The SAWA Corpus: a Parallel Corpus English - Swahili

Guy De Pauw
CNTS - Language Technology Group, University of Antwerp, Belgium
School of Computing and Informatics, University of Nairobi, Kenya
guy.depauw@ua.ac.be

Peter Waiganjo Wagacha
School of Computing and Informatics, University of Nairobi, Kenya
waiganjo@uonbi.ac.ke

Gilles-Maurice de Schryver
African Languages and Cultures, Ghent University, Belgium
Xhosa Department, University of the Western Cape, South Africa
gillesmaurice.deschryver@ugent.be

Abstract

Research in data-driven methods for Machine Translation has greatly benefited from the increasing availability of parallel corpora. Processing the same text in two different languages yields useful information on how words and phrases are translated from a source language into a target language. To investigate this, a parallel corpus is typically aligned by linking linguistic tokens in the source language to the corresponding units in the target language. An aligned parallel corpus therefore facilitates the automatic development of a machine translation system and can also bootstrap annotation through projection. In this paper, we describe data collection and annotation efforts and preliminary experimental results with a parallel corpus English - Swahili.

1 Introduction

Language technology applications such as machine translation can provide an invaluable, but all too often ignored, impetus in bridging the digital divide between the Western world and Africa. Quite a few localization efforts are currently underway that improve ICT access in local African languages. Vernacular content is now increasingly being published on the Internet, and the need for robust language technology applications that can process this data is high.

For a language like Swahili, digital resources have become increasingly important in everyday life, both in urban and rural areas, particularly thanks to the increasing number of web-enabled mobile phone users in the language area. But most research efforts in the field of natural language processing (NLP) for African languages are still firmly rooted in the rule-based paradigm. Language technology components in this sense are usually straight implementations of insights derived from grammarians. While the rule-based approach definitely has its merits, particularly in terms of design transparency, it has the distinct disadvantage of being highly 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 quite rightly 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 empirical approach to language technology is high. The data-driven, corpus-based approach described in this paper establishes such an alternative, so far not yet extensively investigated for African languages. The main advantage of this

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 9–16,
Athens, Greece, 31 March 2009. ©2009 Association for Computational Linguistics

approach is its language independence: all that is needed is (linguistically annotated) language data, which is fairly cheap to compile. Given this data, existing state-of-the-art algorithms and resources can consequently be re-used to quickly develop robust language applications and tools.

Most African languages are however resource-scarce, meaning that digital text resources are few. An increasing number of publications however are showing that carefully selected procedures can indeed bootstrap language technology for Swahili (De Pauw et al., 2006; De Pauw and de Schryver, 2008), Northern Sotho (de Schryver and De Pauw, 2007) and smaller African languages (Wagacha et al., 2006a; Wagacha et al., 2006b; De Pauw and Wagacha, 2007; De Pauw et al., 2007a; De Pauw et al., 2007b).

In this paper we outline on-going research on the development of a parallel corpus English - Swahili. The parallel corpus is designed to bootstrap a data-driven machine translation system for the language pair in question, as well as open up possibilities for projection of annotation.

We start off with a short survey of the different approaches to machine translation (Section 2) and showcase the possibility of projection of annotation (Section 3). We then concentrate on the required data collection and annotation efforts (Section 4) and describe preliminary experiments on sentence, word and morpheme alignment (Sections 5 and 6). We conclude with a discussion of the current limitations to the approach and provide pointers for future research (Section 7).

2 Machine Translation

The main task of Machine Translation (MT) can be defined as having a computer take a text input in one language, the Source language (SL), decode its meaning and re-encode it producing as output a similar-meaning text in another language, the Target language (TL). The idea of building an application to automatically convert text from one language to an equivalent text-meaning in a second language traces its roots back to Cold War intelligence efforts in the 1950's and 60's for Russian-English text translations. Since then a large number of MT systems have been developed with varying degrees of success. For an excellent overview of the history of MT, we refer the reader to Hutchins (1986).

The original dream of creating a fully automatic MT system has long since been abandoned and most research in the field currently concentrates on minimizing human pre- and post-processing effort. A human translator is thus considered to work alongside the MT system to produce faster and more consistent translations.

The Internet brought in an interesting new dimension to the purpose of MT. In the mid 1990s, free on-line translation services began to surface with an increasing number of MT vendors. The most famous example is Yahoo!'s Babelfish , offering on-line versions of Systran to translate English, French, German, Spanish and other Indo-European languages. Currently Google.inc is also offering translation services. While these systems provide far from perfect output, they can often give readers a sense of what is being talked about on a web page in a language (and often even character set) foreign to them.

There are roughly three types of approaches to machine translation:

1. **Rule-based** methods perform translation using extensive lexicons with morphological, syntactic and semantic information, and large sets of manually compiled rules, making them very labor intensive to develop.

2. **Statistical** methods entail the collection and statistical analysis of bilingual text corpora, i.e. parallel corpora. The technique tries to find the highest probability translation of a sentence or phrase among the exponential number of choices.

3. **Example-based** methods are similar to statistical methods in that they are parallel corpus driven. An Example-Based Machine Translator (EBMT) scans for patterns in both languages and relates them in a translation memory.

Most MT systems currently under development are based on methods (2) and/or (3). Research in these fields has greatly benefited from the increasing availability of parallel corpora, which are needed to bootstrap these approaches. Such a parallel corpus is typically aligned by linking, either automatically or manually, linguistic tokens in the source language to the corresponding units in the target language. Processing this data enables the development of fast and effective MT systems in both directions with a minimum of human involvement.

	English Sentences	Swahili Sentences	English Words	Swahili Words
New Testament	7.9k		189.2k	151.1k
Quran	6.2k		165.5k	124.3k
Declaration of HR	0.2k		1.8k	1.8k
Kamusi.org	5.6k		35.5k	26.7k
Movie Subtitles	9.0k		72.2k	58.4k
Investment Reports	3.2k	3.1k	52.9k	54.9k
Local Translator	1.5k	1.6k	25.0k	25.7k
Full Corpus Total	33.6k	33.6k	542.1k	442.9k

Table 1: Overview of the data in the SAWA Corpus

3 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 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., 2007a) 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 parallel corpus. These techniques try to transport the annotation of a well resourced source language, such as English, to texts in a target language. As a natural extension of the domain of machine translation, these methods employ parallel corpora which are aligned at the sentence and word level. 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 unto the text in the target language. The following general principle holds: the closer 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, knowledge transfer is often able to generate a well directed annotation of the target language.

This holds particular promise for the annotation of dependency analyses for Swahili, as exemplified in Figure 1, since dependency grammar focuses on semantic relationships, rather than core syntactic properties, that are much more troublesome to project across languages. The idea is that a relationship that holds between two words in the source language (for instance the *subj* relationship between *cat* and *fell*), also holds for the corresponding linguistic tokens in the target language, i.e. *paka* and *alianguka*.

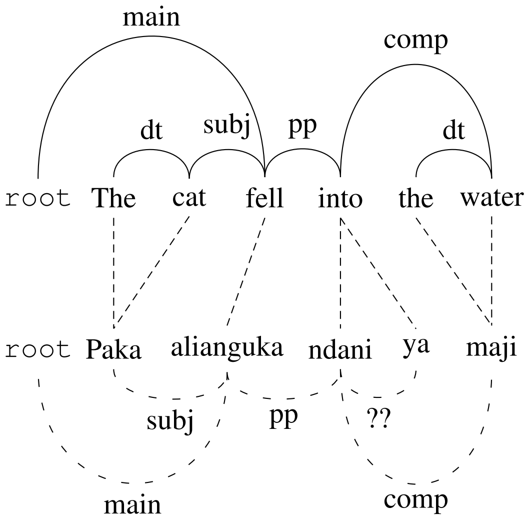
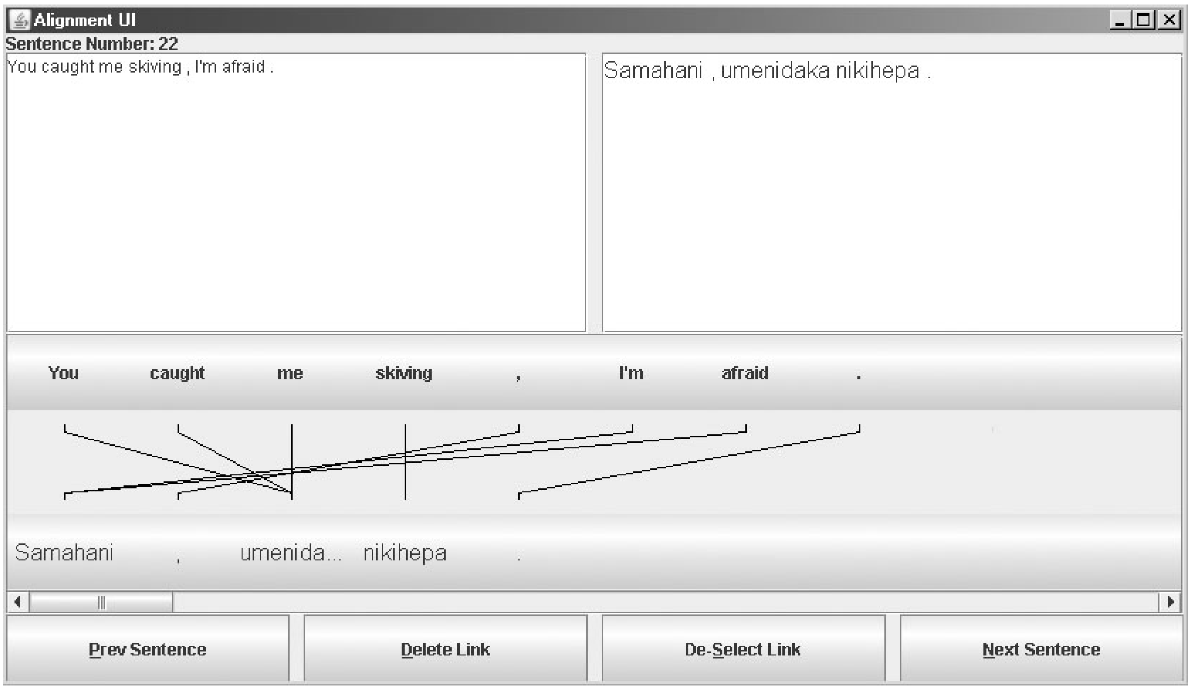


Figure 1: Projection of Dependency Analysis Annotation

In the next section we describe data collection and preprocessing efforts on the SAWA Corpus, a parallel corpus English - Swahili (cf. Table 1), which will enable this type of projection of annotation, as well as the development of a data-driven machine translation system.

4 Data Collection and Annotation

While digital data is increasingly becoming available for Swahili on the Internet, sourcing useful



Sentence Number: 22

You caught me skiving , I'm afraid .

Samahani , umenidaka nikihepa .

Figure 2: Manual word alignment using the UMIACS interface

bilingual data is far from trivial. At this stage in the development of the MT system, it is paramount to use faithfully translated material, as this benefits further automated processing. The corpus-based MT approaches we wish to employ, require word alignment to be performed on the texts, during which the words in the source language are linked to the corresponding words in the target language (also see Figures 1 and 2).

But before we can do this, we need to perform sentence-alignment, during which we establish an unambiguous mapping between the sentences in the source text and the sentences in the target text. While some data is inherently sentence-aligned, other texts require significant preprocessing before word alignment can be performed.

The SAWA Corpus currently consists of a reasonable amount of data (roughly half a million words in each language), although this is not comparable to the resources available to Indo-European language pairs, such as the Hansard corpus (Roukos et al., 1997) (2.87 million sentence pairs). Table 1 gives an overview of the data available in the SAWA Corpus. For each segment it lists the number of sentences and words in the respective languages.

4.1 Sentence-aligned Resources

We found digitally available Swahili versions of the New Testament 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 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 MT system, it does have the advantage of being inherently aligned on the verse level, facilitating further sentence alignment. Another typical bilingual text is the UN Declaration of Human Rights, which is available in many of the world's languages, including Swahili. The manual sentence alignment of this text is greatly facilitated by the fixed structure of the document.

The downloadable version of the on-line dictionary English-Swahili (Benjamin, 2009) 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. These can be extracted from DVDs and while the language is compressed to fit on screen and constitutes scripted language, they nevertheless provide a reasonable approximation of spoken language. Another advantage of this data is that it is inherently sentence-aligned, thanks to the technical time-coding information. It also opens up possibilities for MT systems with other language pairs, since a commercial DVD typically contains subtitles for a large number of other languages as well.

4.2 Paragraph-aligned Resources

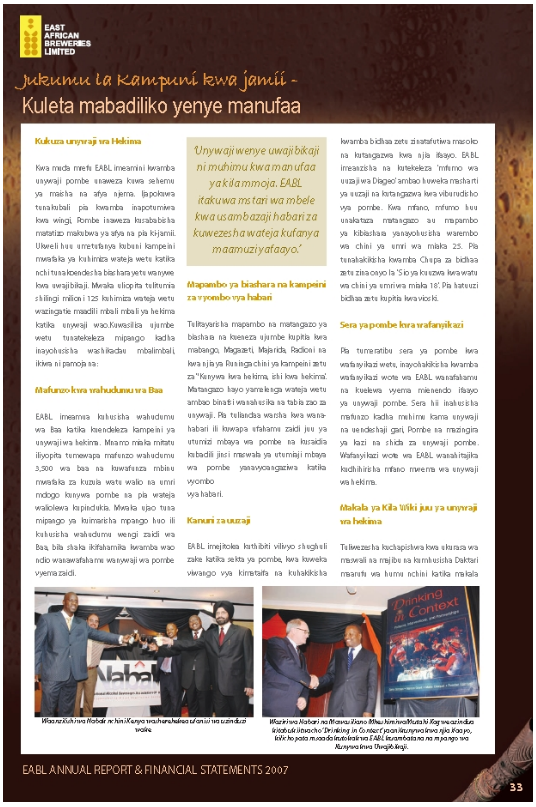
The rest of the material consists of paragraph-aligned data, which was manually sentence-aligned. We obtained a substantial amount of data from a local Kenyan translator. Finally, we also included Kenyan investment reports. These are yearly reports from local companies and are presented in both English 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 manual post-processing and paragraph alignment (Figure 3). 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 MT system may need to process in a practical setting.

The reader may note that there is a very diverse variety of texts within the SAWA corpus, ranging from movie subtitles to religious texts. While it certainly benefits the evaluation to use data from texts in one specific language register, we have chosen to maintain variety in the language data at this point. Upon evaluating the decoder at a later stage, we will however investigate the bias introduced by the specific language registers in the corpus.

4.3 Word Alignment

All of the data in the corpus was subsequently tokenized, which involves automatically cleaning up the texts, conversion to UTF-8, and splitting punctuation from word forms. The next step involved scanning for sentence boundaries in the paragraph-aligned text, to facilitate the automatic sentence alignment method described in Section 5.

While not necessary for further processing, we also performed manual word-alignment annotation. This task can be done automatically, 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 human reference, will allow us to tweak parameters to arrive at the optimal settings for this language pair.



EAST AFRICAN BREWERIES LIMITED

Jukumu la Kampuni kwa jamii -
Kuleta mabadiliko yenye manufaa

Kukuza unywaji wa Hekima

Kwa muda mrefu EABL imeamini kwamba unywaji pombe unaweza kuwa sehemu ya maisha na afya njema. Ijapokuwa tunakubali pia kwamba inapotumiwa kwa wingi, Pombe inaweza kusababisha matatizo makubwa ya afya na pia ki-jamii. Ukweli huu umetufanya kubuni kampeini mwafaka ya kuhimiza wateja wetu katika nchi tunakoendesha biashara yetu wanywe kwa uwajibikaji. Mwaka uliopita tulitumia shilingi milioni 125 kuhimiza wateja wetu wazingatie maadili mbali mbali ya hekima katika unywaji wao.Kuwasilisa ujumbe wetu tunatekeleza mipango kadha inayohusisha washikadau mbalimbali, ikiwa ni pamoja na:

Mafunzo kwa wahudumu wa Baa

EABL imeamua kuhusisha wahudumu wa Baa katika kuendeleza kampeini ya unywaji wa hekima. Mnamo miaka mitatu iliyopita tumewapa mafunzo wahudumu 3,500 wa baa na kuwafunza mbinu mwafaka za kuzuia watu walio na umri mdogo kunywa pombe na pia wateja waliolewa kupindukia. Mwaka ujao tuna mipango ya kuimarisha mpango huo ili kuhusisha wahudumu wengi zaidi wa Baa, bila shaka ikifahamika kwamba wao ndio wanawafahamu wanywaji wa pombe vyema zaidi.

'Unywaji wenye uwajibikaji ni muhimu kwa manufaa ya kila mmoja. EABL itakuwa mstari wa mbele kwa usambazaji habari za kuwezesha wateja kufanya maamuzi yafaayo.'

Mapambo ya biashara na kampeini za vyombo vya habari

Tulitayarisha mapambo na matangazo ya biashara na kueneza ujumbe kupitia kwa mabango, Magazeti, Majarida, Radioni na kwa njia ya Runinga chini ya kampeini zetu za "Kunywa kwa hekima, ishi kwa hekima'. Matangazo hayo yamelenga wateja wetu ambao binafsi wanahusika na tabia zao za unywaji. Pia tuliandaa warsha kwa wanahabari ili kuwapa ufahamu zaidi juu ya utumizi mbaya wa pombe na kusaidia kubadili jinsi maswala ya utumiaji mbaya wa pombe yanavyoangaziwa katika vyombo
vya habari.

Kanuni za uuzaji

EABL imejitolea kuthibiti vilivyo shughuli zake katika sekta ya pombe, kwa kuweka viwango vya kimataifa na kuhakikisha kwamba bidhaa zetu zinatafutiwa masoko na kutangazwa kwa njia ifaayo. EABL imeanzisha na kutekeleza 'mfumo wa uuzaji wa Diageo' ambao huweka masharti ya uuzaji na kutangazwa kwa viburudisho vya pombe. Kwa mfano, mfumo huu unakataza matangazo au mapambo ya kibiashara yanayohusisha warembo wa chini ya umri wa miaka 25. Pia tunahakikisha kwamba Chupa za bidhaa zetu zina onyo la 'Sio ya kuuzwa kwa watu wa chini ya umri wa miaka 18'. Pia hatuuzi bidhaa zetu kupitia kwa vioski.

Sera ya pombe kwa wafanyikazi

Pia tumeratibu sera ya pombe kwa wafanyikazi wetu, inayohakikisha kwamba wafanyikazi wote wa EABL wanafahamu na kuelewa vyema mienendo ifaayo ya unywaji pombe. Sera hii inahusisha mafunzo kadha muhimu kama unywaji na uendeshaji gari, Pombe na mazingira ya kazi na shida za unywaji pombe. Wafanyikazi wote wa EABL wanahitajika kudhihirisha mfano mwema wa unywaji wa hekima.

Makala ya Kila Wiki juu ya unywaji wa hekima

Tuliwezesha kuchapishwa kwa ukurasa wa maswali na majibu na kumhusisha Daktari maarufu wa humu nchini katika makala

Waanzilishi wa Nabalc nchini Kenya washerehekea ufanisi wa uzinduzi wake

Waziri wa Habari na Mawasiliano Mheshimiwa Mutahi Kagwe azindua kitabu kiitwacho 'Drinking in Context' yaani kunywa kwa njia ifaayo, kilicho pata msaada kutoka kwa EABL kwa ushirikiano na mpango wa Kunywa kwa Uwajibikaji.

EABL ANNUAL REPORT & FINANCIAL STATEMENTS 2007

33

Figure 3: Text Extraction from Bilingual Investment Report

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 (Figure 2). Given the linguistic differences between English and Swahili, this is by no means a trivial task. Particularly the morphological richness of Swahili means that there is a lot of convergence from words in English to words in Swahili (also see Section 6). This alignment was done on some of the manual translations of movie subtitles, giving us a gold-standard word-alignment reference of about 5,000 words. Each annotator's work was cross-checked by another annotator to improve correctness and consistency.

5 Alignment Experiments

There are a number of packages available to process parallel corpora. To preprocess the paragraph-aligned texts, we used Microsoft's bilingual sentence aligner (Moore, 2002). The

Precision	**Recall**	**F($\beta = 1$)**
39.4%	44.5%	41.79%

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

Precision	**Recall**	**F($\beta = 1$)**
50.2%	64.5%	55.8%

Table 3: Precision, Recall and F-score for the morpheme/word-alignment task using GIZA++

output of the sentence alignment was consequently manually corrected. We found that 95% of the sentences were correctly aligned with most errors being made on 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. This also explains why there are more Swahili words in the paragraph aligned texts than in English, while the situation is reversed for the sentence aligned data.

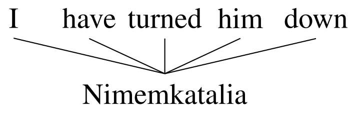
For word-alignment, the state-of-the-art method is GIZA++ (Och and Ney, 2003), which implements the word alignment methods IBM1 to IBM5 and HMM. While this method has a strong Indo-European bias, it is nevertheless interesting to see how far we can get with the default approach used in statistical MT.

We evaluate by looking at the word alignments proposed by GIZA++ and compare them to the manually word-aligned section of the SAWA Corpus. We can quantify the evaluation by calculating precision and recall and their harmonic mean, the F-score (Table 2). The former expresses how many links are correct, divided by the total number of links suggested by GIZA++. The latter is calculated by dividing the number of correct links, by the total number of links in the manual annotation. The underwhelming results presented in Table 2 can be attributed to the strong Indo-European bias of the current approaches. It is clear that extra linguistic data sources and a more elaborate exploration of the experimental parameters of GIZA++ will be needed, as well as a different approach to word-alignment. In the next section, we describe a potential solution to the problem by defining the problem on the level of the morpheme.

6 Alignment into an Agglutinating Language

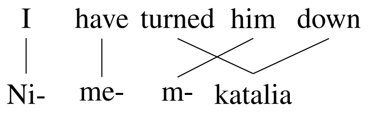
The main problem in training a GIZA++ model for the language pair English - Swahili is the strong agglutinating nature of the latter. Alignment patterns such as the one in Figures 1 and 2 are not impossible to retrieve. But no corpus is exhaustive enough to provide enough linguistic evidence to unearth strongly converging alignment patterns, such as the one in Example 1.

(1)



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

(2)



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 analysis, we developed a machine learning system trained and evaluated on the Helsinki corpus of Swahili (Hurskainen, 2004). Experimental results show that the data-driven approach achieves state-of-the-art performance in a direct comparison with a rule-based method, with the added advantage of being robust to word forms for previously unseen lemmas (De Pauw and de Schryver, 2008). We can consequently use morphological deconstruction as a preprocessing step for the alignment task, similar to the method described by Goldwater and McClosky (2005), Oflazer (2008) and Stymne et al. (2008).

We have no morphologically aligned parallel data available, so evaluation of the morphology-based approach needs to be done in a roundabout way. We first morphologically decompose the Swahili data and run GIZA++ again. Then we recompile the Swahili words from the morphemes and group the word alignment links accordingly. Incompatible linkages are removed. The updated scores are presented in Table 3. While this certainly improves on the scores in Table 2, we need to be aware of the difficulty that the morphological preprocessing 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.

For the purpose of projection of annotation, this is however not an issue. We performed a preliminary experiment with a dependency-parsed English corpus, projected unto the morphologically decompounded tokens in Swahili. We are currently lacking the annotated gold-standard data to perform quantitative evaluation, but have observed interesting annotation results, that open up possibilities for the morphological analysis of more resource-scarce languages.

7 Discussion

In this paper we presented parallel corpus collection work that will enable the construction of a machine translation system for the language pair English - Swahili, as well as open up the possibility of corpus annotation through projection. We are confident that we are approaching a critical amount of data that will enable good word alignment that can subsequently be used as a model for an MT decoding system, such as the Moses package (Koehn et al., 2007). While the currently reported scores are not yet state-of-the-art, we are confident that further experimentation and the addition of more bilingual data as well as the introduction of extra linguistic features will raise the accuracy level of the proposed MT system.

Apart from the morphological deconstruction described in Section 6, the most straightforward addition is the introduction of part-of-speech tags as an extra layer of linguistic description, which can be used in word alignment model IBM5. The current word alignment method tries to link word forms, but knowing that for instance a word in the source language is a noun, will facilitate linking it to a corresponding noun in the target language, rather than considering a verb as a possible match. Both for English (Ratnaparkhi, 1996) and Swahili (De Pauw et al., 2006), we have highly accurate part-of-speech taggers available.

Another extra information source that we have so far ignored is a digital dictionary as a seed for the word alignment. The kamusiproject.org electronic dictionary will be included in further word-alignment experiments and will undoubtedly improve the quality of the output.

Once we have a stable word alignment module, we will further conduct learning curve experiments, in which we train the system with gradually increasing amounts of data. This will provide us with information on how much more data we need to achieve state-of-the-art performance. This additional data can be automatically found by parallel web mining, for which a few systems have recently become available (Resnik and Smith, 2003).

Furthermore, 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 as additional material within a parallel corpus (McEnery and Xiao, 2007) and may further help improve the development of a robust, open-ended and bidirectional machine translation system for the language pair English - Swahili. The most innovative prospect of the parallel corpus is the annotation of dependency analysis in Swahili, not only on the syntactic level, but also on the level of the morphology. The preliminary experiments indicate that this approach might provide a valuable technique to bootstrap annotation in truly resource-scarce languages.

Acknowledgments

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). We are greatly indebted to Dr. James Omboga 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 annotation efforts.

References

M. Benjamin. 2009. *The Kamusi Project*. Available at: http://www.kamusiproject.org (Accessed: 14 January 2009).

G. De Pauw and G.-M. de Schryver. 2008. Improving the computational morphological analysis of a Swahili corpus for lexicographic purposes. *Lexikos*, 18:303–318.

G. De Pauw and P.W. Wagacha. 2007. Bootstrapping morphological analysis of Gĩkũyũ using unsupervised maximum entropy learning. In *Proceedings of the eighth INTERSPEECH conference*, Antwerp, Belgium.

G. De Pauw, G.-M. de Schryver, and P.W. Wagacha. 2006. Data-driven part-of-speech tagging of

Kiswahili. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of Text, Speech and Dialogue, 9th International Conference*, volume 4188 of *Lecture Notes in Computer Science*, pages 197–204, Berlin, Germany. Springer Verlag.

G. De Pauw, P.W. Wagacha, and D.A. Abade. 2007a. Unsupervised induction of Dholuo word classes using maximum entropy learning. In K. Getao and E. Omwenga, editors, *Proceedings of the First International Computer Science and ICT Conference*, pages 139–143, Nairobi, Kenya. University of Nairobi.

G. De Pauw, P.W. Wagacha, and G.-M. de Schryver. 2007b. Automatic diacritic restoration for resource-scarce languages. In Václav Matoušek and Pavel Mautner, editors, *Proceedings of Text, Speech and Dialogue, Tenth International Conference*, volume 4629 of *Lecture Notes in Computer Science*, pages 170–179, Heidelberg, Germany. Springer Verlag.

G.-M. de Schryver and G. De Pauw. 2007. Dictionary writing system (DWS) + corpus query package (CQP): The case of Tshwanelex. *Lexikos*, 17:226–246.

S. Goldwater and D. McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, Canada.

Google. 2009. *Google Translate*. Available at http://www.google.com/translate (Accessed: 14 January 2009.

A. Hurskainen. 2004. HCS 2004 – Helsinki Corpus of Swahili. Technical report, Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC.

W.J. Hutchins. 1986. *Machine translation: past, present, future*. Ellis, Chichester.

R. Hwa and N. Madnani. 2004. *The UMIACS Word Alignment Interface.* Available at: http://www.umiacs.umd.edu/˜nmadnani/alignment/forclip.htm (Accessed: 14 January 2009).

R. Hwa, Ph. Resnik, A. Weinberg, and O. Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 392–399, Philadelphia, PA, USA.

Ph. 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. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

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

R.C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, volume 2499 of *Lecture Notes in Computer Science*, pages 135–144, Berlin, Germany. Springer Verlag.

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

K. Oflazer. 2008. Statistical machine translation into a morphologically complex language. In *Computational Linguistics and Intelligent Text Processing*, pages 376–388, Berlin, Germany. Springer Verlag.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In E. Brill and K. Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics.

Ph. Resnik and N.A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(1):349–380.

S. Roukos, D. Graff, and D. Melamed. 1997. *Hansard French/English.* Available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20 (Accessed: 14 January 2009).

S. Stymne, M. Holmqvist, and L. Ahrenberg. 2008. Effects of morphological analysis in translation between German and English. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 135–138, Columbus, USA.

P.W. Wagacha, G. De Pauw, and K. Getao. 2006a. Development of a corpus for Gĩkũyũ using machine learning techniques. In J.C. Roux, editor, *Proceedings of LREC workshop - Networking the development of language resources for African languages*, pages 27–30, Genoa, Italy, May, 2006. European Language Resources Association, ELRA.

P.W. Wagacha, G. De Pauw, and P.W. Githinji. 2006b. A grapheme-based approach for accent restoration in Gĩkũyũ. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 1937–1940, Genoa, Italy, May, 2006. European Language Resources Association, ELRA.

Yahoo! 2009. *Babelfish*. Available at http://babelfish.yahoo.com (Accessed: 14 January 2009).