22] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999. [Online]. Available: http://dx.doi.org/10.1006/csla.1999.0128

[23] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, ser. SETQA-NLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 49–57. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622110.1622119

[24] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. [Online]. Available: http://dx.doi.org/10.3115/1073083.1073135

[26] J. Olive, "Global Autonomous Language Exploitation (GALE)," DARPA/IPTO Proposer Information Pamphlet, 2005.

[27] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of Association for Machine Translation in the Americas*. Citeseer, 2006, pp. 223–231.

[28] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 160–167. [Online]. Available: http://dx.doi.org/10.3115/1075096.1075117