A Corpus-based Survey of Four Electronic Swahili–English Bilingual Dictionaries

Guy De Pauw, *CLiPS — Language Technology Group, University of Antwerp, Antwerp, Belgium; and School of Computing and Informatics, University of Nairobi, Nairobi, Kenya (guy.depauw@ua.ac.be)*,

Gilles-Maurice de Schryver, *Department of African Languages and Cultures, Ghent University, Ghent, Belgium; Xhosa Department, University of the Western Cape, Bellville, Republic of South Africa; and TshwaneDJe HLT, Pretoria, Republic of South Africa (gillesmaurice.deschryver@UGent.be)*,
and
Peter Waiganjo Wagacha, *School of Computing and Informatics, University of Nairobi, Nairobi, Kenya (waiganjo@uonbi.ac.ke)*

Abstract: In this article we survey four different electronic bilingual dictionaries for the language pair Swahili–English. Aided by a data-driven morphological analyzer and part-of-speech tagger, we quantify the coverage of the dictionaries on large monolingual corpora of Swahili. In a second series of experiments, we investigate how applicable the dictionaries are as a tool in the development of a machine translation system, by evaluating bilingual coverage on the parallel SAWA corpus. At the same time we attempt to consolidate the dictionaries into a unified lexicographic database and compare the coverage to that of its composite parts.

Keywords: LEXICOGRAPHY, EVALUATION, MORPHOLOGY, LEMMATIZATION, PARALLEL CORPORA, MACHINE LEARNING, MACHINE TRANSLATION, SWAHILI (KISWAHILI), ENGLISH

Samenvatting: Een corpusgebaseerde evaluatie van vier bilinguale elektronische woordenboeken Swahili–Engels. In dit artikel evalueren we vier verschillende elektronische woordenboeken voor het talenpaar Swahili–Engels. Met behulp van automatische morfosyntactische analyse, kwantificeren we de dekking van de woordenboeken op basis van grote monolinguale corpora voor het Swahili. In een tweede reeks experimenten onderzoeken we de toepasbaarheid van de woordenboeken als hulpmiddel bij de ontwikkeling van automatische vertaalsystemen, door hun bilinguale dekking te meten op basis van het parallelle SAWA corpus. Tegelijkertijd proberen we de woordenboeken te integreren in een overkoepelende lexicografische databank en vergelijken we de dekking ervan met die van de samenstellende delen.

Sleutelwoorden: LEXICOGRAFIE, EVALUATIE, MORFOLOGIE, LEMMATISERING, PARALLELLE CORPORA, AUTOMATISCHE LEERTECHNIEKEN, AUTOMATISCH VERTALEN, SWAHILI (KISWAHILI), ENGELS

Lexikos 19 (AFRILEX-reeks/series 19: 2009): 340-352

1. Introduction

Bilingual dictionaries are typically used as linguistic aids, providing support for translators, travellers, foreign language students and comparative linguists. In recent years, bilingual dictionaries have also become essential components in the field of machine translation, particularly for the data-driven approaches. Bilingual dictionaries help establish correct word alignment patterns between words in the source and target language, which enables the automated creation of machine translation systems on the basis of language-independent techniques. For this kind of language technological purpose, the bilingual dictionary not only needs to be available in electronic format, but more importantly, needs to have sufficient coverage of the language pair in question.

Unfortunately, not enough research efforts survey and quantitatively compare dictionaries for this kind of task. In the context of ongoing research on machine translation English ↔ Swahili (De Pauw et al. 2009), we set out to compare and evaluate electronically available bilingual dictionaries for this language pair. This will not only enable us to pick the best candidate(s) for the job at hand, but also to consolidate the information sources into a uniform lexicographic database that can serve as a machine translation aid.

We begin this article by reviewing a previous survey of Swahili dictionaries in Section 2. We then provide a brief description and assessment of the currently available electronic dictionaries Swahili–English in Section 3. The lemmatizer used to perform the lookup procedures in the corpus-based evaluation, is described in Section 4. A quantitative assessment of the coverage of the dictionaries is then given in Section 5, after which we conclude this article with a discussion of the results in Section 6.

2. Previous work

This article updates and complements a previous attempt at surveying Swahili dictionaries using a computational method. Hurskainen (2004) considers five different dictionaries, which are converted into *worsened* finite-state morphological analyzers. Their generative power is consequently evaluated on three different corpora and their coverage is compared to that of SWATWOL (Hurskainen 1992), a *comprehensive Swahili parser*.

The publication, however, does not make it clear how the dictionaries were obtained or converted into digital format, nor does it provide any insight into how the morphological information contained in them is translated into a morphological analyzer. That said, it seems counter-intuitive to evaluate a dictionary as a morphological description of a language, rather than as a lexical one.

Our survey employs an alternative computational and corpus-based evaluation technique, one which can easily be replicated and one which addresses some of the issues apparent in Hurskainen (2004). We focus on readily

available digital dictionaries, dictionaries that can easily be converted into a unified database format. As such, only one dictionary is covered by both Hurskainen (2004) and our survey. We add to our comparison three recently published dictionaries, including the expansive *Internet Living Swahili Dictionary*, which was strangely absent from Hurskainen (2004).

Our evaluation method uses a single, comprehensive Swahili lemmatizer, which is used to retrieve lemmas for word forms in a large Swahili corpus. The lemmatizer that was used in our experiments, allows us to simply evaluate the dictionaries in terms of how many lemmas in the corpus they cover, regardless of the morphological information they encode. Contrary to Hurskainen (2004) we also focus our evaluation of bilingual dictionaries on their potential as tools in machine translation, by comparing their coverage on a parallel corpus.

3. Digitally available bilingual dictionaries Swahili–English

A fair number of bilingual dictionaries Swahili–English have been published over the years. They range from early colonial attempts at Swahili lexicography, to simple tourist phrase books, to fully-fledged translation dictionaries. While the source files for most of these dictionaries are typically not digitally available, the current major dictionaries are electronically accessible and can therefore be included in our survey. In this section, we briefly describe the electronically available dictionaries in terms of development history and features, and provide a first qualitative assessment.

3.1 The *Internet Living Swahili Dictionary* [ILSD]

One of the most famous Swahili–English dictionaries is not only available online, but is also largely developed there: the *Internet Living Swahili Dictionary* at KamusiProject.org. Development on this dictionary started in the early 1990s. Apart from the inclusion of Rechenbach (1967), this dictionary is conceived as a community effort, allowing non-expert users to create and update dictionary entries, which are reviewed by an editorial team. The dictionary is not available in print format, but — like the Freedict dictionary (cf. Section 3.2) — the *Internet Living Swahili Dictionary* (henceforth ILSD), has an open development architecture and the data is readily available for download.

The formatting of the entries in the downloadable files is illustrated in Figure 1. A Swahili word is associated with an English translation equivalent, a part-of-speech tag, and possible inflections and derivations. Some entries also include terminology and taxonomy fields. Many dictionary entries also feature example sentences. Even though this format does not rule out cross-referencing as such, it is currently not an active feature in ILSD.

[Swahili Word] -anguka
[English Word] fall
[Part of Speech] verb

[Derived Word] angika V
[Swahili Example] Theluji ikianguka, hatutakwenda baharini.
[English Example] If snow falls, we won't go to the beach.

Figure 1: Dictionary entry for *anguka* in ILSD

Thanks to the community effort, which has Swahili speakers from around the globe contributing lexical entries, the ILSD is by far the largest Swahili–English dictionary available with more than 60 000 entries. These entries are not comparable to dictionary 'articles', however, as each sense of each lemma is given a separate 'entry' — many of which unfortunately overlap. Also, quite a few inconsistencies and untidy entries can be observed in the dictionary. These include some obvious trial entries still remaining in the database, a sloppy definition of the field "Derived Word" (pointing to derivations and inflections alike, while at other times referring to taxonomy or dialectal features), and an inaccurate attribution of dialectal features to words. Furthermore, translation equivalents are often paraphrased, potentially hampering the use of those entries as an aid for word alignment in the context of machine translation.

3.2 The *Freedict Swahili–English Dictionary* [Freedict]

The *Freedict Swahili–English Dictionary* is an attempt to unify and homogenize existing bilingual dictionaries (Bański and Wójtowicz 2009). It is based on a previously published electronic dictionary (Dict 2009) and also includes entries from a Freedict dictionary (Freedict 2009), and a Swahili–Esperanto–English dictionary (Ergane 2009). It uses the open-source Freedict architecture for development and dissemination and sources are therefore freely downloadable.

The latest version includes 2 600 entries, associated with an English translation equivalent and a part-of-speech tag. Figure 2 illustrates the typical layout of the entries. While the dictionary itself is very small and the information provided is scarce, the developers seem to have tried as much as possible to provide single-word translation equivalents that bode well in a machine translation environment.

anguka
fall
(v)

Figure 2: Dictionary entry for *anguka* in Freedict

3.3 The *TshwaneDJe Swahili–English Dictionary* [TeDJe-SED]

The *TshwaneDJe Swahili–English Dictionary* (Hillewaert, Joffe and De Schryver 2009), for short TeDJe-SED, is the most recently published work in our survey. It includes about 16 000 entries and features morphological decomposition, corpus-based example phrases, and an intricate system of cross-references. It

also includes a tool that provides translation equivalents in Microsoft Word, which indicates the developers had (human-aided) machine translation in mind during development.

The dictionary was created using TshwaneDJe's in-house lexicography tool *TshwaneLex* (Joffe et al. 2009). It can be accessed on-line through a web interface and is available as a stand-alone download as well. The actual source files can only be accessed through *TshwaneLex*. Even though it features a fairly limited number of lemmas, it provides by far the most detailed lexicographic information of all the dictionaries in this survey, as illustrated in Figure 3.

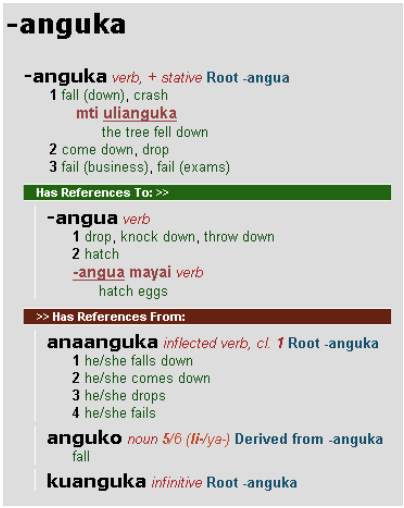

Figure 3: Dictionary entry for *anguka* in TeDJe-SED

3.4 The TUKI *Swahili–English Dictionary* [TUKI]

The *Kamusi ya Kiswahili–Kiingereza/Swahili–English Dictionary* (TUKI 2006[2]) was developed at the then Institute of Kiswahili Research of the University of Dar es Salaam, in Tanzania, over the course of three years and was first published in hard copy in 2001. It is currently the best-selling paper bilingual dictionary Swahili–English. It includes lemmas, part-of-speech tags, translation equivalents, sporadic example phrases, and derivations.

In 2003 a digital version of the dictionary became available on CD-ROM. It consists of formatted HTML files with the same content as the hard copy. In principle, the HTML formatting tags should allow us to extract the required information for the purpose of our experiments, as illustrated in Figure 4. Unfortunately, a large number of inconsistent formatting issues, not present in the hard copy, can be observed in the digital version.

anguk.a *kt* [*sie*] 1 come down, fall down, drop, crash. 2 lose in a business. 3 fail: *Ame~ mitihani yake* he failed his examinations. *(tde)* **angukia**, *(tden)* **angukiana**, *(tdew)* **angukiwa**; *(tdk)* **angukika**. **anguko** *nm.*

```
<p align="JUSTIFY"><b>anguk.a</b> <i>kt</i>
[<i>sie</i>] 1 come down, fall down, drop, crash.
2 lose in a business. 3 fail: <i>Ame~ mitihani
yake</i> he failed his examinations. <i>(tde)</i>
<b>angukia</b>, <i>(tden)</i> <b>angukiana</b>,
<i>(tdew)</i> <b>angukiwa</b>; <i>(tdk)</i>
<b>angukika</b>. <b>anguko</b> <i>nm.</i></p>
```

Figure 4: Dictionary entry for *anguka* in TUKI
[*NB! Dictionary article: commas vs. semicolons to separate the run-ons as in original.*]

Since no updated copies were available, we decided to semi-automatically clean up the HTML files. A range of scripts scanned for unusual patterns in the HTML formatting. These were automatically converted and consequently proof-read by a human annotator. This resulted in a cleaner and more consistently formatted electronic dictionary, which can be converted into the database format required for our experiments.

3.5 Consolidation of sources

To perform a quantitative survey, we had to convert the four different lexicographic sources to a uniform database format. For this we propose the format illustrated in Table 1. The main field is the lemma, from which we removed root indications (e.g. hyphens in ILSD and TeDJe-SED; dots in TUKI). We add part-of-speech tag information and noun class information where applicable. We include a field with different English translations for the lemma at hand. If inflections of a lemma are listed, they are included in the "Related words" field. Finally, each database record is associated with its source. Note that multiword expressions are not included in the consolidated database.

Table 1: Consolidated lexical database for four dictionaries

Lemma	POS-tag	Class	English translation	Related words	Source
geuko	noun	ma-	change, transformation		TUKI
geuza	verb	–	change, modify, ...	aligeuza, anageuza, kugeuza, niligeuza, ...	TeDJe-SED
geuzo	noun	5/6	change	mageuzo	ILSD

Table 2 shows some quantitative information about the converted dictionaries.

Table 2: Quantitative information for four dictionaries and consolidated database (approximate numbers)

	ILSD	Freedict	TeDJe-SED	TUKI	ALL
Swahili entries	61 000	2 600	15 500	14 500	—
Number of lemmas	17 000	2 500	2 500	13 000	21 000
Unique lemmas	8 000	100	150	3 900	—

The first row displays the number of dictionary entries as advertised by the developers. The second row shows the number of orthographically distinct lemmas per dictionary (not taking into account homographs with different morphosyntactic or semantic features). The last row shows how many lemmas are exclusive (i.e. unique) to that dictionary. Both the larger dictionaries as well as TeDJe-SED and Freedict include a fair number of unique lemmas, so unifying the different sources can lead to a rich lexicographic database.

4. Swahili morphological analyzer

In Section 5 we will compute the coverage of the dictionaries on the basis of large corpora. Given the rich morphological features of Swahili, however, we first need to lemmatize the word forms in the corpus to be able to match them to the lemmas in the database (as illustrated in Table 1). In De Pauw and De Schryver (2008) we introduced the first data-driven morphological analyzer for a Bantu language. We described how the lemmatized Helsinki Corpus of Swahili (Hurskainen 2004a) can be used as an information source that powers a lemmatizer using the machine learning technique of memory-based learning. We proceeded to show in a number of different experiments how the analyzer can be observed to significantly outperform a meticulously designed rule-based approach.

Since then, we continued development of the system and made some significant changes, including the introduction of trigram-based classification, which has previously shown to be beneficial for morphological processing (Van den Bosch and Daelemans 2005). The consolidated database described in Section 3.5 also yielded extra "word form — lemma" pairs that further complemented the training data. Finally, we replaced the machine learning technique of memory-based learning with that of maximum entropy learning. All of the above tweaks serve to further improve the accuracy of the lemmatizer from 88% (De Pauw and De Schryver 2008: 312) to 91% for morphologically complex words.

5. Computing the coverage

In this section we attempt to quantify the coverage of the respective dictionaries on the basis of a large monolingual Swahili corpus (Section 5.1). We also

investigate the usability of the dictionaries as a tool in machine translation, by looking at their coverage on a parallel corpus English–Swahili (Section 5.2).

5.1 Monolingual corpus

We used several textual sources to compute the coverage of the dictionaries. These include:

— The Helsinki Corpus of Swahili, HCS (Hurskainen 2004a) consisting of more than 9 million words.

— The TshwaneDJe Kiswahili Internet Corpus, TeDJe-KIC (De Schryver and Joffe 2009) of more than 20 million words.

— The Swahili part of the parallel SAWA corpus (De Pauw et al. 2009), containing ± 0.5 million words.

— Wikipedia in Swahili: almost 12 000 Internet pages, good for more than 1 million words.

We pre-processed the texts by uniformly converting them into UTF-8, tokenizing the data and lemmatizing them using the automatic morphological analyzer described in Section 4. The data was also part-of-speech tagged using the method described in De Pauw et al. (2006). The Helsinki Corpus of Swahili already has lemmatization and part-of-speech tag information available. We nevertheless chose to process it again using our own techniques, for reasons of annotation accuracy and consistency across the data sets.

We then proceeded to compute coverage. We used a purely quantitative approach for this, which checks for each word in the corpus whether its lemma (for the given part-of-speech) can be retrieved in the dictionaries. Table 3 displays the scores for the different dictionaries and corpora.

Table 3: Coverage scores on monolingual corpora (in %)

	ILSD	**Freedict**	**TeDJe-SED**	**TUKI**	**ALL**
HCS	87.9	50.7	60.4	73.4	90.0
TeDJe-KIC	88.2	51.0	61.2	74.1	90.5
SAWA corpus	85.5	50.2	60.2	71.9	89.9
Wikipedia	83.4	48.9	59.4	69.8	85.2
ALL DATA	87.9	50.8	60.9	73.7	90.2

The ILSD has the highest coverage across the board, but loses a lot of coverage for the more recent texts in the recently developed and noisier Wikipedia pages. TUKI follows the same trend, while the TeDJe-SED dictionary hardly loses coverage. The latter's smaller set of lemmas consistently covers the most frequent words in the corpus and is therefore not as vulnerable to change of

register and publication date. The complete consolidated dictionary database performs quite well with an overall coverage of about 90.2%.

To study the effect of publication date, we calculated the coverage per year of the periodicals included in the Helsinki Corpus of Swahili (1990 → 2002, no data for 1995, 1996, 1997). The downward trend in coverage is visible for all four dictionaries; see Figure 5. The most frequent items not covered by the dictionaries are named entities, foreign words and IT terminology. The need to update dictionaries consistently is therefore high. The open architectures of the Freedict and ILSD projects are in this sense suitable solutions.

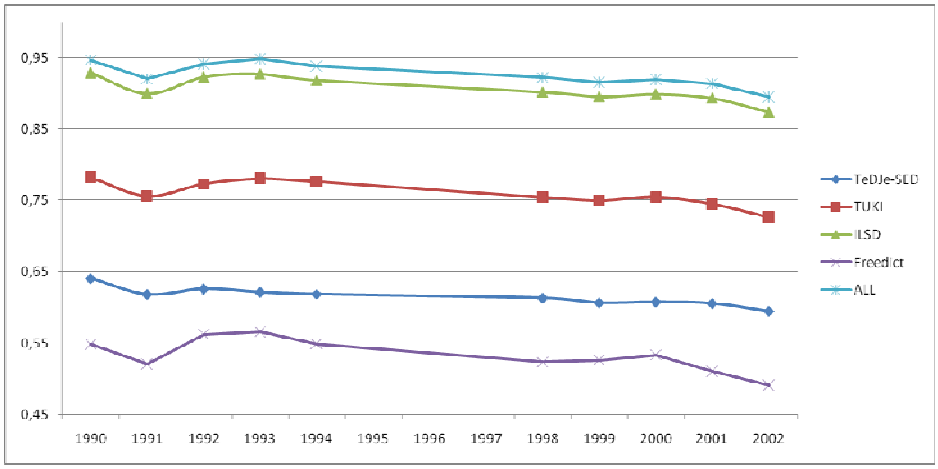

Figure 5: Coverage shift over time

Interestingly, the scores reported here differ significantly from those in Hurskainen (2004). Overall, coverage scores are lower than those reported in the previous survey, which may be due to differences in evaluation metrics. Stranger however is that Hurskainen (2004) observes higher recall scores for more recent documents, whereas Figure 5 shows a definite downward trend over time. These discrepancies warrant further investigation.

We also calculated the coverage of the dictionaries disregarding the frequency of the lemma. In this calculation, covering a highly frequent word like *lakini* 'but; however; nevertheless' scores the same as covering a hapax. Table 4 shows that in this experiment ILSD is trailing TUKI, indicating that even though ILSD contains many more entries, TUKI seems to cater for a wider range of words.

A final experiment counts for how many lemmas in the respective dictionaries evidence can be found in the corpus. The last line of Table 4 shows that TeDJe-SED has all lemmas covered by real-world data. About 30% of the entries in TUKI are not found in the data, while only two thirds of ILSD is covered by the corpus. Comparing the data in Tables 3 and 4 shows that while TUKI trails in comparison to ILSD in terms of raw coverage (Table 3), it does

seem to strike a better balance in terms of both lexical richness and empirical evidence (Table 4).

Table 4: Coverage scores for unique lemmas and reverse coverage (in %)

	ILSD	Freedict	TeDJe-SED	TUKI
ALL DATA	21.7	5.7	7.8	22.7
Reverse coverage	67.4	95.5	100	70.5

Owing to the massive amount of data, it is impossible to check whether the lemma retrieved in the dictionary is indeed the one intended in the text. It might indeed be the case that a particular lemma-tag combination retrieved in the dictionary, does not describe the correct meaning in its actual context. We manually checked a small section of the corpus (±2 000 words) and found only two occasions of such an error. We are therefore confident that our scores are reliable in the context of the comparison between the dictionaries.

In De Pauw and De Schryver (2008) we presented our morphological analyzer as a way to unearth undiscovered lemmas in the corpus data. Our approach indeed has the distinct advantage that it is not dependent on a preset list of roots or lemmas, and is thus capable of lemmatizing word forms for previously unseen lemmas. The experiments outlined in this section have further underlined this property, as we now have at our disposal a list of word forms and associated candidate lemmas (roughly put the remaining 10% not covered by the consolidated dictionary) that need to be lexicographically described.

5.2 Parallel corpus

So far we have only evaluated the dictionaries in a monolingual context. We have calculated the raw coverage of the dictionaries, but this does not provide any insight into the suitability of the dictionary as a bilingual information source for (machine) translation purposes. To properly estimate this, we used the parallel SAWA corpus (De Pauw et al. 2009).

The SAWA corpus is a one million word bilingual corpus, consisting of political documents, religious texts, movie subtitles, investment reports and other documents in both English and Swahili. They were semi-automatically sentence aligned and a small portion was manually word aligned.

The third line in Table 3 gives us some insight into how many lemmas are covered in the Swahili portion of the corpus. We then counted how many times the *actual translation* provided by the dictionary can be found in the associated English part. This gives us some insight into how useful the bilingual dictionary is in the context of machine translation and how appropriate the translation equivalents are that are provided by the respective dictionaries from an empirical point of view.

The first row of Table 5 shows that the small TeDJe-SED dictionary out-

performs the other dictionaries in terms of accuracy. The ILSD has a surprisingly low score. While it covers more than 85% of the lemmas in the SAWA corpus (cf. Table 3), this is only useful 59% of the time in the context of machine translation. Even TUKI's translation equivalents seem to be better suited to the task. The smaller dictionaries have a higher score, since they tend to cover more frequent words, which they describe better.

In a further experiment we take the subset of lemmas that is shared by all dictionaries and investigate their bilingual coverage in the SAWA corpus. Hereby we level the playing field in terms of dictionary size and only compute the usability of the respective bilingual dictionaries as a machine translation tool. The results can be found on the last row in Table 5. While the differences are definitely more narrow, Freedict and TeDJe-SED surprisingly keep the edge over the larger dictionaries, indicating they truly provide more useful translation equivalents.

Table 5: Bilingual coverage scores

	ILSD	**Freedict**	**TeDJe-SED**	**TUKI**
SAWA corpus (all)	58.5%	67.6%	69.4%	60.5%
SAWA corpus (common subset)	64.5%	67.6%	69.4%	65.6%

This experiment shows that evaluating the coverage of bilingual dictionaries needs to be performed on different fronts. Raw coverage scoring on monolingual data does indeed give us some insight into the scope of the dictionary and in this sense both TUKI and ILSD score very well. In terms of how useful the dictionary is as a tool in machine translation, computing the coverage of bilingual documents provides an interesting, alternative insight into the matter, and here Freedict and especially TeDJe-SED score admirably.

6. Discussion

In this article we compared four different electronically available dictionaries for Swahili: one ported from a standard hard copy dictionary (TUKI), two developed and distributed electronically as a community effort (Freedict and ILSD), and a small but accurate electronic one developed by a lexicographic company (TeDJe-SED). From the various analyses presented, it is clear that none of the dictionaries by themselves offer a one-stop solution for machine translation work.

Of course, it is important to keep in mind that none of these four dictionaries was conceived with the aim to function as a component in an automated translation system. The four dictionaries in this study mostly have different aims and target user groups in mind, apart from being compiled in quite different ways. Therefore, comparing them from a very specific angle, an angle that was not intended by the compilers, does not really do justice to any of them. As such, the fact that the reverse coverage of TUKI is 'only' 70% for

example (cf. Table 4) may be seen as a positive aspect, as the compilers surely attempted to cover as wide a range of vocabulary as possible, and may even have included obsolete terms on purpose. Conversely, the 100% reverse coverage for TeDJe-SED is exactly a design feature, given the compilation of that dictionary is directly inspired by corpus facts (cf. De Schryver et al. 2006).

We nevertheless made several attempts at unifying the different dictionaries into a consolidated lexicographic database. This indeed improved monolingual coverage to more than 90%. To compute bilingual corpus coverage of the consolidated database, we needed to first resolve conflicts between database fields for similar lemmas. Our experiments showed that the order of preference TeDJe-SED → Freedict → TUKI → ILSD yielded the best results, with about 70% of word pairs retrieved.

We believe that the consolidated database will be of great value to our machine translation system, as it helps link the English words to the associated Swahili words in the translation pairs. Reversely, the parallel corpus also contributes to the discovery of new, previously unrecorded translation pairs. Future research will investigate how this iterative procedure can be maximally exploited in a lexicographic, as well as a language technological context.

The biggest challenge however remains the development of a large coverage and effective machine translation system for Swahili. Even the recently released Google Translation System for Swahili seems to suffer from some apparent gaps in the vocabulary. We are confident that a machine translation system built using the consolidated database described in this article can significantly alleviate this problem.

It is actually interesting to note that Swahili, the widest spoken Bantu language, still does not have a fully functional bilingual dictionary available that is applicable in the context of machine translation. The vast coverage of ILSD is somewhat hampered by the noise in the database fields and the often impractical translation equivalents. TeDJe-SED seems highly accurate and lexicographically sound, but is so far lacking in terms of raw coverage. Freedict is a simple and effective dictionary, but suffers from its limited scope. Finally, TUKI strikes a nice balance between size and lexicographic scope, but is seemingly in arrested development, which is unfortunate in light of the graphs displayed in Figure 5. We are however confident that the electronic availability of these dictionaries and our evaluation thereof might help lead the way for lexicographers to develop a new, accurate and large-coverage bilingual Swahili–English dictionary, one that does not only serve as a human translation aid, but distinctly moves forward towards machine translation as well.

Acknowledgements

Guy De Pauw is funded as a Postdoctoral Fellow of the *Research Foundation — Flanders* (FWO). Gilles-Maurice de Schryver would like to thank *Ghent University* for its continued support of his field trips in Africa.

References

Bański, P. and B. Wójtowicz. 2009. A Repository of Free Lexical Resources for African Languages: The Project and the Method. De Pauw, G. et al. (Eds.). 2009: 89-95.

De Pauw, G. and G.-M. de Schryver. 2008. Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes. *Lexikos* 18: 303-318.

De Pauw, G., G.-M. de Schryver and L. Levin (Eds.). 2009. *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages*. Athens: Association for Computational Linguistics.

De Pauw, G., G.-M. de Schryver and P.W. Wagacha. 2006. Data-Driven Part-of-Speech Tagging of Kiswahili. *Lecture Notes in Artificial Intelligence* 4188: 197-204.

De Pauw, G., P.W. Wagacha and G.-M. de Schryver. 2009. The SAWA Corpus: A Parallel Corpus English–Swahili. De Pauw, G. et al. (Eds.). 2009: 9-16. [SAWA corpus]

De Schryver, G.-M. and D. Joffe. 2009. *TshwaneDJe Kiswahili Internet Corpus.* Pretoria: TshwaneDJe HLT. [TeDJe-KIC]

De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

Dict. 2009. The DICT Development Group [online]. http://dict.org

Ergane. 2009. A Multilingual Translation Dictionary for Windows [online]. http://download.travlang.com/Ergane

Freedict. 2009. Free Bilingual Dictionaries [online]. http://freedict.org [Freedict]

Google Translate. 2009. Google's free online language translation service instantly translates text and web pages [online]. http://translate.google.com [Google]

Hillewaert, S., P. Joffe and G.-M. de Schryver. 2009. *Kamusi ya Kiswahili–Kiingereza Katika Mtandao/Online Swahili–English Dictionary* [online]. http://africanlanguages.com/swahili/ [TeDJe-SED]

Hurskainen, A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1: 87-119.

Hurskainen, A. 2004. Computational Testing of Five Swahili Dictionaries. Karlsson, F. (Ed.). 2004. *Proceedings of the 20th Scandinavian Conference of Linguistics, Helsinki, January 7–9, 2004*. Department of General Linguistics Publications No. 36 [online]. Helsinki: University of Helsinki. http://www.ling.helsinki.fi/kielitiede/20scl/proceedings.shtml

Hurskainen, A. 2004a. HCS 2004 — Helsinki Corpus of Swahili. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC. [HCS]

Joffe, D. et al. 2009. *TshwaneLex Suite* [online]. http://tshwanedje.com/tshwanelex/

Kamusi Project. 2009. The Internet Living Swahili Dictionary [online]. http://kamusiproject.org [ILSD]

Rechenbach, C.W. 1967. *Swahili–English Dictionary.* Washington: Catholic University of America Press.

TUKI. 2006². *Kamusi ya Kiswahili–Kiingereza/Swahili–English Dictionary*. Dar es Salaam: Taasisi ya Uchunguzi wa Kiswahili, Chuo Kikuu cha Dar es Salaam. [TUKI]

Van den Bosch, A. and W. Daelemans. 2005. Improving Sequence Segmentation Learning by Predicting Trigrams. *Proceedings of the Ninth Conference on Natural Language Learning*: 80-87. Ann Arbor: Association for Computational Linguistics.

Wikipedia. 2009. Wikipedia in Swahili [online]. http://sw.wikipedia.org [Wikipedia]