

Exploring the SAWA corpus: collection and deployment of a parallel corpus English—Swahili

Guy De Pauw · Peter Waiganjo Wagacha · Gilles-Maurice de Schryver

Published online: 19 July 2011
© Springer Science+Business Media B.V. 2011

Abstract Research in machine translation and corpus annotation has greatly benefited from the increasing availability of word-aligned parallel corpora. This paper presents ongoing research on the development and application of the SAWA corpus, a two-million-word parallel corpus English—Swahili. We describe the data collection phase and zero in on the difficulties of finding appropriate and easily accessible data for this language pair. In the data annotation phase, the corpus was semi-automatically sentence and word-aligned and morphosyntactic information was added to both the English and Swahili portion of the corpus. The annotated parallel corpus allows us to investigate two possible uses. We describe experiments with the projection of part-of-speech tagging annotation from English onto Swahili, as well as the development of a basic statistical machine translation system for this language pair, using the parallel corpus and a consolidated database of existing English—Swahili translation dictionaries. We particularly focus on the difficulties

The research presented in this paper was made possible through the support of the VLIR-IUC-UON program and was partly funded by the SAWA BOF UA-2007 project. The first author is funded as a Postdoctoral Fellow of the Research Foundation—Flanders (FWO).

G. De Pauw (✉)
CLiPS, Department of Linguistics, University of Antwerp, Antwerp, Belgium
e-mail: guy.depauw@ua.ac.be

G. De Pauw · P. W. Wagacha
School of Computing and Informatics, University of Nairobi, Nairobi, Kenya

P. W. Wagacha
e-mail: waiganjo@uonbi.ac.ke

G.-M. de Schryver
Department of African Languages and Cultures, Ghent University, Ghent, Belgium
e-mail: gillesmaurice.deschryver@ugent.be

G.-M. de Schryver
Xhosa Department, University of the Western Cape, Cape Town, South Africa

of translating English into the morphologically more complex Bantu language of Swahili.

Keywords Parallel corpus · Swahili · English · Machine translation · Projection of annotation · African language technology

1 Introduction

Typical language technology applications such as information extraction, spell checking and machine translation can provide an invaluable—but all too often ignored—impetus in bridging the digital divide between the Western world and developing countries. In Africa, quite a few localization efforts are currently underway that allow improved ICT access in local African languages (e.g. ANLoc). Vernacular content is increasingly being published on the Internet and the need for robust language technology applications that can process this data is obviously high.

For a language like Swahili, spoken by more than fifty million people in East and Central Africa, digital resources have become increasingly important in everyday life, both in urban and rural areas, thanks to the growing number of web-enabled mobile phone users in the language area and increased bandwidth, courtesy of broadband and the terrestrial and undersea optical fiber cables. The prominence of regional economic blocks such as the East African Market and the growing popularity of the expanded media in the region further underline the need for African language technology tools.

Most research efforts in the field of natural language processing for African languages are rooted in the rule-based paradigm. Language technology components in this sense are usually straight implementations of insights derived from grammarians. Albeit often highly accurate and intricately designed, the rule-based approach has the distinct disadvantage of being language-dependent and costly to develop, as it typically involves a lot of expert manual effort.

Furthermore, many of these systems are decidedly *competence*-based. The systems are often tweaked and tuned towards a small set of ideal sample words or sentences, ignoring the fact that real-world language technology applications have to be principally able to handle the *performance* aspect of language. Many researchers in the field are growing weary of publications that ignore quantitative evaluation on real-world data or that report incredulously high accuracy scores, excused by the erroneously perceived *regularity* of African languages.

In a linguistically diverse and increasingly computerized continent such as Africa, the need for a more empirically motivated and less *resource-heavy* approach to language technology is high. The data-driven, corpus-based approaches described in this paper, establish such an alternative, so far not yet extensively investigated for African languages. The main advantage of this approach is its language independence: all that is needed is (linguistically annotated) data, which is cheaper to compile than it is to design a rule-based system. Given this data, existing state-of-the-art algorithms and resources can easily be applied to quickly develop robust language applications and tools.

Most African languages are resource-scarce, meaning that digital text resources are few. An increasing number of publications however are showing that carefully selected corpus-based procedures can indeed bootstrap language technology for languages such as Amharic (Gambäck et al. 2009), Northern Sotho (de Schryver and De Pauw 2007; Faaß et al. 2009), Swahili (De Pauw et al. 2006; De Pauw and de Schryver 2008; Steinberger et al. this volume), Tswana (Groenewald 2009) and even very resource-scarce African languages (De Pauw and Wagacha 2007; De Pauw et al. 2007; Badenhurst et al., this volume; Scannell, this volume). This paper continues this novel and promising new trend in African language technology research, by presenting the development and deployment of the SAWA corpus, a two-million-word parallel corpus English—Swahili.

This paper starts off by outlining the data collection and annotation efforts needed to compile the SAWA corpus (Sect. 2). We particularly zero in on the difficulties of finding appropriate and easily accessible data for this language pair and introduce a novel, supervised sentence-alignment method. In Sect. 3 we explore different approaches to word-alignment for the language pair English—Swahili. The resulting multi-tiered annotated corpus allows us to investigate two possible practical uses for the data. Section 4 describes experiments with the projection of part-of-speech tagging information from English onto Swahili, while Sect. 5 presents a first, basic bidirectional statistical machine translation system based on the SAWA corpus data. We conclude with a discussion of the current results and limitations and provide pointers for future research in Sect. 6.

2 Data collection and annotation

While digital data in Swahili is abundantly available on the Internet, sourcing useful bilingual English—Swahili data is far from trivial. Even countries that have both English and Swahili as their official languages, such as Tanzania, Kenya and Uganda, do not tend to translate and/or publicly publish all government documents bilingually. While non-parallel, i.e. *comparable*, corpora are now increasingly being researched in the context of machine translation, we deemed it appropriate to try and source faithfully translated material in the initial stages of the SAWA corpus development. Restricting ourselves to purely parallel data enables the straightforward deployment of standard statistical machine translation tools (Sect. 5) and allows us to investigate the possibility of projection of annotation (Sect. 4).

Table 1 gives an overview of the data currently available in the SAWA corpus. It consists of a reasonable amount of data (roughly two and a half million tokens), although this is not comparable to the resources available for Indo-European language pairs, such as the Hansard corpus (Roukos et al. 1997). Although religious material constitutes three quarters of the corpus at this point, we attempted to get data from other domains as well, such as economic and political documents.

We found digitally available Swahili versions of the Bible and the Quran for which we sourced the English counterparts. This is not a trivial task when, as in the case of the Swahili documents, the exact source of the translation is not provided. By carefully examining subtle differences in the English versions, we were however

Table 1 Overview of the sentence-aligned data in the SAWA corpus

	Sentences	English tokens	Swahili tokens
Bible	52.4k	944.9k	751.2k
Quran	14.2k	177.1k	137.7k
Politics	3.8k	69.2k	62.5k
Kamusi.org	5.7k	41.6k	29.8k
Movie subtitles	11.2k	70.0k	58.1k
Local translator	1.3k	24.9k	24.2k
Investment reports	6.4k	137.8k	135.9k
Full corpus total	73.7k	1.463M	1.201M

Scores in bold indicate manually sentence-aligned portions

able to track down the most likely candidate. While religious material has a specific register and may not constitute ideal training material for an open-ended machine translation system, it does have the advantage of being inherently aligned on the verse level, facilitating further sentence-alignment. The political portion of the corpus consists of the UN Declaration of Human Rights and the 2009 Draft Constitution of Kenya.

The downloadable version of the on-line dictionary English—Swahili (Benjamin 2011) contains individual example sentences associated with the dictionary entries. These can be extracted and used as parallel data in the SAWA corpus. Since at a later point, we also wish to study the specific linguistic aspects of spoken language, we opted to have some movie subtitles manually translated. Movie subtitles can be easily downloaded from OpenSubtitles.org and while the language is compressed to fit on screen and constitutes scripted language, this data nevertheless provides a reasonable approximation of spoken language. Another advantage of working with subtitles is that it is inherently sentence-aligned, thanks to the technical time-coding information. It also opens up possibilities for machine translation systems for other language pairs, since a commercial feature film typically has subtitles available for a large number of other languages as well.

We also obtained a substantial amount of data from a local Kenyan translator, non-governmental organization leaflets on social welfare. Finally, we also included Kenyan investment reports. These are yearly reports from local companies and are presented in both English and Swahili. A major difficulty was extracting the data from these documents. The company reports are presented in colorful brochures in PDF format, meaning automatic text exports require significant manual post-processing and paragraph alignment. They nevertheless provide a valuable resource, since they come from a fairly specific domain and are a good sample of the type of text the projected machine translation system may need to process in a practical setting.

All of the data in the corpus was tokenized, which involves automatically cleaning up the texts, conversion to UTF-8 and sentence boundary detection. Each text in the SAWA corpus was subsequently automatically part-of-speech tagged and lemmatized. For Swahili we used the systems described in De Pauw et al. (2006) and De Pauw and de Schryver (2008). For the English data we used the TreeTagger (Schmid 1994). The goal is to provide a multi-tier representation of the sentences in

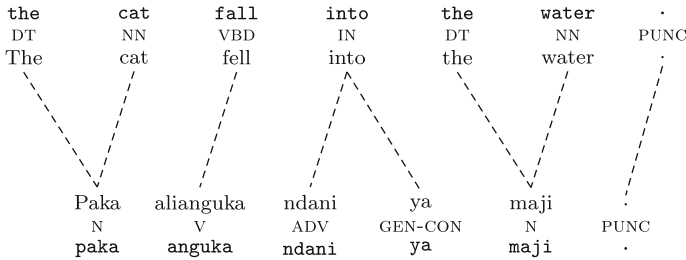


Fig. 1 Multi-tier annotation in the sawa corpus

both languages, as illustrated in Fig. 1, which includes lemmatization, part-of-speech tags and word-alignment (see Sect. 3).

The next annotation step involved sentence-alignment of the data, during which we establish an unambiguous mapping between the sentences in the source text and the sentences in the target text. We did this semi-automatically, using the Microsoft Bilingual Sentence Aligner (Moore 2002) as a pre-processing step. The output of the sentence-alignment was subsequently manually corrected, with the exception of the Old Testament data, which was processed fully automatically (cf. infra). We found that only about 5% of the sentences in the data needed to be manually corrected. Most errors can be attributed to sentences that were not present in English, i.e. instances where the translator decided to add an extra clarifying sentence to the direct translation from English. Where possible, such sentences were removed from the corpus.

Having a data set of manually sentence-aligned words provides us with the option to develop a sentence-alignment method that can *learn* from examples. Similar work was described in Zhao et al. (2003); Ceașu et al. (2006), but we were not able to get hold of either system. We therefore decided to develop such a tool from scratch, using a Maximum Entropy Learning method as the backbone.

To train the sentence-alignment method, we extract each manually aligned pair of sentences in the sawa corpus, as well as n sentences before and after the aligned pair. This is illustrated in Fig. 2 for $n = 1$. Each of the $(4n + 1)$ pairs is represented as a bilingual bag of words (EW and SW), part-of-speech tags (ET and ST) and (non-function word) lemmas (EL and SL). Other *tell-tale* alignment signs, such as verse indications in religious text, are removed.

Two example training instances are presented in Fig. 3. Negative examples of alignments (dashed lines in Fig. 2) receive class “0”, while positive examples are marked as class “1”. In addition a unique index is appended to each class (indicated as $-n$ in Fig. 3). A Maxent classifier (Le 2004) is subsequently trained on this data,

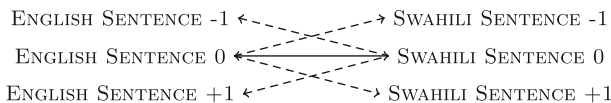


Fig. 2 Sentence-alignment candidates

1-1 EW=The EW=cat EW=fell EW=into EW=the EW=water EW=. ET=DT ET=NN ET=VBD
 ET=IN ET=DT ET=NN ET=Punc EL=cat EL=fall EL=water SW=Paka SW=alianguka
 SW=ndani SW=ya SW=maji ST=N ST=V ST=ADV ST=GEN-CON ST=N ST=Punc SL=paka
 SL=anguka SL=maji
 0-2 EW=The EW=cat EW=fell EW=into EW=the EW=water EW=. ET=DT ET=NN ET=VBD
 ET=IN ET=DT ET=NN ET=Punc EL=cat EL=fall EL=water SW=Na SW=mbwa SW=pia SW=;
 ST=CC ST=N ST=ADV ST=Punc SL=mbwa

Fig. 3 Instances for Maximum Entropy Sentence Aligner

which establishes a maximum entropy model that uniquely identifies the sentence pairs in the training data.

During classification of a new, previously unseen sentence pair, the model outputs the probabilities of all of the classes in the training model. The presence of individual features that match cross-lingually, trigger a higher probability for positive “1-*n*”-type classes, whereas sentence pairs with mismatching features skew classification towards the more common negative “0-*n*”-type classes. Essentially the probability of a given class expresses the similarity of the current sentence pair to the associated sentence pair in the training data. As such it functions not unlike a kNN type classifier with *k* equal to the number of training instances. We chose the maximum entropy classifier due to its ability to effectively handle both large sets of classes as well as sparse vectors.

The best alignment pattern for a particular paragraph is then established by maximizing the overall probability of sentence-alignment classification decisions through dynamic programming. While computationally heavy and rather slow¹, sentence-alignment accuracy was increased to 98.4% in a ten-fold cross validation experiment using an *n* value of 4. We used this method to perform automatic sentence-alignment of the Old Testament data, where the paragraph level is equal to the verse level.

While not essential for further processing, we also created a small manually word-aligned evaluation set. This task can be performed automatically using standard tools (Sect. 3), but it is useful to have a gold-standard reference against which we can evaluate the automated method. Monitoring the accuracy of the automatic word-alignment method against the reference material, allows us to tweak parameters to arrive at the optimal settings for this language pair.

We used the UMIACS word-alignment interface (Hwa and Madnani 2004) for this purpose and asked the annotators to link the words between the two sentences, as illustrated in Fig. 1. Given the linguistic differences between English and Swahili, this is by no means a trivial task. Particularly the agglutinating nature of Swahili morphology means that there is a lot of convergence from (multiple) words in English to words in Swahili (also see Sect. 3). This alignment was done on some of the manual translations of movie subtitles, giving us a small gold-standard word-alignment reference of about 5,000 words. Each annotator’s work was cross-checked by another annotator to improve correctness and overall consistency. In the next section, we will explore automatic approaches to word-alignment.

¹ On an Intel Xeon 2.4Ghz system with 8Gb RAM, training took about 36h. The classification phase fares better, taking only a few seconds per paragraph.

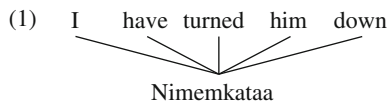
3 Alignment experiments

There are a number of packages available to process parallel corpora. For word-alignment, the state-of-the-art method is GIZA++ (Och and Ney 2003), which implements among others the word-alignment methods IBM1 to IBM5 and HMM. While this method is particularly well suited to handle closely related languages, it is interesting to see the performance of the default approach for the distant language pair English—Swahili.

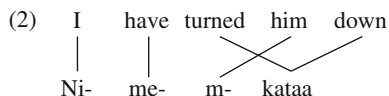
We performed exploratory experiments with different alignment models and found that using factored data (i.e. including part-of-speech tags and lemma information) yielded the highest accuracy. We evaluate the approach by looking at the word-alignments proposed by GIZA++ and comparing them to the manually word-aligned section of the SAWA corpus. Following the recommendations of Fraser and Marcu (2007), we quantify the evaluation by calculating precision and recall and their harmonic mean, the *F*-score. Precision expresses how many links between words are correct, divided by the total number of links suggested by GIZA++. Recall is calculated by dividing the number of correct links, by the total number of links in the manual annotation.

The underwhelming results presented in the first row of Table 2 can be attributed to the Indo-European bias of the GIZA++ approach. It is primarily used to align related languages on the word level. For our language pair, it is clear that extra linguistic data sources and a more elaborate exploration of the experimental parameters of GIZA++ is required.

The main problem in training a GIZA++ model for the language pair English—Swahili is the strong agglutinating nature of the latter. Alignment patterns such as the one in Fig. 1 are not impossible to retrieve, but no corpus is exhaustive enough to provide enough linguistic evidence to unearth strongly converging alignment patterns, such as the one in Example 1.



Morphologically deconstructing the Swahili word however can greatly relieve the sparse data problem for this task:



The isolated Swahili morphemes can more easily be linked to their English counterparts, since there will be more linguistic evidence in the parallel corpus, linking for example *ni* to *I* and *m* to *him*. To perform this kind of morphological segmentation, we used the system that provides lemmatization information in the SAWA corpus (De Pauw and de Schryver 2008). By identifying the base form of the word, we can also distinguish a prefix group and a suffix group. Since affixes in Swahili are monosyllabic, we can proceed by syllabifying these two groups to arrive

Table 2 Precision, recall and *F*-score for the word-alignment task using GIZA++

	Precision (%)	Recall (%)	<i>F</i> ($\beta = 1$) (%)
Word model	39.4	44.5	41.8
Morpheme model	50.2	64.5	55.8
Morpheme model + dictionary	66.5	72.6	69.4

at a complete morphological segmentation of the word form. Introducing such morphological features in statistical machine translation has previously been attempted with varying degrees of success for other morphologically complex languages as well (Bojar 2007; Minkov et al. 2007; Ramanathan et al. 2008; Szymne et al. 2008; Oflazer 2008; Diaz de Ilarraza et al. 2009).

We have no morphologically aligned gold standard data available, so evaluation of the morpheme-based approach needs to be done in a roundabout way. We first morphologically decompose the Swahili data and run GIZA++ again. Next we recompile the Swahili words from the morphemes and group the word-alignment links accordingly. Incompatible linkages are removed and simple majority voting resolves ambiguous alignment patterns. The updated scores are presented in the second row of Table 2 and show that this type of processing is highly beneficial for this language pair.

We also have at our disposal a consolidated database of four electronic English—Swahili translation dictionaries (De Pauw et al. 2009a), containing 21,000 lemmas. By introducing this information source in the morpheme-based alignment process as well, we are able to substantially improve on the word-alignment scores (third row Table 2).

While morpheme-based word-alignment certainly improves on the scores of the word-based system, we need to be aware of the difficulty that this morphological pre-processing step will introduce in the decoding phase, necessitating the introduction of a language model that not only works on the word level, but also on the level of the morpheme, as well as a morphological generation component for English → Swahili translation. Although the morpheme-based language model can provide useful additional linguistic information to the machine translation system, there is no quick fix for the latter problem. For the purpose of projection of annotation, this is however not an issue, since this type of processing typically occurs on the word level.

4 Projection of annotation

While machine translation constitutes the most straightforward application of a parallel corpus, projection of annotation has recently become an interesting alternative use of this type of resource. As previously mentioned, most, if not all African languages are resource-scarce: annotated data is not only unavailable, but commercial interest to develop these resources is limited. Unsupervised approaches

can be used to bootstrap annotation of a resource-scarce language (De Pauw and Wagacha 2007; De Pauw et al. 2007) by automatically finding linguistic patterns in large amounts of raw text.

Projection of annotation attempts to achieve the same goal, but through the use of a word-aligned parallel corpus. These techniques try to transport the annotation of a well-resourced source language, such as English, to texts in a target language. The direct correspondence assumption coined in Hwa et al. (2002), hypothesizes that words that are aligned between source and target language, must share linguistic features as well. It therefore allows for the annotation of the words in the source language to be projected onto the text in the target language. The following general principle holds: the more closely the source and target language are related, the more accurate this projection can be performed. Even though lexical and structural differences between languages prevent a simple one-to-one mapping, this type of knowledge transfer is often able to generate a fairly well-directed annotation of the target language (De Pauw et al. 2010).

To investigate the applicability of this technique to resource-scarce languages, we performed an experiment to see how well the projection of part-of-speech tag information is handled from English to Swahili. We word-aligned the SAWA corpus again without using factored data, as this is typically not available to resource-scarce languages. We then project the English tags along the word-alignment links onto the Swahili words. Since the tag sets are different, we also needed to design a conversion table that maps the English part-of-speech tags to their Swahili counterparts, finally allowing us to evaluate the result of the projection against the part-of-speech tags of the silver standard (Fig. 1), i.e. the tags provided by the Swahili memory-based tagger (De Pauw et al. 2006).

Table 3 outlines the result of this experiment. The first row shows the performance of the projection on the manually word-aligned gold standard set. 90.1% tagging accuracy is far below that of the data-driven Swahili part-of-speech tagger (De Pauw et al. (2006) reports over 98% tagging accuracy), but gives a good indication of how well projection works when word-alignment is optimal.

The fully automatic projection of part-of-speech tags, which projects tags through automatically induced word-alignment links, scores almost 75%. Error analysis showed that most of the tagging errors were made on Swahili words that were not aligned to an English counterpart and had therefore not received any part-of-speech tag. Many of those constitute function words, that can be easily tagged using a table look-up post-processing technique. Furthermore, De Pauw et al. (2010) show that tagging accuracy and overall coverage can be further increased by training a morphologically aware machine learning classifier on top of the projected annotation.

Table 3 Projection of part-of-speech tags from English onto Swahili

	Tagging accuracy (%)
Gold-standard data	90.1
Automatically aligned data	74.8

While 70% may seem like a rather modest result for a part-of-speech tagger, it is important to point out that this result was obtained without extra linguistic information sources for Swahili, purely on the basis of existing annotation tools for English and the automatically word-aligned parallel corpus data. The fact that these languages are linguistically very different, further underlines the robustness of the projection technique. This *knowledge light* approach to corpus annotation can thus be considered as a promising technique to provide annotated data for resource-scarce languages.

5 Machine translation

The most straightforward and practical application of a parallel corpus is undoubtedly as a resource to build a statistical machine translation (SMT) system. In this section we outline a preliminary SMT experiment using the resources that the SAWA corpus has to offer. Apart from a very early contribution (Woodhouse 1968), there are no published papers on Swahili machine translation, although an earlier version of the SAWA corpus was described in De Pauw et al. (2009). In the summer of 2009, Google released a Swahili version of their on-line machine translation system, which clearly uses many of the same resources described in this publication. In this section, we will compare the output of Google's system to that of our SMT approach.

As our decoder we used the standard MOSES package (Koehn et al. 2007), which takes the alignment patterns generated by GIZA++ to construct a (possibly phrase-based) machine translation system. To construct an n-gram language model with the SRILM toolkit (Stolcke 2002), we used the twenty-million-word *TshwaneDJe Kiswahili Internet Corpus* (de Schryver and Joffe 2009), which contains a similar spread in document types as the SAWA corpus. For English we used the Gigaword corpus (Graff 2003) as the basis for our language model. In both cases we opted for a simple trigram language model.

We did not perform extensive parameter tweaking and tuning on either the SMT or language model side, mostly restricting ourselves to the default settings. Therefore the experimental results presented in this section still leave considerable room for improvement.

From each section of the SAWA corpus we randomly extracted a 10% test set, which was held out during training of the SMT system. We unfortunately lack the resources to perform extensive human evaluation, but as an alternative, we can evaluate the automatically generated translations by comparing them to the original, reference translations. The quality of the output can thus be quantified using the standard machine translation evaluation measures BLEU, NIST, WER (Word Error Rate) and PER (Position-Independent Word Error Rate). Note that this experimental setup puts the SAWA/MOSES at an inherent disadvantage, because we can only guarantee that the test set constitutes unseen data for *our* system, while the same data may have been used by Google's system to train their machine translation system.

The experimental results can be found in Tables 4 and 5. Interestingly, the results vary according to the direction of the translation. For English → Swahili translation

Table 4 Quantitative evaluation of machine translation task: English—Swahili

	GOOGLE				SAWA/MOSES			
	BLEU	NIST	WER	PER	BLEU	NIST	WER	PER
Bible	0.16	4.62	67.44	55.55	0.15	4.24	71.01	56.98
Quran	0.15	4.58	68.41	55.19	0.15	4.18	71.25	57.88
Politics	0.15	4.55	68.31	55.52	0.14	4.24	71.56	57.75
Kamusi.org	0.13	4.39	69.14	56.03	0.10	4.34	71.21	58.03
Subtitles	0.10	4.23	72.10	58.14	0.10	4.22	73.51	60.08
Translator	0.10	4.19	72.31	58.41	0.10	4.14	72.13	59.05
Investment	0.12	4.44	72.14	56.95	0.14	4.21	71.53	57.50
Total	0.15	4.56	68.45	55.47	0.14	4.23	71.30	57.41

Table 5 Quantitative evaluation of machine translation task: Swahili—English

	GOOGLE				SAWA/MOSES			
	BLEU	NIST	WER	PER	BLEU	NIST	WER	PER
Bible	0.19	4.65	70.92	57.13	0.23	4.81	66.50	52.96
Quran	0.18	4.59	71.90	57.92	0.23	4.70	67.01	52.87
Politics	0.16	4.42	72.30	60.99	0.21	4.24	68.31	53.14
Kamusi.org	0.17	4.62	72.52	60.42	0.23	4.60	69.52	53.14
Subtitles	0.14	4.12	73.99	62.14	0.22	4.72	70.19	55.07
Translator	0.14	4.14	73.31	61.22	0.21	4.40	69.96	55.20
Investment	0.17	4.18	72.01	60.86	0.23	4.79	66.78	53.05
Total	0.18	4.54	71.92	58.57	0.23	4.74	67.04	53.13

the SAWA/MOSES system underperforms compared to Google Translate's system. This may be partly attributed to the experimental setup, but is also likely due to extra linguistic sources that the latter system uses on the target language side, such as morphological generation (Denis Gikunda (Google Inc. East Africa), personal communication). For Swahili → English translation, our system fares better, not hampered by the morphological generation issues of the target language. On all of the evaluation metrics and for all of the subsections of the SAWA corpus, the SAWA/MOSES approach significantly outperforms the Google system.

When we look at the experimental results in a bit more detail, some general tendencies appear. Religious material in general is translated more accurately by both systems. For the SAWA/MOSES system this is no surprise, as three quarters of the training material is indeed in this particular register. It is encouraging however, that the religious bias of the SAWA corpus does not seem to yield a huge drop in accuracy on other document types. Even the pseudo-spoken language of the subtitles documents is handled fairly well. When inspecting the output of the SAWA/MOSES system, the most significant problems at this point seem to be on the level of morphological processing, rather than being intrinsic lexical problems due to the dominant register in the training data.

6 Discussion

In this paper we presented the development and deployment of a parallel corpus English—Swahili. The current version of the *SAWA* corpus has more than two million words, part-of-speech tagged, lemmatized and sentence and word-aligned. To our knowledge, this is the only such resource available for a sub-Saharan African language. As new resources, such as legal documents, investments reports and translated Wikipedia pages are increasingly being made available, we are confident that the *SAWA* corpus will significantly grow in size in years to come. Furthermore, advances in parallel web mining (Resnik and Smith 2003) will further contribute to the range of data in this parallel corpus, reducing its religious bias.

We introduced projection of annotation as one of the possible uses of the *SAWA* corpus. We presented a proof-of-the-principle experiment that showed that annotation of a target language can be bootstrapped, relying solely on word-alignment patterns and the annotation modules of a resource-rich source language. This is a particularly promising result for the annotation of other resource-scarce African languages that have parallel data, typically the Bible and the Quran, at their disposal. We will also investigate the possibility of projecting dependency analyses from English onto Swahili, allowing us to bootstrap the development of a dependency parser for the latter language.

Furthermore, this paper presented the first published experimental results of a statistical machine translation system for a Bantu language. We are confident that the quality of the translations can be significantly improved by performing an extensive exploration of algorithmic parameters for both *GIZA++* and *MOSES*, as well as through the inclusion of more data. We will particularly need to focus on the morphological generation component for translation into Swahili, as this is currently the primary bottleneck for the *SAWA* system.

We will also explore other machine translation decoders, as well as alternative approaches to word and morpheme alignment, including an adaptation of the maximum entropy approach used to perform sentence-alignment in the *SAWA* corpus. Significant advances may also be made by exploiting the word-reordering capabilities of the *MOSES* package. Word-reordering attempts to pre-process the source data to mimic the word order of the target language before decoding. This is particularly useful for distant language pairs that have significant differences in word order and it is clear that the language pair English—Swahili can benefit from such an approach as well.

Finally, we will also look into the use of comparable corpora, i.e. bilingual texts that are not straight translations, but deal with the same subject matter. These have been found to work well as additional material within a parallel corpus and may further help improve the development of a robust, open-ended and bidirectional machine translation system for the language pair English—Swahili.

Demo and data: A demonstration machine translation system and non-copyrighted parts of the *SAWA* corpus will be made publicly available through AfLaT.org.

Acknowledgments We are very grateful for the insightful and useful comments from the reviewers, which helped shape the final version of this paper. We are also greatly indebted to Dr. James Omboza Zaja for contributing some of his translated data, to Mahmoud Shokrollahi-Far for his advice on the Quran and to Anne Kimani, Chris Wangai Njoka and Naomi Maajabu for their tireless annotation efforts.

References

- ANLoc. (2011). *The African network for localization*. Available at: <http://www.africanlocalisation.net>. Accessed: 10 June 2011.
- Benjamin, M. (2011). *The Kamusi project*. Available at: <http://www.kamusiproject.org>. Accessed: 10 June 2011.
- Bojar, O. (2007). English-to-Czech factored machine translation. In *Proceedings of the second workshop on statistical machine translation* (pp. 232–239). Morristown, USA: Association for Computational Linguistics.
- Ceașu, A., Ștefănescu, D., & Tufiș, D. (2006). Acquis communautaire sentence alignment using support vector machines. In *Proceedings of the 5th international conference on language resources and evaluation* (pp. 2134–2137). Genoa, Italy: ELRA—European Language Resources Association.
- De Pauw, G., & Wagacha, P. (2007). Bootstrapping morphological analysis of Gikūyū using unsupervised maximum entropy learning. In *Proceedings of the eighth INTERSPEECH conference*. Antwerp, Belgium: International Speech Communication Association.
- De Pauw, G., & de Schryver, G.-M. (2008). Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos*, 18, 303–318.
- De Pauw, G., de Schryver, G.-M., & Wagacha, P. (2006). Data-driven part-of-speech tagging of Kiswahili. In P. Sojka, I. Kopeček, & K. Pala (Eds.), *Proceedings of text, speech and dialogue, ninth international conference* (pp. 197–204). Berlin, Germany: Springer.
- De Pauw, G., Wagacha, P., & Abade, D. (2007). Unsupervised induction of Dholuo word classes using maximum entropy learning. In K. Getao & E. Omwenga (Eds.), *Proceedings of the first international computer science and ICT conference* (pp. 139–143). Nairobi, Kenya: University of Nairobi.
- De Pauw, G., de Schryver, G.-M., & Wagacha, P. W. (2009a). A corpus-based survey of four electronic Swahili–English bilingual dictionaries. *Lexikos*, 19, 340–352.
- De Pauw, G., Wagacha, P., & de Schryver, G.-M. (2009b). The SAWA corpus: A parallel corpus English–Swahili. In G. De Pauw, G.-M. de Schryver, & L. Levin (Eds.), *Proceedings of the first workshop on language technologies for African languages (AfLaT 2009)* (pp. 9–16). Athens, Greece: Association for Computational Linguistics.
- De Pauw, G., Maajabu, N., & Wagacha, P. (2010). A knowledge-light approach to Luo machine translation and part-of-speech tagging. In G. De Pauw, H. Groenewald, & G.-M. de Schryver (Eds.), *Proceedings of the second workshop on African language technology (AfLaT 2010)* (pp. 15–20). Valletta, Malta: European Language Resources Association (ELRA).
- de Schryver, G.-M., & De Pauw, G. (2007). Dictionary writing system (DWS) + corpus query package (CQP): The case of TshwaneLex. *Lexikos*, 17, 226–246.
- de Schryver, G.-M., & Joffe, D. (2009). *TshwaneDJe Kiswahili internet corpus*. Pretoria, South Africa: TshwaneDJe HLT.
- Diaz de Ilaraza, A., Labaka, G., & Sarasola, K. (2009). Relevance of different segmentation options on Spanish-Basque SMT. In L. Mrquez & H. Somers (Eds.), *Proceedings of the 13th annual conference of the European association for machine translation* (pp. 74–80). Barcelona, Spain: European Association for Machine Translation
- Faaß, G., Heid, U., Taljard, E., & Prinsloo, D. J. (2009). Part-of-speech tagging of Northern Sotho: Disambiguating polysemous function words. In G. De Pauw, G.-M. de Schryver, & L. Levin (Eds.), *Proceedings of the first workshop on language technologies for African languages (AfLaT 2009)* (pp. 38–45). Athens, Greece: Association for Computational Linguistics.
- Fraser, A., & Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3), 293–303.
- Gambäck, B., Olsson, F., Argaw, A. A., & Asker, L. (2009). Methods for Amharic part-of-speech tagging. In G. De Pauw, G.-M. de Schryver, & L. Levin (Eds.), *Proceedings of the first workshop on language technologies for African Languages (AfLaT 2009)* (pp. 104–111). Athens, Greece: Association for Computational Linguistics.

- Graff, D. (2003). *English Gigaword*. [Online]. Available: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05>. Accessed: 10 June 2011.
- Groenewald, H. J. (2009). Using technology transfer to advance automatic lemmatisation for Setswana. In G. De Pauw, G.-M. de Schryver & L. Levin (Eds.), *Proceedings of the first workshop on language technologies for African languages(AfLaT 2009)* (pp. 32–37). Athens, Greece: Association for Computational Linguistics.
- Hwa, R., & Madnani, N. (2004). *The UMIACS Word alignment interface*. Available at: <http://www.umiacs.umd.edu/~nmadnani/alignment>. Accessed: 10 June 2011.
- Hwa, R., Resnik, P., Weinberg, A., & Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 392–399). Philadelphia, USA: Association for Computational Linguistics.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). MOSES: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics.
- Le, Z. (2004). *Maximum entropy modeling toolkit for Python and C++*. Available at: http://home pages.inf.ed.ac.uk/s0450736/maxent_toolkit.html. Accessed: 10 June 2011.
- Minkov, E., Toutanova, K., & Suzuki, H. (2007). Generating complex morphology for machine translation. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 128–135). Prague, Czech Republic: Association for Computational Linguistics.
- Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. In S. Richardson (Ed.), *Proceedings of the fifth conference of the association for machine translation in the Americas on machine translation: From research to real users* (pp. 135–144). Berlin, Germany: Springer.
- Och, F., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51.
- Oflazer, K. (2008). Statistical machine translation into a morphologically complex language. In *Computational linguistics and intelligent text processing* (pp. 376–388). Berlin, Germany: Springer.
- OpenSubtitles.org. (2011). *OpenSubtitles*. Available at <http://www.opensubtitles.org>. Accessed: 10 June 2011.
- Ramanathan, A., Hegde, J., Shah, R., Bhattacharya, P., & Sasikumar, M. (2008). Simple syntactic and morphological processing can help English–Hindi statistical machine translation. In *Third international joint conference on natural language processing* (pp. 513–520). Hyderabad, India: Asian Federation of Natural Language Processing.
- Resnik, P., & Smith, N. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(1), 349–380.
- Roukos, S., Graff, D., & Melamed, D. (1997). *Hansard French/English*. [Online]. Available at: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20>. Accessed: 10 June 2011.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. Jones (Ed.), *Proceedings of the international conference on new methods in language processing* (pp. 44–49). Manchester, UK: UMIST.
- Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In J. Hansen & B. Pellom (Eds.), *Proceedings of the international conference on spoken language processing* (pp. 901–904). Denver, USA: International Speech Communication Association.
- Stymne, S., Holmqvist, M., & Ahrenberg, L. (2008). Effects of morphological analysis in translation between German and English. In *Proceedings of the third workshop on statistical machine translation* (pp. 135–138). Columbus, USA: Association for Computational Linguistics.
- Woodhouse, D. (1968). A note on the translation of Swahili into English. *Mechanical Translation and Computational Linguistics*, 11, 75–77.
- Zhao, B., Zechner, K., Vogel, S., & Waibel, A. (2003). Efficient optimization for bilingual sentence alignment based on linear regression. In *Proceedings of the HLT-NAACL 2003 workshop on building and using parallel texts: Data driven machine translation and beyond* (Vol. 3, pp. 81–87). Morristown, USA: Association for Computational Linguistics.