Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography*

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)*,

David Joffe, *TshwaneDJe HLT, Pretoria, Republic of South Africa (david.joffe@tshwanedje.com)*,

Pitta Joffe, *Pretoria, Republic of South Africa (leonjoffe@gmail.com)*, and

Sarah Hillewaert, *Department of Anthropology, University of Michigan, Michigan, United States of America (sarahhil@umich.edu)*

Abstract: An innovative online Swahili–English dictionary project is presented. A careful study of some of the log files attached to this reference work reveals some hitherto unknown aspects of true dictionary look-up behaviour, which results in the depreciation of the importance of corpora for dictionary making. Three lexicography software modules are advanced to further enhance the success of the online dictionary.

Keywords: LEXICOGRAPHY, SOFTWARE, ONLINE, DICTIONARY, LOG FILE, CORPUS, FREQUENCY, RANK, CORRELATION, SWAHILI, ENGLISH, TSHWANELEX

Samenvatting: Zoeken woordenboekgebruikers werkelijk frequente woorden op? — Over de overschatting van de waarde van corpusgebaseerde lexicografie. Een vernieuwend online Swahili–Engels woordenboekproject wordt voorgesteld. Een minutieuze studie van enkele van de log bestanden gekoppeld aan dit referentiewerk onthult tot dusver onbekende aspecten van het echte opzoekgedrag van woordenboekgebruikers, wat leidt tot een devaluatie van het belang van corpora voor het maken van woordenboeken. Drie lexicografische softwaremodules worden naar voor geschoven om het succes van het online woordenboek verder te vergroten.

Sleutelwoorden: LEXICOGRAFIE, SOFTWARE, ONLINE, WOORDENBOEK, LOG BESTAND, CORPUS, FREQUENTIE, RANG, CORRELATIE, SWAHILI, ENGELS, TSHWANELEX

* Sections of this article are based on a presentation by G.-M. de Schryver at the Tenth International Conference of the African Association for Lexicography, organised by the Sesiu sa Sesotho Lexicography Unit, University of the Free State, Bloemfontein, Republic of South Africa, 13–15 July 2005.

1. Corpus-based Lexicography

Since the beginning of modern corpus-based lexicography with the COBUILD project in the early 1980s (Sinclair 1987), hardly anyone has doubted the value of using electronic corpora on both the macrostructural and microstructural levels during the entire compilation process of reference works. Literally hundreds of research papers exalt the benefits of using corpora in lexicography, and upon studying the arguments one also immediately and intuitively tends to agree.

In oversimplified terms the main advantages can be summarised as follows. When one draws a lemma-sign list from the top section of a corpus-derived (lemmatised) frequency list, then the resulting macrostructure (i.e. 'the list of headwords' plus their parts of speech and some morphological guidance) will also provide the dictionary user with what he/she is most likely to want to look up. Focussing on the microstructure of reference works, it further seems sound to accept that when senses are arranged according to their occurrences in corpora, that when examples are chosen from the living language, and that when also all other aspects such as collocationality issues are based on corpus statistics, that the user will again be most satisfied.

When, back in 2003, two of the current authors applied for the *Kernerman Dictionary Research Grants* to support 'The Creation of an Innovative Kiswahili–English Online Dictionary', those same assumptions could also be descried. Given all the data in the current article will be based on this online dictionary project, a summary of that proposal seems in order.

2. Project Proposal Summary

Swahili (or Kiswahili in the language itself) is one of Africa's major languages, is the official language of Tanzania, and is also spoken throughout East Africa as a lingua franca by several dozen million people. Since well over a century ago, numerous mono- and bilingual dictionaries have been compiled for Swahili. Given its official status, the substantial number of speakers, and a relatively long lexicographic tradition, one would assume an advanced state of lexicographical research as well as the availability of modern and up-to-date dictionaries for Swahili. This is unfortunately not really the case. Western compilation principles were largely transferred, and up to today the most commonly used Swahili dictionaries remain rooted in lexica originally compiled by missionaries. Overall, one also notices a restricted to non-existent dictionary culture.

Swahili is an agglutinating language, which means that morphemes are juxtaposed to form linguistic words. In all current Swahili dictionaries, 'orthographic words' have been decomposed into their formatives, with only the latter being lemmatised. Not all primary speakers of Swahili can look up 'words' in their own language (as this implies being able to cut off pre- and suffixes),

and even trained learners and scholars often need more than one look-up round before they find what they are looking for (as sound changes between formatives are not always predictable).

In this research project the idea is to deal with all these problems simultaneously. The aim is to create the first corpus-based dictionary that is also intuitive in nature, and to research the feasibility of this approach in real time. Instead of lemmatising stems as in traditional Swahili dictionaries, the suggestion is to lemmatise full orthographic words in addition to stems, and to provide full translations for these strings. In order to sensibly limit the number of items one can physically treat, the intention is to select the items from a frequency list derived from a large corpus. Concordance lines will be culled from the corpus for each frequent orthographic word, and the various translations will be recorded in order of frequency.

A user will thus be able to directly look up words as they are spoken or written, and the translations found will be from most likely to least likely. An English search index will additionally enable searches in the other direction. Such an approach will obviously require much more 'space' than in a traditional stem-based dictionary, which is why the dictionary will be developed and made available in an electronic environment right from the start. This environment will primarily be on the Internet, where it is possible to keep a log of all searches. The analysis of such log files will enable the team to research whether or not this hybrid approach is feasible and to amend the approach if need be.

Given the intuitive lemmatisation approach, especially primary speakers and learners at the elementary and intermediate levels will for the first time be able to effectively look up words, and find meanings of 'real' words, which should come some way in combating the lack of a dictionary culture. Furthermore, log files in an electronic environment are a notoriously underused tool, a tool that will be utilised to its full potential in this project. Each visitor to the dictionary will automatically receive a user ID with which dictionary-using behaviour, including vocabulary retention, will be tracked. For the first time, truly unobtrusive data will be collected and true look-up behaviour in an electronic environment will be recorded. Finally, this project will ensure that Swahili, an increasingly popular language on the Internet, is also kept alive in a modern online reference work based on sound lexicographical principles.

3. The Result: An Innovative Online Swahili–English Dictionary

Three years later, this online Swahili–English dictionary has indeed been produced, and is available at <http://africanlanguages.com/swahili/>. As is clear from the project proposal summary above, it has always been (and it still remains) the intention to view the project as a 'work in progress' — or even a 'research environment' — where the project members can freely try out

different approaches to making better (online) dictionaries. Adaptations and changes are based on the searches and look-up behaviour seen in the log files, and involve the development of new lexicography software modules.

As a random example, Addendum 1 shows a screenshot of what a primary speaker of Swahili will see when looking up the English phrase 'that one over there'. One may firstly notice that the entire dictionary interface text and the dictionary's metalanguage have both been 'adapted' (customised) to Swahili, the language of the user. If one browses the same dictionary in English, as is done in Addendum 2, all of these aspects will of course be in English. This is an example of 'dynamic metalanguage customisation', as explained in De Schryver and Joffe (2005).

In a stem-based dictionary for Swahili, only the root *-le* would have been lemmatised, with, in the better dictionaries, also an indication that this root is used to form 'demonstratives of position 3'. If one does not know the concordial agreement system of Swahili, however, this information is not of much help, as during actual usage, this root takes different forms according to the class (cl.) of the noun: *yule* (cl. 1), *wale* (cl. 2), *ule* (cl. 3 and 11), *ile* (cl. 4 and 9), *lile* (cl. 5), *yale* (cl. 6), *kile* (cl. 7), *vile* (cl. 8) and *zile* (cl. 10).

The data shown in Addendum 1 were directly output from *TshwaneDJe HLT*'s dictionary compilation software TshwaneLex, and as one can see, each of the possible (frequent) forms has been lemmatised, *in addition* to the inclusion of the root. Cross-references (or 'hyperlinks' online) moreover link each of the various forms with the root. Related material is shown on the same output screen, which means that no matter whether users start at any of the full forms (as in Addendum 2) or at the root, they will receive guidance, see the meaning, and be provided with (corpus-based) usage examples.

This hybrid approach to lemmatisation has the advantage that words are 'restored' to their actual appearance in written text, and thus that meaningful and pronounceable text strings are shown rather than mere linguistic concepts (i.e. roots) only.

Showing full forms has another advantage, as seen in Addendum 2, from which one can derive that *ule* as 'that one over there' (for class 3 and class 11 nouns) is homonymous with another *ule* with the meaning 'that you eat' (derived from, and linked to, the root *-la* 'eat'). The latter is exactly one of the innovative aspects introduced into this project from the early stages. Given there are literally thousands of possible forms for each verb, the idea was to only treat the frequent full forms.

Likewise for the other parts of speech: In each case the team would 'focus on' and physically 'enter' full forms only when the corpus frequencies would warrant doing so. The corpus would thus be used as the ultimate arbiter, and with a balanced and representative Swahili corpus of around fifteen million running words, the assumption was that this approach would indeed also answer most users' queries. The remainder of this article now studies how successful this assumption was.

4. Bergenholtz and Johnsen (2005): Furthering the Discussion

In response to an article by De Schryver and Joffe (2004), the metalexicographers Bergenholtz and Johnsen (2005: 122) wrote:

> There are only a few published scholarly descriptions of internet dictionary log files. The most interesting contribution from de Schryver/Joffe (2004) describes the log file for a South-African bilingual dictionary, a Sesotho sa Leboa–English dictionary. The number of visitors and the number of lookups is not very high: 21,337 lookups made by 2,530 different visitors. [...] De Schryver/Joffe (2004) fail to mention one very interesting point: With 28,000 English lemmas and 25,000 Sesotho sa Leboa lemmas, the users cannot have looked up all lemmas (with only 21,337 lookups). It would be most interesting to know which types of words are not looked up: Is about 90% or 80% of the dictionary never used at all? The very limited number of lookups indicates that no more than 40–50% of the dictionary is actually being used. Will all lemmas in the dictionary be looked up in time when the dictionary has had many more users? Or are there some lemmas that will never be looked up? If future dictionary makers knew the answers to those questions, they would not have to waste time describing words of no interest to the users.

These are indeed intriguing questions, and questions one can answer once one has enough data. These questions also link in well with what the current project team wanted to find out through a thorough study of the log files of the online Swahili–English dictionary.

A preliminary remark is in order, however, and it concerns word-status. As is well known, users of electronic dictionaries, whether on CD-ROM, an intranet, or the Internet, search for much more than just 'words'. Log files attached to such dictionaries clearly show that users increasingly assume that electronic dictionaries behave like Web search engines such as *Google*, and type in concatenations of keywords, combinations and phrases surrounded by quotes, entire sentences, and even dump full paragraphs (lifted from other sources) into the search field. In addition to that, an increasing number of people do not care about spelling, even type in SMS-like words and smileys, and search for a variety of languages other than the one(s) the dictionary is treating.

For languages such as Swahili and Northern Sotho (Sesotho sa Leboa), there is the additional problem of word-status on word-level itself, with a difference between 'linguistic word' and 'orthographic word'. With for instance thousands of verbal forms for a single verb root, is one dealing with thousands of (orthographic) words, or with just one (linguistic) word? When is which 'word' described? With this in mind, the (Indo-European-biased) statement by Bergenholtz and Johnsen above is at least slightly naive.

In other words, in order to 'solve' the problem posed, one must actually make sure that users of a dictionary 'understand' the lemmatisation approach used in a certain dictionary. Imagine a user wants to find the translation of the

Swahili *mchezaji* 'player'. In a stem-based dictionary, the user will have to know that this word is a deverbative, and thus look up the verbal root *-cheza* 'to play', but in the hybrid approach of the online Swahili–English dictionary either a search for the verbal root or the full noun will reveal the answer. This thus means that one should, in order to see which percentage of this dictionary is being returned over time, also take the cross-referenced material into account, as this material is shown concurrently with the searched-for-item.

The Swahili dictionary was placed online in May 2004, and a little over two years later, over half a million searches were made in both 'directions'. Observe that the latter is actually another novelty introduced in this online dictionary, as only one side is being prepared: Swahili *to* English. To allow for English searches to be made, a smart English index is generated. This is described online as follows:

> Note: This is not your 'standard' English to Swahili dictionary. Instead, the results you see here have been generated following a search in the Swahili to English side. This is a novel type of electronic dictionary we are experimenting with, one that shows how senses in one language are spread all over the lexicon in another, and how these then again spread out, etc.

At the time of writing, there are around 6 500 articles in the Swahili dictionary, and around 11 500 items in the English search index. Each of the Swahili lemma signs treated was chosen for its (high) frequency in the corpus. Assuming that the current (in-progress) data would have been online since 'day one', one can simulate — using the *real* search queries, as the logs have been kept since the very beginning — which percentage of the dictionary is being returned as the number of searches grows. Taking a snapshot every one thousand searches, which at the current look-up rate roughly means 'one snapshot a day', the graph shown in Figure 1 is obtained.

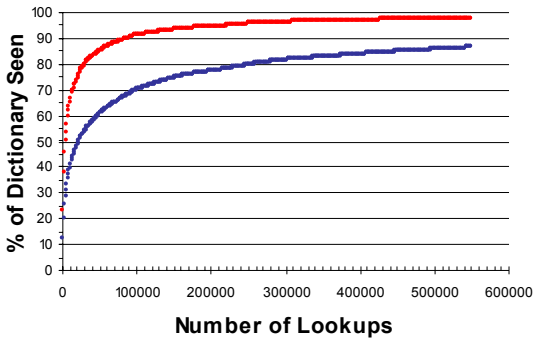


Figure 1: Percentage of dictionary returned ('seen') in function of the total number of searches (bottom: directly, top: with cross-references)

The bottom line indicates that over 86% (86.55%) of the material has been searched for directly, while the top line indicates that close to 98% (97.81%) of the dictionary data have been returned when one also includes the cross-referenced material. Looking at the trend of these lines, it should be clear that *all* dictionary data will indeed be seen over time.

Users search for far more than what is returned, of course, so one would like an indication of the 'hit rate'. Here one can again use the actual half a million searches done so far. The outcome is that 'only' 53.1% of the Swahili searches return one or more hits, while this value climbs to 68.5% for searches in the English index. These values are in line with what was observed for the online Northern Sotho–English dictionary. According to De Schryver and Joffe (oral communication at EURALEX 2004, 7 July 2004), only about 16% of all the misses are 'real misses'. This value was arrived at following a detailed study of each and every miss in their Northern Sotho dictionary. Real misses are those items that should/could have been in the dictionary, and these are basically easy to handle, as one must only make time to compile the necessary articles for them.

Based on all the above and taken at face value, therefore, it seems as if treating just the top-frequent orthographic words in a dictionary will indeed satisfy most users, and this in turn seems to indicate that a corpus-based approach to the macrostructural treatment of the 'words' of a language is an excellent strategy. This conclusion, however, is *not* correct, as will be shown in the next section.

5. The Relation between Corpus Ranks and Actual Dictionary Lookup Ranks

With over half a million real dictionary searches at one's disposal on the one hand, and with corpus-derived frequencies on the other, it becomes possible to calculate various correlation coefficients between the two sets of data. Reformulated, one can effectively take a corpus list of words, and compare that list word for word with actual dictionary searches, and/or one can take searched-for items in a dictionary, and compare those with the corpus. In a way, De Schryver and Joffe (2004: 190) already tried to look into this type of correlation when they sought to answer the following research question:

> 'Are the top 100 searches also the top 100 in a corpus?' If it would turn out that there is indeed a large overlap, this finding would provide substantial support for the practice of including or omitting lemma signs in a dictionary based on frequency considerations (and by extension for corpus-based lexicography in general).

Based on the fact that 30 of the top 100 Northern Sotho searches could also be found in the corpus top 100, while as many as 63 could be found in the corpus

top 1 000, they came to the conclusion that users indeed look up the frequent words of the language. While this observation is also true for the Swahili–English data at hand, it is only part of the story. It is and remains true that the top few thousand words of a language are also those that users most frequently look up, but the real question one wishes to answer is what happens beyond that point. In the bold words of Bergenholtz and Johnsen one would like to know whether there are indeed words that lexicographers should 'not have to waste time describing' as they are 'of no interest to the users'.

There are different ways to approach this question, but one of the most straightforward ones is as follows. In a two-dimensional plane, one could have the corpus data (as frequencies or ranks) on one axis, and the corresponding actual dictionary lookups (expressed as a count or also as a rank) on the other axis. If corpus-based lexicography indeed reflects (or rather 'pre-empts') what users look up (or 'will look up') in a real dictionary, then the most frequent word in the corpus should also correspond with the word most frequently searched for, the tenth most frequent corpus item should correspond with the tenth most frequent lookup, the one hundredth with the one hundredth, etc. In this ideal situation, the result would be a straight line out of the intersection of the axes in the two-dimensional plane. Allowing for (small) deviations, the straight line would turn into a 'scatter plot', with a cloud of dots 'around' the imaginary straight line. Mathematically, the straight line corresponds with a Pearson correlation coefficient of 1.0, while deviations result in lower values.

Before turning to the actual scatter plot, Table 1 lists the top 10 Swahili words, according to their frequency in the fifteen-million-word TshwaneDJe Swahili Corpus (TSC), and the ranks of these are contrasted with the lookup ranks derived from the actual searches made in the online Swahili dictionary.

Table 1: Comparing corpus ranks with dictionary lookup ranks for Swahili

Item	Corpus frequency	Lookup frequency	Corpus rank	Lookup rank
na	399 663	1 236	1	2
ya	384 813	781	2	6
wa	282 625	683	3	9
kwa	190 645	980	4	4
katika	104 859	472	5	17
za	88 488	244	6	57
ni	87 585	1 173	7	3
kuwa	70 267	469	8	18
la	68 857	239	9	59
hiyo	55 888	117	10	173

A brief look at this top 10 seems to indicate that the full scatter plot might indeed revolve around a straight line. However, and again using the ranks, the outcome of this exercise on a much larger scale as displayed in Figure 2, is at least highly surprising. Note that, given full orthographic words are entered/ treated in the online Swahili dictionary, these are compared with *un*lemmatised corpus data. In Figure 2, each dot thus represents the dictionary lookup rank versus the unlemmatised corpus rank of a particular word.

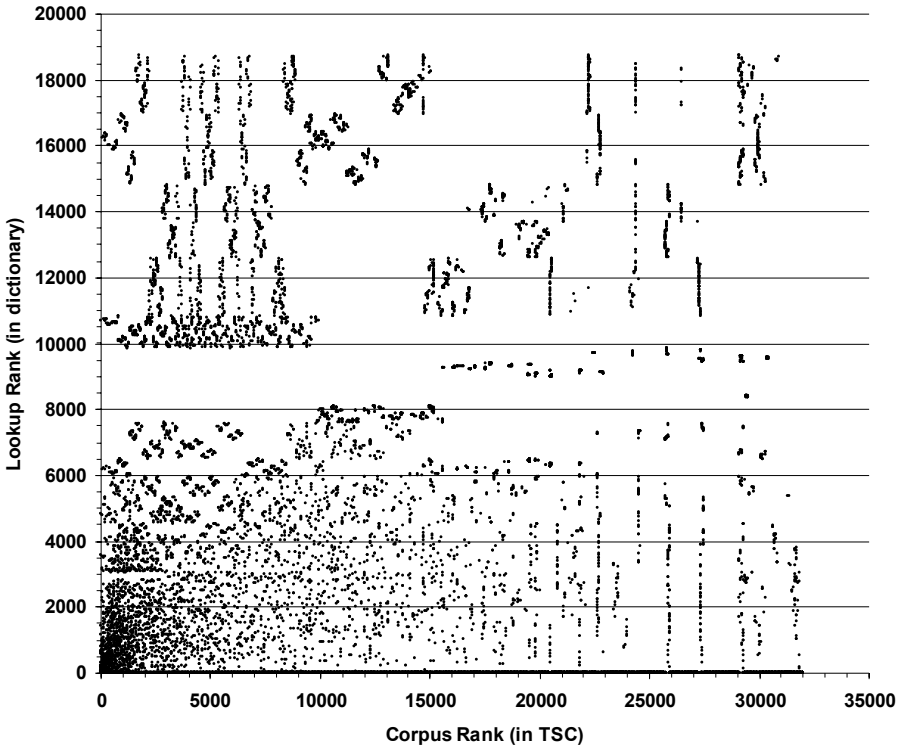


Figure 2: Ranks of the Swahili 'dictionary lookups' versus their corresponding 'corpus frequency' ranks in the TshwaneDJe Swahili Corpus (TSC)

These data for Swahili can immediately be contrasted with those for English. In Table 2, the top 10 English corpus words as seen in the one-hundred-million-word British National Corpus (BNC) are contrasted with their lookup occurrences in the English index, and Figure 3 displays the scatter plot for English. Note, again, that given the nature of the online Swahili–English dictionary, whereby users are 'allowed' to search for non-canonical dictionary forms in both directions, the *un*lemmatised corpus statistics were used in both Table 2 and Figure 3.

Table 2: Comparing corpus ranks with dictionary lookup ranks for English

Item	Corpus frequency	Lookup frequency	Corpus rank	Lookup rank
the	6 187 925	1 100	1	12
of	2 941 786	652	2	41
and	2 682 874	681	3	36
to	2 560 344	1 050	4	14
a	2 150 872	553	5	56
in	1 883 290	642	6	43
that	1 115 377	343	7	127
it	1 089 558	395	8	103
is	998 857	1 037	9	15
was	923 972	200	10	258

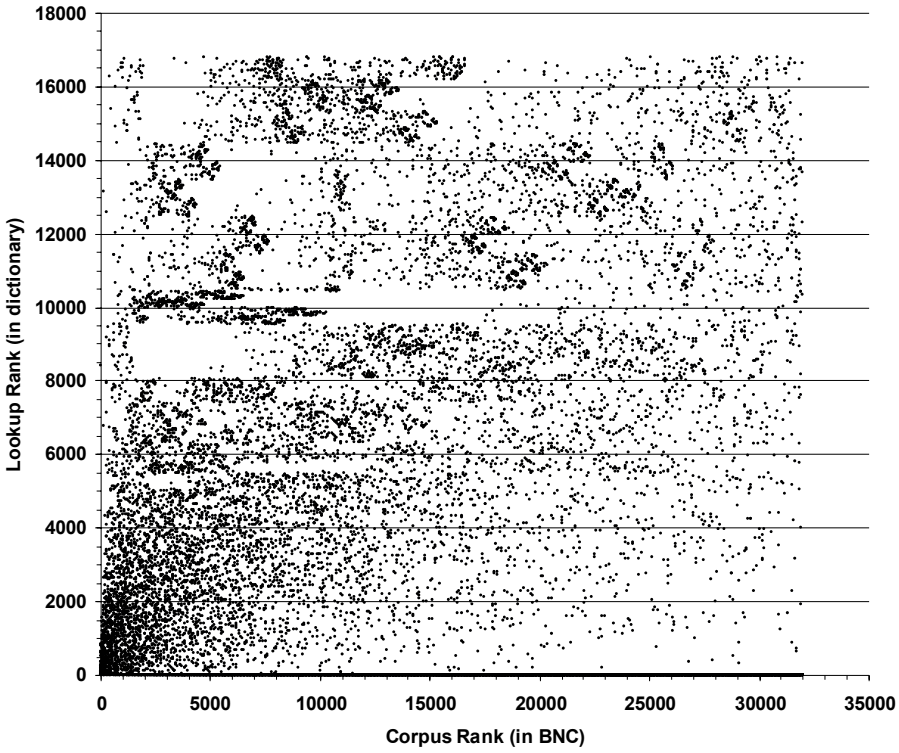


Figure 3: Ranks of the English 'dictionary lookups' versus their corresponding 'corpus frequency' ranks in the British National Corpus (BNC)

If one zooms in on the area around the intersection of the axes in Figures 2 and 3, or thus the top ranks, then one *does* see some kind of vague correlation, but as one moves along the axes, this correlation vanishes entirely. Actually, this too can conveniently be displayed. Figure 4 shows a graph plotting the Pearson correlation coefficient for Swahili word rankings, with each point being the correlation recalculated from 1 to the N*100th point (so each one 'includes' the previous one). Figure 5 shows the equivalent for English.

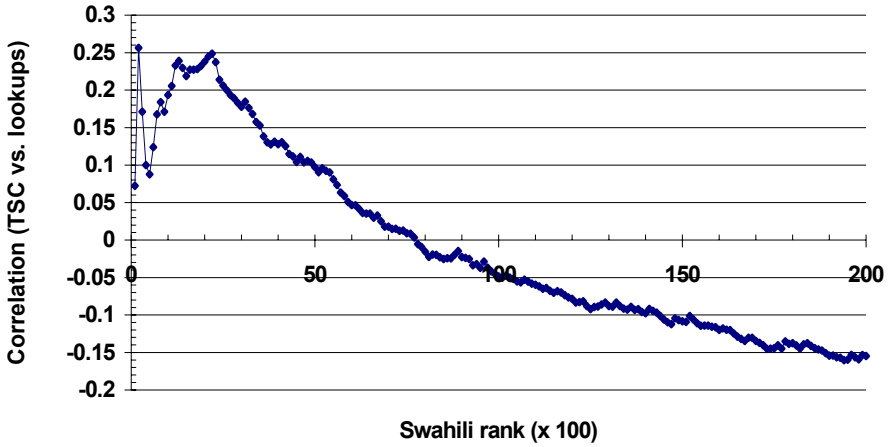


Figure 4: Correlation between corpus ranks and actual dictionary lookup ranks for Swahili (recalculated after every increase of the rank with one hundred)

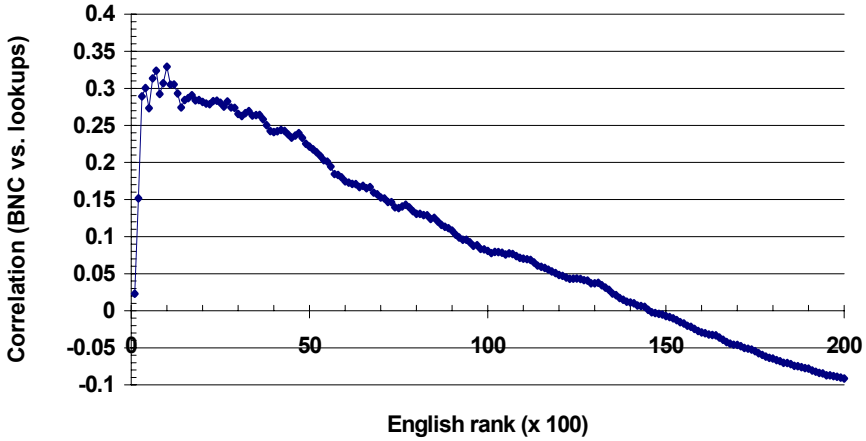


Figure 5: Correlation between corpus ranks and actual dictionary lookup ranks for English (recalculated after every increase of the rank with one hundred)

Figures 4 and 5 clearly reveal that there is indeed some minor correlation between corpus ranks and actual dictionary lookup ranks for the first few thousand words (up to around 3 000 for Swahili, and up to around 5 000 for English), but beyond that point there simply is no correlation whatsoever.

This is a hugely important — albeit shocking — revelation, as it means that it is simply impossible to 'predict' which words will be of interest to the dictionary user. Given the nature of the Internet, it is safe to assume that this dictionary user is a 'general dictionary user' with 'general needs'. To make this conclusion more tangible, take for example Figure 3 at the BNC rank 15 000, which could be the cut-off point for a dictionary with an upper limit of roughly fifteen thousand entries. Looking upwards from that point in Figure 3, it should be clear that it is unfortunately so that virtually any word may be looked up with any frequency at this cut-off point.

6. Additional Lexicography Software Modules

If one were to summarise the outcomes of the research so far — and against the background of other existing studies into (paper) dictionary use, an overview of which may be found in De Schryver and Joffe (2004: 187-188) — then one can make two statements:

(a) If one needs to prepare a small dictionary, for example a pocket school dictionary, with only a few thousand entries, then it is indeed good practice to base the selection of the lemmas on corpus data.

(b) If one needs to prepare a large dictionary, for example a large desktop dictionary, with several tens of thousands of entries, then the use of a corpus as an arbiter on what to include in and what to exclude from the dictionary makes little sense for all low-frequency lemmas.

These conclusions, then, pose great difficulties to lexicographers, as the corpus does not provide the 'magic answer' every dictionary maker was hoping for. When one compiles a small (school) dictionary one tends to 'throw out'/'skip' function words and so-called easy and basic words, but those are precisely the ones needed in such a dictionary. When one compiles a large (desktop) dictionary one has become accustomed to using corpus frequencies for selecting material, but it turns out that this is by no means a guarantee for look-up success. For want of any better/other approach at this stage, however, the corpus 'may' continue to be used, but as a guidance only.

Bringing these outcomes back to the online Swahili–English dictionary, and to conclude, it is obvious that 'progressing down the (unlemmatised) corpus frequency list' when selecting lemmas to be treated during compilation is not the way to go (anymore). Of course, in order to increase the hit rate (cf. the

16% 'real misses', mentioned in section 4) one must carry on with the addition of more lemmas. Instead of continuing the manual treatment of 'full orthographic forms', however, it seems more advantageous to call in an extra lexicography software module that could do some level of morphological analysis of those lookups that are not treated in the dictionary. Given a detailed and linguistically exhaustive decomposition is not needed for lexicographic purposes, the project team is currently experimenting with what could be termed 'clumped morphotactic decomposition'. Whenever a particular search is 'not found' in the dictionary, the clumped morphotactics module kicks in, and tries to decompose the search item. This is best illustrated with an example. An actual word that was searched for but not found in the past is the Swahili *yakaisha* '(and then) they stopped'. From Addendum 3 one can see how guidance is currently given in this regard. At the time of writing, around 100 'Swahili rules' have been stored in the 'clumped morphotactic decomposition' module, with which the Swahili hit rate has increased with another 4%.

From Addendum 1 it can be derived that multi-words have also been stored in the English index. Whenever such items have been stored, the respective Swahili article(s) is (are) simply offered to the user. All Swahili and English combinations can also directly be looked up. However, and as pointed out earlier (cf. section 4 above), users also increasingly search for phrases and entire sentences, and these are of course more often than not missing from the dictionary/index. In order to meet the dictionary user halfway in this regard, another lexicography software module was written that takes the input text, and when no matches are found, presents 'answers' (i.e. displays articles) for up to the first 10 words of an input string. This is illustrated in Addendum 4 for 'I love chicken' (which is an actual search string that was flagged as 'not found' earlier).

A third new lexicography software module that has been developed is an additional custom search index, which aims to re-route frequent misspellings to the most likely form. With this module the hit rate continues to climb. A further extension of this module is illustrated in Addendum 5. The dictionary team is rather sure that the gross majority of the dictionary users who look up *jambo* 'matter, affair, thing' are actually searching for the meaning of the pair *hujambo* 'how are you?' / *sijambo* 'I'm fine!'. Without further ado, by presenting the pair *hujambo/sijambo* together with *jambo*, the user will be in a position to distinguish between these forms, and will hopefully start using the non-corrupted forms.

Looking back, therefore, it is clear that it is simply impossible to know in advance which words users will want to look up in a large dictionary. Corpus frequencies do not predict look-up behaviour beyond the top few thousand words of a language. There is thus no such thing as words a lexicographer better not treat. Instead, and in an electronic environment, it will be more advantageous to add lexicography software modules that help increase the hit rate. As these modules may reuse the already compiled material, properly treating

and covering the top few thousand words of a language, however, remains an important core component of any reference work. Beyond that, lexicographers will have to be inventive.

Acknowledgements

Gilles-Maurice de Schryver and Sarah Hillewaert would like to extend their gratitude to Ari Kernerman, who made the 2003 *Kernerman Dictionary Research Grants* available to them for their three-year project 'The Creation of an Innovative Kiswahili–English Online Dictionary'. The work on this dictionary has since been transferred to Pitta Joffe and David Joffe, and the continued development is currently sponsored by *TshwaneDJe HLT*.

References

Bergenholtz, Henning and Mia Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes, Journal of Linguistics* 34: 117-141.

BNC. 1995–2006. British National Corpus [online]. Available: <http://www.natcorp.ox.ac.uk>.

De Schryver, Gilles-Maurice and David Joffe. 2004. On How Electronic Dictionaries are Really Used. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 187-196. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

De Schryver, Gilles-Maurice and David Joffe. 2005. Dynamic Metalanguage Customisation with the Dictionary Application TshwaneLex. Kiefer, F., G. Kiss and J. Pajzs (Eds.). 2005. *Papers in Computational Lexicography, COMPLEX 2005*: 190-199. Budapest: Linguistics Institute, Hungarian Academy of Sciences.

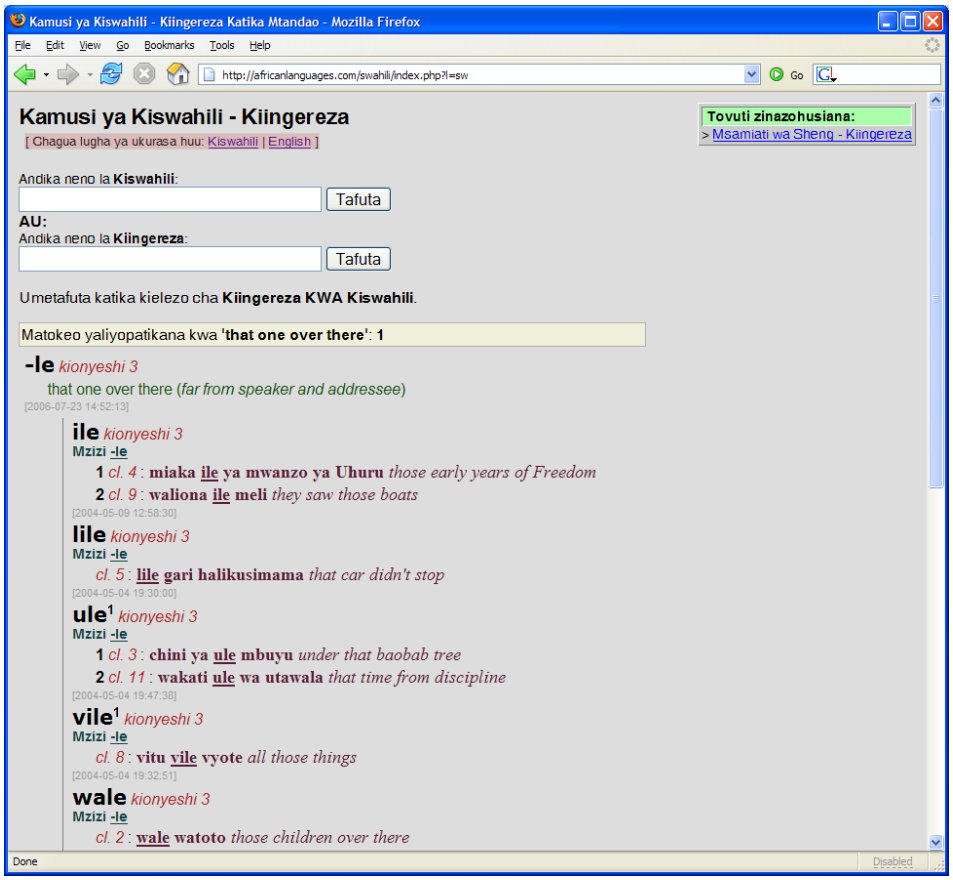
Google. 1998–2006. Google Search Engine [online]. Available: <http://www.google.com/>.

Hillewaert, Sarah, Pitta Joffe and Gilles-Maurice de Schryver. 2004–2006. *Kamusi ya Kiswahili–Kiingereza Katika Mtandao/Online Swahili–English Dictionary* [online]. Available: <http://africanlanguages.com/swahili/>.

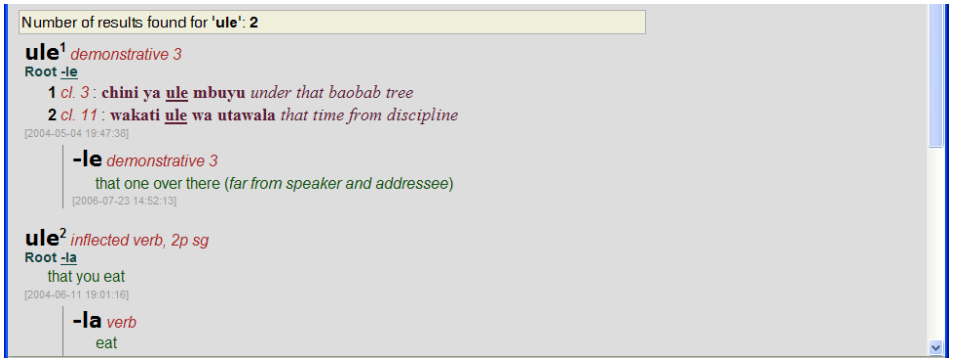
Sinclair, John McH. (Ed.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London: Collins ELT.

TshwaneDJe HLT. 2002–2006. TshwaneDJe Human Language Technology [online]. Available: <http://tshwanedje.com/>.

Addendum 1: Looking up 'that one over there' in the online English to Swahili index (with interface in Swahili)

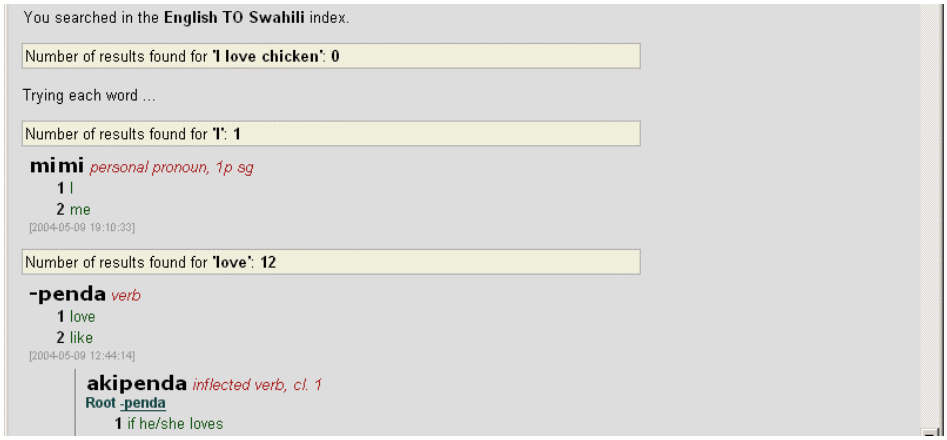


Addendum 2: Looking up *ule* in the online Swahili to English dictionary (with interface in English)



Addendum 3: Looking up *yakaisha* '(and then) they stopped' in the online Swahili to English dictionary (Note the morphological decomposition)

Addendum 4: Looking up 'I love chicken' in the online English to Swahili index (Note that each word is being handled separately)



Addendum 5: Looking up *jambo* 'matter, affair, thing' in the online Swahili to English dictionary (Note that also the pair *hujambo* 'how are you?' / *sijambo* 'I'm fine!' is shown, thanks to the re-router)