The Relationship Between Dictionary Look-up Frequency and Corpus Frequency Revisited: A Log-File Analysis of a Decade of User Interaction with a Swahili-English Dictionary

Gilles-Maurice de Schryver
gillesmaurice.deschryver@UGent.be
BantUGent, Ghent University, Belgium
&
Department of African Languages,
University of Pretoria, South Africa

Sascha Wolfer
wolfer@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Germany

Robert Lew
rlew@amu.edu.pl
Adam Mickiewicz University, Poland

ABSTRACT

In an earlier publication it was claimed that there is no useful relationship between Swahili-English dictionary look-up frequencies and the occurrence frequencies for the same wordforms in Swahili-English corpora, at least not beyond the top few thousand wordforms. This result was challenged using data for German by a different team of researchers using an improved methodology. In the present article the original Swahili-English data is revisited, using ten years' worth of it rather than just two, and using the improved methodology. We conclude that there is indeed a positive relationship. In addition, we show that online dictionary look-up behaviour is remarkably similar across languages, even when, as in our case, one is dealing with languages from very dissimilar language families. Furthermore, online dictionaries turn out to have minimum look-up success rates, below which they simply cannot go. These minima are language-sensitive and vary depending on the regularity of the searched-for entries, but are otherwise constant no matter the size of randomly sampled dictionaries. Corpus-informed sampling always improves on any random method. Lastly, from the point of view of the graphical user interface, we argue that the average user of an online bilingual dictionary is better served with a single search box, rather than separate search boxes for each dictionary side.

Keywords: lexicography; online dictionaries; log files; corpus frequencies; Swahili; English; language universals

BACKGROUND

COMPUTERS AND LEXICOGRAPHY

In recent decades, computers have revolutionized many aspects of our lives to a hitherto unseen degree. The change has not escaped lexicography (Lew & De Schryver, 2014), with dictionary publishers and users moving to the digital medium. Dictionary users appreciate the

affordances offered by digital dictionaries (Tan & Woods, 2008), which finally seem to be evolving away from their print predecessors (De Schryver, 2003) towards innovative digital tools embracing the new platform. At the same time, Lew and De Schryver (2014) note a shift in how users see the dictionary: no longer as a revered authority, but as a tool to use.

The fact that dictionaries are now increasingly offered and consulted in digital format is evident. Less evident to the uninitiated is the impact that computers have made to the very process of making a dictionary. It is this aspect, and more specifically the use of computer corpora in selecting what goes into the dictionary, that we focus on in the present contribution. We do so by using an approach to dictionary user studies inspired by the internet era and Big Data approaches: by investigating patterns of dictionary look-ups on a complete data set of dictionary searches, by any and all users over a massive period of ten years.

TEXT CORPORA AS DATA SOURCES FOR LEXICOGRAPHY

Dictionaries are now essentially seen as lexical tools designed to serve all sorts of users, informing them about various lexical aspects of language. How do dictionaries make sure that what they have to say about language is correct? Traditionally, dictionary-making has been based primarily on the intuitions of lexicographers. In the more ambitious projects this intuition was supplemented with manual excerptions placed on index cards, which contributed an element of objectivity, as long as the excerptions represented authentic use of language. In actual practice, the selection of material to be excerpted tended to be subjective and fragmentary. Yet perhaps the single most significant source for dictionary content was the content of existing dictionaries: copying from earlier entries was not only thought to be harmless, but in fact a display of good practice and diligence; much as we view research publications today, except with no concern (due to limited space) for acknowledging citations.

New lexicographic methods were made possible by the introduction of the computer into the process. A major turning point here was the COBUILD project (Sinclair, 1987). Foremost amongst its innovations was the introduction of a corpus of texts as the primary data to drive dictionary compilation. The COBUILD corpus was initially 7.3 million words, which by today's standards would be considered inadequate. In addition, it was far from being balanced, with journalistic texts and fiction overrepresented.

Despite these reservations, the COBUILD project is generally seen as revolutionary. The corpus was utilized for a number of purposes; the one of most immediate interest in the context of the present work was that corpus frequency was employed as the basis in decisions for inclusion amongst the lemmas of the future dictionary. Given that dictionaries should be useful, the assumption is that the more frequent vocabulary items (as evidenced by their uses attested in text corpora) would, on average, be the ones to generate the most interest from dictionary users, and thus would be looked up more often than words which are rare in the language. To generalize, the question of interest here is whether there is a positive relationship between corpus frequency and dictionary look-up frequency. In online dictionaries, the frequency with which specific items are looked up by users may be estimated through the examination of server log files, which record details of user visits.

LOG FILES AS A RECORD OF ONLINE DICTIONARY USER BEHAVIOUR

A log file for an online dictionary is a machine-readable, automatically generated record of the interaction of the user with the website-based dictionary. From the point of view of internet technology, the 'user' is not so much the human trying to use the dictionary, but, more directly, the browser client (or 'user agent') which the human user utilizes to communicate with the server hosting the dictionary.

eISSN: 2550-2131
ISSN: 1675-8021

Insofar as log files hold a systematic record of the interaction between the dictionary and its user, details contained therein are a potential source of information about the consultation behaviour of online dictionary users. Note that these files collect, amongst other information the input in the dictionary search box(es), and thus the focus is on what the users intend to look up, rather than the content of the dictionary proper. In other words, the online dictionary is merely used as a 'catch' for the study of dictionary look-up behaviour. In our case, we are dealing with two languages with very different grammatical structures, i.e. Swahili and English, which allows for a comparison of results across those two languages.

PREVIOUS STUDIES OF LOG FILES ATTACHED TO ONLINE DICTIONARIES

An early suggestion to utilize computer records of the interaction between a human user and a digital dictionary was made by Knowles (1983), though this method was not pursued except for isolated attempts involving records of dictionary look-ups generated by software installed locally on a PC. Other early suggestions include Crystal's (1986, p. 80) 'users of computerised dictionaries can have their procedures logged on the computer istself [sic]', while Abate (1985) suggested using such feedback to enhance aspects such as the structuring of data and access time in his projected dictionary's database of the future (De Schryver, 2003).

The real potential of log files came into focus with the introduction and growing importance of web-based online dictionaries, resulting in a series of studies of dictionary log files (Bergenholtz & Johnson, 2005; De Schryver & Joffe, 2004; De Schryver, Joffe, Joffe, & Hillewaert, 2006; Koplenig, Meyer, & Müller-Spitzer, 2014; Lemnitzer, 2001; Lorentzen & Theilgaard, 2012; Müller-Spitzer, Wolfer, & Koplenig, 2015; Schoonheim, Tiberius, Niestadt, & Tempelaars, 2012; Verlinde & Binon, 2010).

Lemnitzer (2001) may be seen as a pioneer study of this type. In this early study, logs from a small suite of bilingual dictionaries were used to identify a list of failed searches: strings that were searched by users but did not have a corresponding dictionary entry. A sample of 500 of those were subsequently manually classified, showing that most were misspellings, followed by actual words missing from the dictionary. De Schryver and Joffe (2004) focus on the distribution and nature of user lookups in an online Northern Sotho dictionary, reporting patterns across time and from specific regular visitors. The authors note users' particular interest in words frequent in the language on the one hand, and items of a sexual and offensive nature on the other. Of more central interest in the present context was the observation that 30 out of the 100 most frequent Northern Sotho searches were among the top 100 items according to corpus frequency. Bergenholtz and Johnson (2005) give the statistics for an online dictionary of Danish, reporting a failure rate of about 20 percent. These are discussed further in greater detail and with copious illustration. Log files for another dictionary of Danish are analysed in Lorentzen and Theilgaard (2012), and this is done prior to and following an update to the dictionary. The initial failure rate here is again reported at close to 20 percent, with the most frequent reasons being misspellings and typos, followed by unindexed inflected forms. After an update driven by these results, the proportion of search failures dropped to just 10 percent. The main goal of Schoonheim et al. (2012) was to gauge the effect of promoting a language game on the usage statistic of an online dictionary of Dutch.

All of these results are of both academic and practical interest, but our main focus in the present study is on the role of corpora in predicting what is likely to be looked up by dictionary users. Therefore, we now turn to the studies that have looked into the relationship between corpus frequency and look-up frequency.

The first in-depth examination of the relationship between look-up frequency and corpus frequency was De Schryver et al. (2006). The analysis was based on the log files

holding two years' worth of search strings entered in the search box of an innovative (wordform-based) Swahili to English dictionary (Hillewaert & De Schryver, 2004–), which also allowed 'inverse' searches in English using a different search box. Following up on their earlier observation (De Schryver & Joffe, 2004), the authors compute Pearson correlations between the ranks of search frequencies and ranks of corpus frequencies (in effect, these correspond to Spearman rank-order correlations on the raw frequencies), computing and plotting the correlation coefficients at increments of 100 ranks. Having done this for both Swahili and English items, the authors report low positive correlation values up to corpus frequency ranks of 3,000 (Swahili) and 5,000 (English), but no positive correlation for higher ranks. This finding is interpreted to mean that – as far as predicting user usage goes – corpus-derived frequencies are only of (limited) use for the few thousand most frequent items but represent no value for higher frequencies. Such an interpretation would put into question the rationale for a corpus-based methodology of identifying potential dictionary lemmas.

The findings of De Schryver and Joffe (2004) were subsequently replicated by Verlinde and Binon (2010). The latter authors, working with log files for the *Base lexicale du français*, found similar correlation coefficients between corpus frequency and look-up frequency, never in excess of 0.3. Few details are given in the paper, but the bar chart included (p. 1148) appears to suggest a tapering of the correlation above the rank of about 3,000.

These two studies thus suggest that corpus frequency is a relatively poor predictor of the frequency of word look-up. Such negative findings would appear to deal a rather serious blow to a central tenet of the mainstream corpus-based methodology characterizing much of modern lexicography, and so the claim warranted further corroboration. A detailed re-examination of the issue, this time using two dictionaries of German, followed in Koplenig et al. (2014), and then the related Müller-Spitzer et al. (2015). In these two contributions, the authors argue that the correlation coefficient adopted as a measure of the relationship between corpus frequency and look-up frequency may distort the picture, since it assumes a linear relationship between the variables of interest, which is not a realistic model here. Furthermore (and perhaps more importantly), in any set of data listing word occurrences or look-ups, there will be a long tail of numerous *rare events*, in particular one-time look-ups of very rare words, and these numerous data points will mask any systematic relationship of interest. The two newer studies propose, and subsequently adopt, an alternative simulation-based approach, whereby virtual dictionaries of varying sizes are generated, and their lemma lists are checked against corpus data. Employing this methodology, the studies paint a more positive picture of the corpus-based approach to selecting lemma candidates: a consistent and non-trivial effect of corpus frequency is noted. Another study covering three years and nearly 30 million lookups in an online Danish dictionary (Trap-Jensen, 2014; Trap-Jensen, Lorentzen, & Sørensen, 2014) used a similar approach, and found corpus-based frequencies to be a good positive predictor of look-up frequency, up to at least the top 100,000 lemmas, but most strikingly for the top 20,000 lemmas. The present study attempts to replicate the above results on log files spanning ten years of look-ups in an online Swahili-English dictionary: the same dictionary that was used by De Schryver et al. (2006), but covering a much longer period.

THE STUDY

RESEARCH QUESTIONS

Our main research question is: *To what extent does corpus frequency predict dictionary look-up frequency?* The direction of our online dictionary is from Swahili to English, but it also

makes sense to consider an inverse virtual dictionary going from English to Swahili. Therefore, a further question we would like to address is: *Is the predictive power (i.e., corpus-frequency predicting look-up frequency) language universal?*

THE SWAHILI-ENGLISH ONLINE DICTIONARY

Swahili (or Kiswahili in the language) is a Bantu language spoken by up to 100 million first- and second-language speakers. It is *the* lingua franca of East Africa, spoken in especially Tanzania and Kenya, but also in the neighbouring countries to their west and south (Mohamed, 2009, pp. iv-v). The existing lexicographic output for Swahili is the result of a century-and-a-half-old *craft*, overwhelmingly in traditional paper format (De Schryver, 2018). One notable exception is the online dictionary used in the present article. The front page of this dictionary is at http://africanlanguages.com/swahili/ (up since 13 May 2004). There is also an alternative (mirror) search page at http://www.goswahili.org/dictionary/ (up since March 2011). The search results of both are logged into the same database.

This dictionary is, at its core, a unidirectional Swahili-to-English dictionary aimed at general users, with an English index which allows visitors to search the entire microstructure and thus also to return Swahili equivalents for English (*inverse*) searches. This dictionary is 'special' in that the macrostructure for Swahili systematically includes both the lemmatised as well as all the unlemmatised orthographic wordforms that are frequent. An example is shown in **Figure 1**, whereby a user first searched for the English verb 'say' and then clicked on the cross-reference link to the Swahili root *-sema*.

FIGURE 1. Search in the online Swahili-English dictionary

As one may see from the screenshot in **Figure 1**, once one clicks on a root of a word, all full orthographic wordforms that are 'derived' from it and which are included in the dictionary are also listed (in this case even including a deverbative noun, *msemaji* 'speaker; political spokesperson'). A traditional dictionary of Swahili will only include the verb

root -*sema* 'say, speak'; in this online dictionary the main purpose of including roots is actually to bring all related forms together in a convenient way (i.e., using a hub-and-spoke model, with roots the hubs, and the cross-references from all the related full orthographic words the spokes). Of course, this automatically also means that this dictionary caters for users who wish to look up roots as in traditional dictionaries of Swahili.

To use this dictionary, little to no knowledge of Swahili grammar is therefore required; users can simply search for words the way they find them in written form or the way they hear them spoken. Examples from **Figure 1** include: *akasema* 'he/she said, he/she was saying'; *atasema* 'he/she will say; he/she will speak'; *hasemi* 'he/she does not say; he/she does not speak'; *husema* 'always says; usually says; always speaks; usually speaks'; *inasema* 'it says'; or *waliosema* 'they who said; they who spoke'. For the inverse English index, this further means that most morphological forms of English lemmata may be found, in this case 'say, says, saying, said'. Reformulated, the English index thus also contains both lemmatised as well as unlemmatised forms. Users of the dictionary quickly realise this, and indeed search for orthographic words as spoken or found in texts; or in corpus terms, as found in an unlemmatised corpus.

Over the first ten years the dictionary has undergone a number of changes:[1]

1. **more language data** was included (these incremental updates were only effected during the first few years, see Addendum 1)

2. auto-broken up **sentence search** was introduced (2006-07-29); this means that multi-word queries were now automatically partitioned into word-length strings, which were subsequently followed up in the dictionary

3. basic **morphological decomposition** was added (2006-07-31) [2 & 3 may apply at the same time]

4. frequent **misspellings** were **re-routed** to the most likely form (2006-10-25: 71 English-to-Swahili, 5 Swahili-to-English; 2009-05-22: 154 English-to-Swahili, 9 Swahili-to-English)

5. **Google adds** were added (2006-11-05)

6. the **feedback form** was taken off (2007-01-22)

7. the dictionary content was offered as a **downloadable dictionary** (around 2009-02-19)

8. the previous two language-specific **search boxes** were replaced with a single combined search box (2009-05-23)

In order to be able to interpret the search data correctly, one also needs to know that:

1. **logs** are **unavailable** for the periods 2009-05-23 – 2009-06-26 and 2011-12-06 – 2012-03-07 (see **Figure 2**)
2. two attempts at **harvesting** (i.e., stealing) the complete dictionary **content** may be identified, and these automated 'searches' thus have to be removed from the logs (see **Figure 2**); the first (from Russia, successful) happened on 2011-04-29, the second (from Lyon, failed) on 2014-06-05

eISSN: 2550-2131
ISSN: 1675-8021

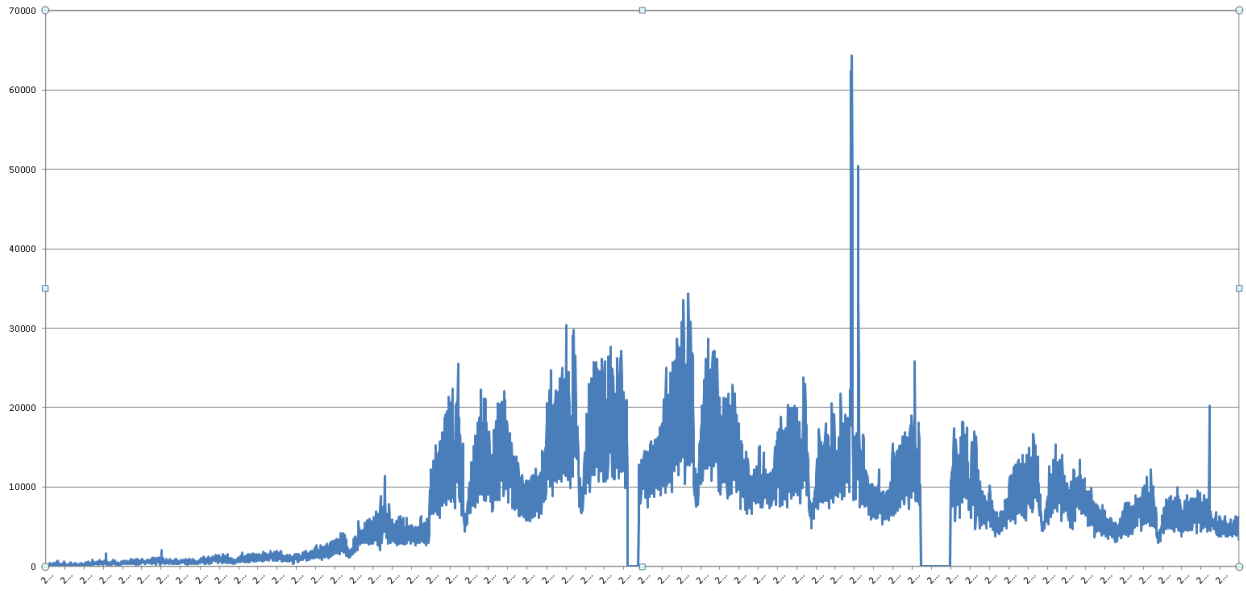

FIGURE 2. Number of overall searches per day over the first ten years in the online Swahili-English dictionary

FORM AND CONTENT OF THE SERVER LOGS

Our complete log files include the following ten fields:

0. *Visitor ID*
1. *Localization language selected,* with the following values: blank/en/sw; note that this is *not* equivalent to the search language, but rather refers to the page metalanguage
2. *Lookup type*, which could take any of the following six values:
   - left blank for a default Swahili-to-English single-word search in Swahili;
   - 'inverse' for an English-to-Swahili search;
   - 'word' for a component word of a Swahili-to-English multi-word search; for a search containing a string of multiple words, that string was broken down into single-word components;
   - 'inv_word' for a component word of an English-to-Swahili multi-word search;
   - 'link' for a cross-reference click;
   - 'decomp' for a match in the basic Swahili morphological decomposition search engine; for example, a search for *ukalia* returns no direct hits, but shows matches for *uka-* 'and then you' and the existing entry *-lia* 'cry/weep/despair/lose hope'.
3. *IP address*
4. *IP address to hostname lookup*
5. *Search term*
6. *Number of hits for the search term*
7. *Timestamp*
8. *Site ID*
9. *User Agent string* (identifying the Web browser)[2]

These complete records were subsequently reduced by deleting items 0, 8, and 9 for further analysis.

THE DICTIONARY LOGS

The logs analysed cover the first ten-year period, between 2004-05-13 and 2014-09-06, and include data from about 30 million searches. Halfway, i.e. after about 5 years (on 2009-05-23), the search interface was modified, from then on offering one combined search box for searches both in Swahili and English. As a consequence, all log files after that date do not mark the source language (however, it is also true that users did not always respect the intended dictionary side, so some of the source language information in the logs will still be incorrect). Should one wish to identify the language of the look-up from the unified search-box period, then the options available would be to match the search term against:

1. a corpus of English, Swahili, or both; or
2. the search terms for English and Swahili obtained from the data in the separate-search-boxes period.

Whatever strategy is adopted, some search items will inevitably not be identifiable as uniquely English or Swahili, as the words may exist in both languages, or neither. For the ones present in both languages, look-up frequencies may be:

1. assigned to neither; or
2. assigned randomly; or
3. assigned proportionally based on the respective corpus frequencies in both languages; or
4. assigned to one language based on the respective corpus frequencies in both languages.

In our analysis we followed option 4 (see the Section **Heuristic for the combined searches**). Whenever a search request was entered into the combined search box, we first checked whether the corresponding token is present in only one of the two corpora. If this is the case, the search request is counted for this language. If both corpora contain the token from the search request, the language with the higher frequency 'wins' and the search request is assigned accordingly. In very few cases (0.088% of all search request tokens), the frequency figures from both corpora are identical. We excluded these from the dataset. After associating a search request from the combined search box to one of the languages, we did not distinguish between the different kinds (i.e. coming from a single-language search box or the combined search box) of Swahili or English search requests anymore.

MULTI-WORD LOOKUPS

The dictionary engine breaks down all multi-word lookups into individual orthographic words and looks up all these components. Log file lines that originate from this 'auto sentence breakdown' are identified as such in the files. These lines have to be processed separately or excluded altogether. We should not include them or at least not mix them up with single-word look-ups, as that would, for example, inflate the look-up frequencies for function words. We could possibly study the differences, although they seem rather predictable (multi-word searches will be more like pieces of actual text). For our present analysis, multi-word searches were not included.

GEOGRAPHICAL INFORMATION

The IP addresses in the logs have been replaced with unique non-identifying numbers in order to comply with the privacy regulation in force since the GDPR.[3]

eISSN: 2550-2131
ISSN: 1675-8021

LISTS OF WORDS LOOKED UP

The lists of the looked-up words for Swahili and English were directly inferred from the processed server logs as described in the Sections **Form and content of the server logs** and **The dictionary logs**. In these logs, each row in the file corresponds to one search request. To carry out the analyses presented below, we had to transform this table to a 'type-based' format, i.e. one line per search term with the associated number of searches over a period of ten years and the language of the search term. The language of the search term was determined by the type of search box the users used (Swahili or English) or, after the interface changed to a combined search box, by the procedure we presented in the Section **The dictionary logs**.

CORPUS FREQUENCY LISTS

Given that our primary research question is whether occurrence of a word in every-day language is related to the look-up frequency of this word in a dictionary, we operationalized the occurrence of a word by corpus frequencies.

SWAHILI

An in-house corpus of 22 million words (22,030,608 tokens) was used, from which the different orthographic words, about half a million of them (522,132 types), were extracted. Most of the corpus material came from the Internet, but is not necessarily online anymore (especially not the data from the 1990s), and about 50 books were included as well.

ENGLISH

An unlemmatized (wordform) wordlist for English was generated, covering the top 200,000 most frequent entries. This list was generated using the SketchEngine (Kilgarriff, Rychlý, Smrž, & Tugwell, 2004) from the very large English enTenTen12 corpus, measuring 11 billion words.[4]

To minimize dependence on corpus size, a random sample was drawn from this enTenTen12 corpus the size of the Swahili corpus, so that the Swahili and English corpora for the analysis would be equal in size, while the English sample would retain the relative lexical frequencies of the original enTenTen12 corpus.

ANALYSIS AND RESULTS

Our primary research question, whether search frequency can be predicted by corpus frequency, has already been answered in the affirmative by Koplenig et al. (2014) and was elaborated on by Müller-Spitzer et al. (2015) for the German Wiktionary and the *Digitales Wörterbuch der Deutschen Sprache* (DWDS). However, this does not automatically imply that the same holds for the dictionary and languages we are looking at in the present article. To keep things comparable between analyses, we replicate the methods used by Koplenig et al. (2014) as closely as possible. Also, given that the online Swahili-English dictionary mainly entered full orthographic words as entries (see the Section **The Swahili-English online dictionary**), rather than word stems only, comparisons may and should be made with corpus items as seen in *unlemmatised* corpora.

INITIAL DATA PREPARATION

As a first step, we aggregated the raw log files. This led to a data set where each search term is associated with the number of searches over the whole period covered by the log files. All search requests longer than 80 characters (0.074%) and those starting with a hyphen were excluded (0.035% for English search requests, 0.21% for Swahili).[5] Further, all search

requests containing numbers and special characters[6] were excluded from the analysis (6.98% for English, 6.73% for Swahili). Since we are only interested in unigram search requests at the moment, we also excluded all search requests containing the space character (44.5% for English, 17.0% for Swahili). Each search term is then associated with the corpus frequency from the corpus frequency list (see the Section **Corpus frequency lists**) for the appropriate language (that is, the English frequency list for English searches, and the Swahili frequency list for Swahili searches).

Koplenig et al. (2014) introduced the notion of *search requests per one million search requests* ('poms') to keep the number of search requests comparable between different dictionaries and/or between different types of searches within one dictionary. To get the variable search requests *poms*, we multiplied the raw frequency of a query by 1,000,000 and divided by the number of all query tokens, rounding the resulting figure to the nearest integer. Following Koplenig et al. (2014), we then assigned the search terms to the categories 'regularly', 'frequently', and 'very frequently' based on its *poms*. We assigned a search term to the category 'regularly', if it received one or more per million search requests. If a term had two or more than two search requests *poms*, it was assigned to the category 'frequently'. For terms with more than ten search requests *poms*, we used the category 'very frequently'. In addition, the more frequent categories included the less frequent ones, so that any 'frequently' searched term was by definition also searched 'regularly', and a term searched for 'very frequently' was also included in the categories 'frequently' and 'regularly'. Koplenig et al. (2014) used this strategy to address several problems associated with the bivariate distribution of search-term frequencies and corpus frequencies. For example, the relationship between the two variables of interest is clearly not linear, which is a problem for ordinary least squares (OLS) regression. (Log-)Transforming the variables does not resolve this problem. To make things worse, we are dealing with a large number of rare events (LNRE), both for the search-term-frequency and the corpus-frequency distributions. Given the underlying distributions, Spearman's rank correlation coefficient is also not a good solution, because it assumes more or less equidistant ranks: an assumption that is not satisfied for either the search term frequencies or for corpus frequencies.

We will first present data for the Swahili search terms (*standard* search) of the Swahili-to-English dictionary before turning to the English search terms (*inverse* search).

SWAHILI-TO-ENGLISH SEARCHES

After preparing and selecting the data as presented in the previous section, the final data set included 711,987 Swahili-English search request types and 14,572,388 search request tokens. **Table 1** shows the number and ratios of search request types for the three searches *poms* categories. Keep in mind that whenever a search request is categorized as 'very frequently', it is also categorized as 'frequently' and 'regularly'. The table shows implicitly that 86.46% (100% minus 13.54%) of all search request types are searched for less than 0.5 times per million (everything above 0.5 is rounded to 1 and thus counts as regularly searched for).

TABLE 1. Number and ratio of search request types for the Swahili search for the three searches *poms* categories

Searches *poms* category	no. search request types	% search request types
regularly (>= 1 searches *poms*)	96,373	13.54
frequently (>= 2)	48,572	6.82
very frequently (>= 11)	12,633	1.77

For our primary research question – whether words that are frequent in general language (as measured by corpus frequencies) are searched for frequently in the Swahili-English dictionary – we needed to select those search request types that have frequency

information available. For the standard search direction, this is the case for 261,790 search request types (36.8%). These types comprise 86.3% of the search request *tokens*. This means that 13.7% of all search request tokens had to be excluded due to missing frequency figures.

This exclusion ratio is due to two major factors. First, unlike for English, German, or any other 'high-resource language', suitable corpora of Swahili that would allow us to generate similarly comprehensive frequency lists are simply not available. Second, Swahili, like all Bantu languages, is agglutinative in nature, in that especially verb forms may have hundreds of inflections. Third, and most significantly, many dictionary users seem to have misunderstood the standard search function and have entered search terms in another language than in Swahili. We went through a randomly sampled list of 100 search requests in the Swahili search box that were not found in the Swahili frequency list. Only *two* of these requests were legitimate Swahili search requests, while 41 were queries that could be identified as coming from another language (plus 8 misspelled foreign-language queries), 24 were proper names, 19 were misspelled Swahili requests and 6 were nonsense requests. See Addendum 2 for the details. From the above it seems quite clear that the vast majority of items without a match in the frequency list are, indeed, not legitimate Swahili items.

To answer our question if frequent words (in terms of occurrences in the corpus) are also often searched for, we start off with two simple tests: first we look at a number of actual corpus occurrences vs. actual dictionary searches, and next we present a simple visualisation of all the actual data. Immediately below, we show the three most frequent words in our 22m Swahili corpus, and compare this top 3 with the actual search ranks:

- *na* 'and, with' is the most frequent word in our Swahili corpus. In terms of dictionary look-ups we note:
  - 62,920 searches
  - 4,303.2718 searches per million
  - Search rank: 2
- *ya* 'of' is the 2nd most frequent word in our Swahili corpus. In terms of dictionary look-ups we note:
  - 61,347 searches
  - 4,195.6900 searches per million
  - Search rank: 3
- *wa* 'of' is the 3rd most frequent word in our Swahili corpus. In terms of dictionary look-ups we note:
  - 31,288 searches
  - 2,139.8723 searches per million
  - Search rank: 6

Quite astonishingly, the corpus top 3 corresponds to ranks 2, 3 and 6 in terms of dictionary look-ups.

This exercise may also be turned around, comparing the top 3 Swahili searches with the corresponding corpus ranks:

- *i* 'I / (i)' is the most searched-for 'Swahili' word. In terms of our Swahili corpus we note:
  - Frequency: 13,032
  - Frequency per million: 591.5406
  - Frequency rank: 513
- *na* 'and, with' is the 2nd most searched-for Swahili word. In terms of our Swahili corpus we note:
  - Frequency: 852,676
  - Frequency per million: 38,704.1520
  - Corpus frequency rank: 1
- *ya* 'of' is the 3rd most searched-for Swahili word. In terms of our Swahili corpus we note:
  - Frequency: 783,545
  - Frequency per million: 35,566.1995
  - Corpus frequency rank: 4

eISSN: 2550-2131
ISSN: 1675-8021

Clearly, there is a problem with what is apparently the most searched-for 'Swahili' word, which is not Swahili but mainly the English first-person-singular pronoun 'I'. This problem is likely the result of the design of the early version of the graphical user interface — for which, see the Section **Heuristic for the combined searches**. Search ranks 2 and 3 correspond with corpus ranks 1 and 4, which looks excellent.

As a further illustration, a few of the first Swahili words that are only 'regularly' and 'frequently' searched for, but not 'very frequently' are:

- *jama* EITHER the (family) name *Jama*; OR short for *jamani* = interjection for drawing attention or for expressing wonder
  - 153 searches
  - 10.46409 searches per million
  - Corpus frequency rank: 24,590
- *tutu* EITHER the family name *Tutu*; OR 'pimple'
  - 153 searches
  - 10.46409 searches per million
  - Corpus frequency rank: 40,057
- *piki* short for *pikipiki* 'motorcycle'
  - 153 searches
  - 10.46409 searches per million
  - Corpus frequency rank: 63,739
- *nitaku* [is only part of a word, as a verb root is missing] 'I will [verb] you'
  - 153 searches
  - 10.46409 searches per million
  - Corpus frequency rank: 169,440

A few of the first Swahili words that are only 'regularly' searched for, but not 'frequently' and 'very frequently' are:

- *Swaziland* 'Kingdom of eSwatini'
  - 21 searches
  - 1.436248 searches per million
  - Corpus frequency rank: 11,948
- *vilo* typo for vile = 'those'
  - 21 searches
  - 1.436248 searches per million
  - Corpus frequency rank: 20,3061
- *yakwamba* 'that'; *ya kwamba* = 'mostly'
  - 21 searches
  - 1.436248 searches per million
  - Corpus frequency rank: 99,556

As one may see, once beyond the top ranks, comparisons on the level of individual words are not very meaningful, hence the need for a simple visualization of the two factors at play (i.e. corpus frequency and search frequency), which we both categorized into three rough bins each. **Figure 3** is a bar plot that shows how the two factors cross for Swahili.[7] Specifically, the proportion of words that are very popular in searches (red bars) is largest for the high-frequency words, but smallest amongst the low-corpus-frequency items. Conversely, the proportion of words that are very unpopular in searches (yellow bars) is largest for the low-frequency words, but smallest amongst the high-frequency items.

eISSN: 2550-2131
ISSN: 1675-8021

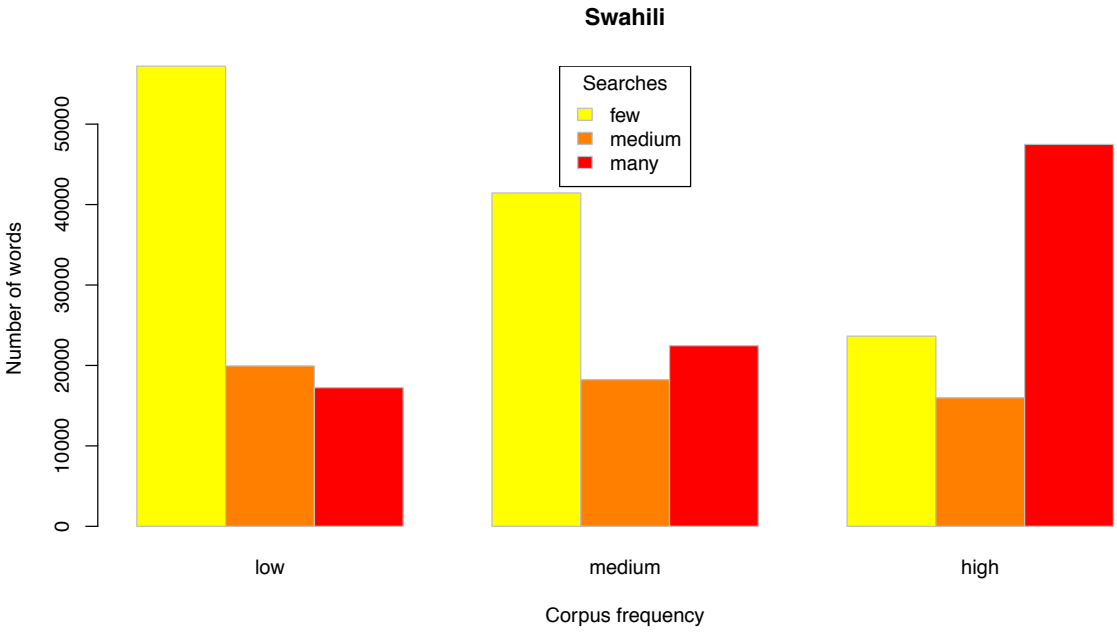

FIGURE 3. Numbers of Swahili words coming from three corpus frequency bands that register three levels of search intensity

Now, for a more fine-tuned analysis, we employ a strategy whereby incremental virtual dictionaries are created on the fly, using the corpus frequency list, and simulating look-ups in those virtual dictionaries of all the terms from the search requests in the log files, in each case noting the inclusion or otherwise in the virtual dictionary. Thus, in line with the tenets of corpus-based lexicography, the corpus frequency rank of the search term determines if a word is included in the dictionary or not. At each step, more and more words from the top of the frequency list are included in the dictionary. If corpus frequency were unrelated to search frequency, the number of entries included in the virtual dictionary should have no effect on the proportion of words that are searched for regularly, frequently and very frequently. If, however, corpus frequency and search frequency were indeed related, we would expect higher inclusion rates of regularly, frequently and very frequently searched-for words for a virtual dictionary including a smaller number of items coming from the top of the frequency list. With a greater number of lower-frequency items included, fewer words should be searched for regularly, frequently and very frequently.

Figure 4 shows that this is, indeed, the case. For example, as long as only the top 50 search requests from the corpus frequency list are included, **all** of these entries are searched for regularly, frequently and very frequently. Even for the top 1,000 corpus frequency ranks, the figures are still very high (100% of the entries are searched for regularly, 99.6% frequently and 92.4% very frequently). However, as more corpus frequency ranks are included, these figures decline and deviate from one another. For example, if the top 30,000 search requests in terms of corpus frequency are included, 75.1% are searched for regularly, 58.6% frequently and 26.7% very frequently.

eISSN: 2550-2131
ISSN: 1675-8021

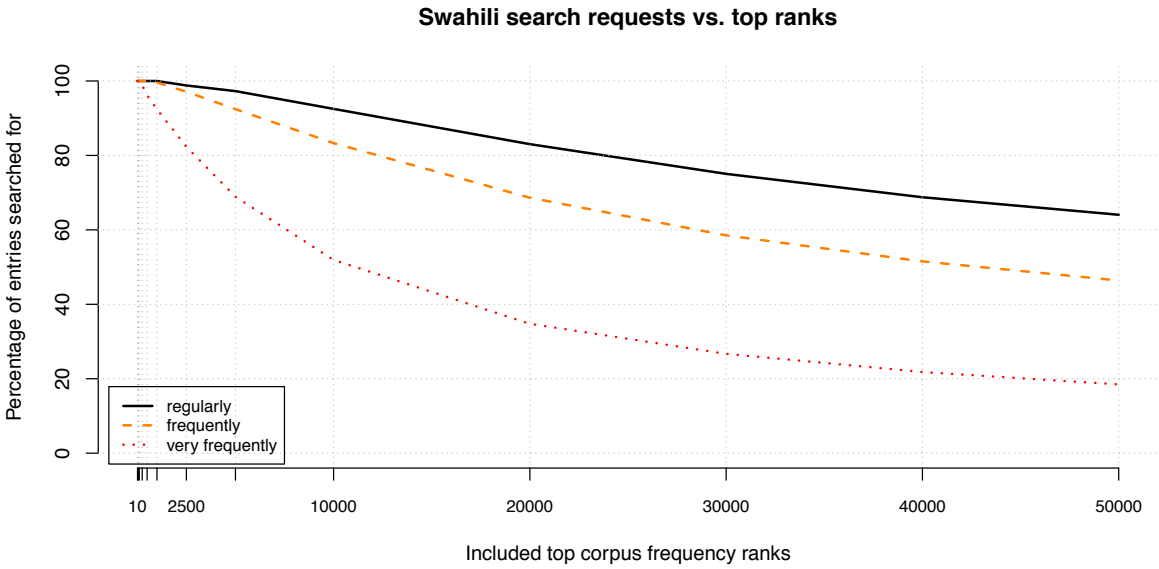

FIGURE 4. Relationship between the number of included frequency ranks from the top of the wordform frequency list for Swahili and the percentage of entries that are searched for regularly, frequently and very frequently in Swahili

Such a declining pattern is in sharp contrast to randomly sampled dictionaries of varying sizes (**Figure 5**). For up to 100 entries, some fluctuation due to the sampling process can be observed. Upwards from 250 entries, however, the ratios for randomly sampled dictionaries stabilize around 30% for regularly searched-for entries, 17% for frequently searched-for entries and 5% for very frequently searched-for entries. Whatever the size of the virtual dictionary, these ratios are lower for randomly sampled dictionaries than for dictionaries based on a corpus frequency list (**Figure 5** vs. **Figure 4**, respectively). These differences, very substantial, represent the advantage of using a frequency list to select words for dictionary inclusion.

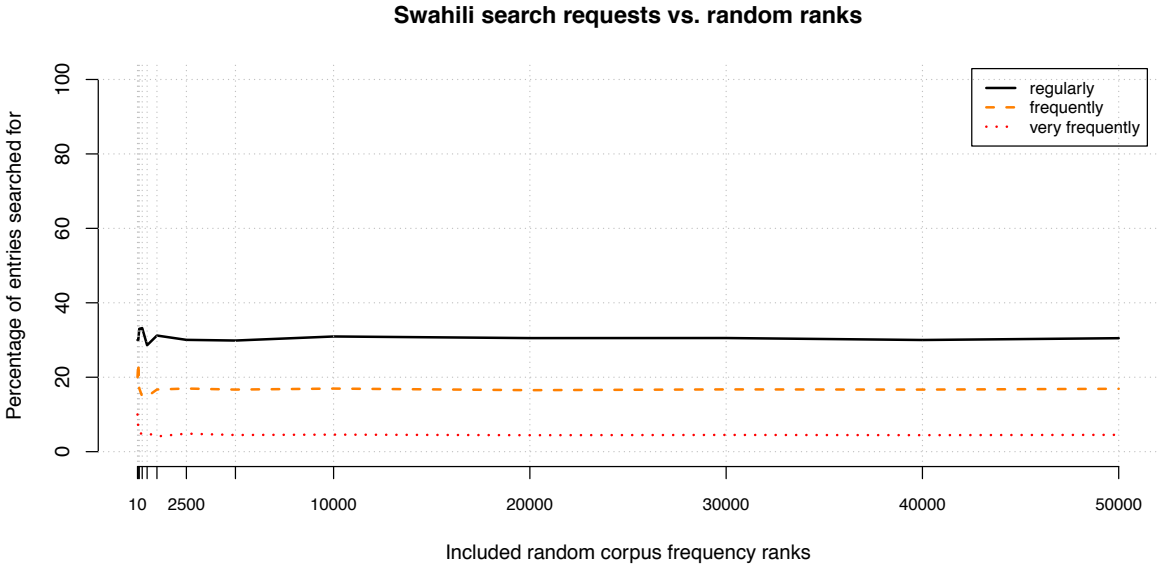

FIGURE 5. Relationship between the number of items *randomly drawn* from the Swahili wordform list and the percentage of entries that are searched for regularly, frequently and very frequently in Swahili

eISSN: 2550-2131
ISSN: 1675-8021

Another approach, used by Müller-Spitzer et al. (2015), also involves virtual dictionaries, but generated somewhat differently. This approach was inspired by the statement in De Schryver et al. (2006, p. 79) that '[c]orpus frequencies do not predict look-up behaviour beyond the top few thousand words of a language'. To test this hypothesis with the Swahili-to-English dictionary, we excluded the top 5,000, and then the top 10,000 items from the corpus frequency list. For both data sets of remaining items, we created two types of virtual dictionaries: a dictionary that consists of the next 5,000 or 10,000 items in the corpus frequency list; and another dictionary that is randomly sampled from among the rest. If corpus frequency were unrelated to search frequency beyond the top few thousand words, rank-based and random dictionaries should perform equally well with regard to frequently and very frequently searched-for entries (to keep the presentation clearer, we did not include the 'regularly searched-for' category in this analysis).

Figure 6 shows that this is clearly not the case. The dictionaries based on frequency perform clearly better than the randomly sampled dictionaries, even though the top 5,000 or 10,000 corpus frequency items are excluded. This holds both for frequently-searched-for words (orange) and words that are very frequently searched for (red shades). For example, in a dictionary that consists of the corpus frequency ranks 5,001 to 10,000 (left-most bar), 74.2% of the entries are frequently searched for. A virtual dictionary of 5,000 entries that has been randomly sampled among all the items below rank 5,000 in the frequency list only covers 15.3%[8] of the search terms. Additionally, given that a dictionary of as many as 10,000 entries but which excludes the top 10,000 corpus frequency ranks, covers fewer search terms than a dictionary which contains only half as many entries but which excludes the top 5,000 corpus frequency ranks — see the rank-based and random bars for '5,000 excluded' vs. the rank-based and random bars for '10,000 excluded' in **Figure 6** — this simulation again illustrates the advantage of using top frequencies for dictionary compilation.

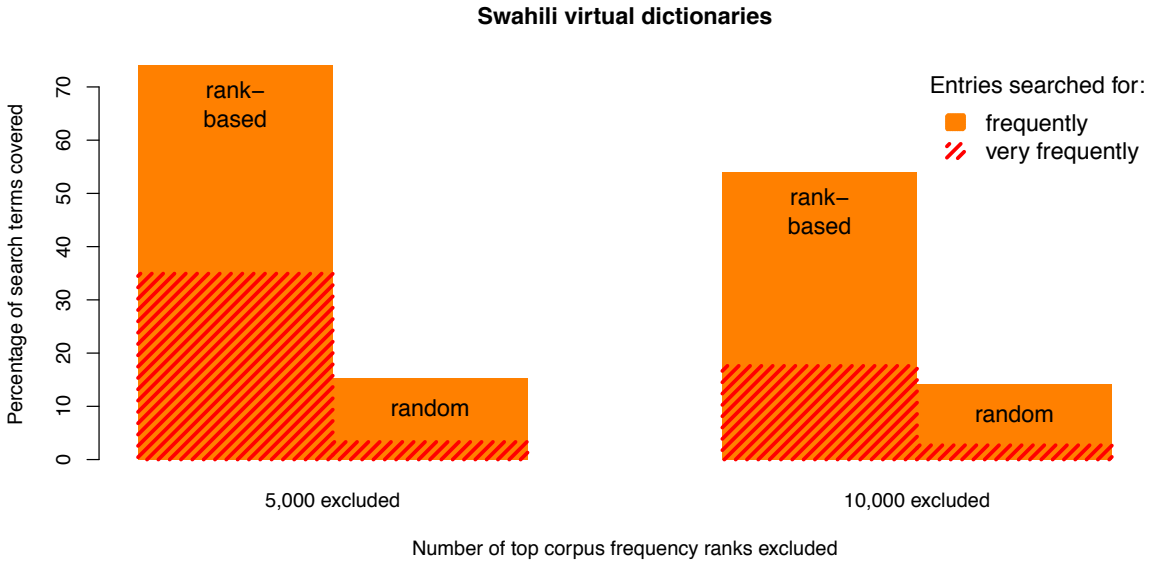

FIGURE 6. Ratio of entries frequently (orange) and very frequently (red shades) searched for in virtual dictionaries consisting of 5,000 (bars on left) or 10,000 (right) Swahili entries after the top 5,000 (left) or 10,000 (right) items have been removed from the corpus frequency list. The rank-based dictionaries include the next 5,000/10,000 items from the frequency list, while the random dictionaries contain 5,000/10,000 items randomly sampled from the rest of the frequency list

eISSN: 2550-2131
ISSN: 1675-8021

ENGLISH-TO-SWAHILI SEARCHES

After preparing and selecting the data as presented in the Section **Initial data preparation**, the final data set included 220,996 English-Swahili search request types and 9,664,213 search request tokens.

Table 2 gives the number and ratios of search request types for the three categories. It can be inferred from the table that 75.89% (100% minus 24.11%) of the search request types are searched for less than 0.5 times per million.

TABLE 2. Number and ratio of search request types for the English search for the three searches *poms* categories

Searches *poms* category	no. search request types	% search request types
regularly (>= 1 searches *poms*)	53,282	24.11
frequently (>= 2)	30,517	13.81
very frequently (>= 11)	10,706	4.84

As was the case for the Swahili searches, corpus frequency data had to be assigned to the English look-up data. English wordform corpus frequency data is available for 102,954 (46.5%) of the request types. In terms of search request *tokens*, frequency data was available for 97.5% of all tokens. So, we had to exclude a lower percentage (2.5%) of search tokens compared to the Swahili-to-English data (13.7%, see previous section).

Again, we begin with two simple tests. First, we show the three most frequent English words in the enTenTen12 corpus, and compare this top 3 with the actual search ranks:

- 'the' is the most frequent English word. In terms of dictionary look-ups we note:
  - 19,684 searches
  - 2,199.409 searches per million
  - Search rank: 10
- 'and' is the 2nd most frequent English word. In terms of dictionary look-ups we note:
  - 9,805 searches
  - 1,095.570 searches per million
  - Search rank: 43
- 'to' is the 3rd most frequent English word. In terms of dictionary look-ups we note:
  - 25,920 searches
  - 2,896.194 searches per million
  - Search rank: 7

Here, the corpus top 3 does not map as beautifully as was the case for Swahili, but it must be said that it is rather amazing to see that so many people actually do search for the function words 'the', 'and' and 'to'.

Turning the comparison around, we now list the top 3 English searches with the corresponding corpus ranks:

- 'you' is the most searched-for word. In terms of the English corpus we note:
  - Frequency: 255,568
  - Frequency per million: 11,616.7273
  - Frequency rank: 19
- 'I' is the 2nd most searched-for word. In terms of the English corpus we note:
  - Frequency: 370,980
  - Frequency per million: 16,862.7273
  - Frequency rank: 14
- 'love' is the 3rd most searched-for word. In terms of the English corpus we note:
  - Frequency: 9,209
  - Frequency per million: 418.5909
  - Frequency rank: 503

eISSN: 2550-2131
ISSN: 1675-8021

Here, the top 2 searches map rather well, the third one not at all.

As a further illustration, a few of the first English words that are only 'regularly' and 'frequently' searched for, but not 'very frequently' are:

- 'photo'
  - 93 searches
  - 10.39144 searches per million
  - Corpus frequency rank: 2,090
- 'lately'
  - 93 searches
  - 10.39144 searches per million
  - Corpus frequency rank: 9,237
- 'bun'
  - 93 searches
  - 10.39144 searches per million
  - Corpus frequency rank: 45,009
- 'motherland'
  - 93 searches
  - 10.39144 searches per million
  - Corpus frequency rank: 61,467

A few of the first English words that are only 'regularly' searched for, but not 'frequently' and 'very frequently' are:

- 'digital'
  - 13 searches
  - 1.452567 searches per million
  - Corpus frequency rank: 2,395
- 'decentralize'
  - 13 searches
  - 1.452567 searches per million
  - Corpus frequency rank: 83,935
- 'grandmama'
  - 13 searches
  - 1.452567 searches per million
  - Corpus frequency rank: 90, 467

As was the case for Swahili, once beyond the top ranks for English, comparisons on the level of individual words are not very meaningful, hence the need for our bar plot of categorized corpus frequency and search frequency, this time for English. **Figure 7** shows that the association between corpus frequency and search frequency is even clearer than in the case of Swahili: note the mirror-image reversal of the patterns between the leftmost (low corpus frequency) and rightmost (high corpus frequency) cluster of bars.

eISSN: 2550-2131
ISSN: 1675-8021

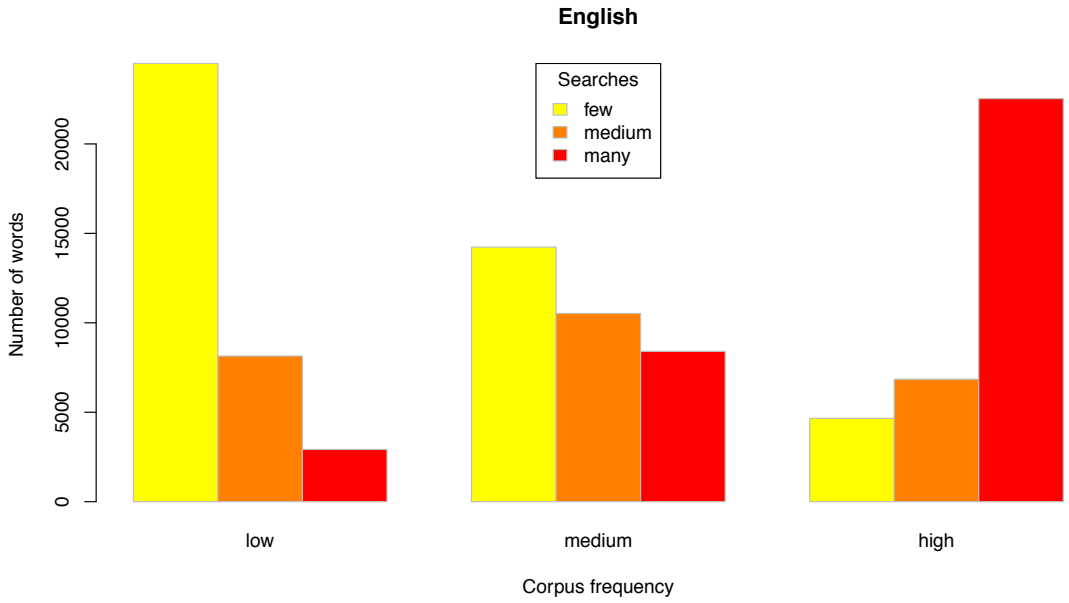

FIGURE 7. Numbers of English words coming from three corpus frequency bands that register three levels of search intensity

The finer-grained dictionary-simulation approach in **Figure 8** confirms that the positive relationship between corpus frequency and search frequency also holds for English searches. If, for example, we include the top 100 words from the corpus frequency list in our virtual dictionary, all of the entries are searched for regularly, frequently and very frequently. If we compare this to a virtual dictionary consisting of the top 30,000 words from the frequency list, the numbers diverge considerably. Not only do the ratios for all categories decline, but the categories drift apart, too. 80.8% of the entries are still searched for regularly, but only 63.7% and 30.6% of the entries are searched for frequently and very frequently, respectively. In a virtual dictionary of the top 50,000 words in the corpus frequency list, 69.3% of the entries are still searched for regularly, 49.5% of the entries are searched for frequently, and 20.4% are searched for very frequently.

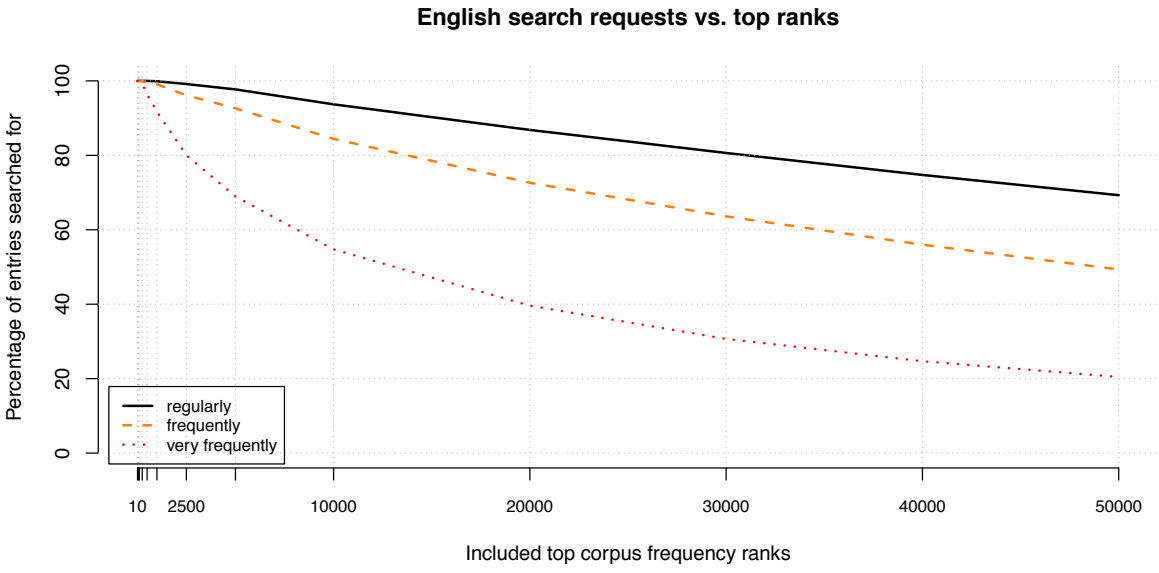

FIGURE 8. Relationship between the number of included frequency ranks from the top of the wordform frequency list for English and the percentage of entries that are searched for regularly, frequently and very frequently in English

eISSN: 2550-2131
ISSN: 1675-8021

This pattern stands in contrast to a virtual dictionary where an increasing number of entries are sampled randomly from the corpus (**Figure 9**). Apart from some fluctuation for virtual dictionaries with very few entries, which is due to the sampling process, the values for such a dictionary stabilize at around 45% for the category regularly, 28% for frequently, and 10% for very frequently. This means that a dictionary consisting of randomly sampled corpus items never outperforms a dictionary based on a corpus frequency list in terms of successful searches (**Figure 9** vs. **Figure 8**, respectively).

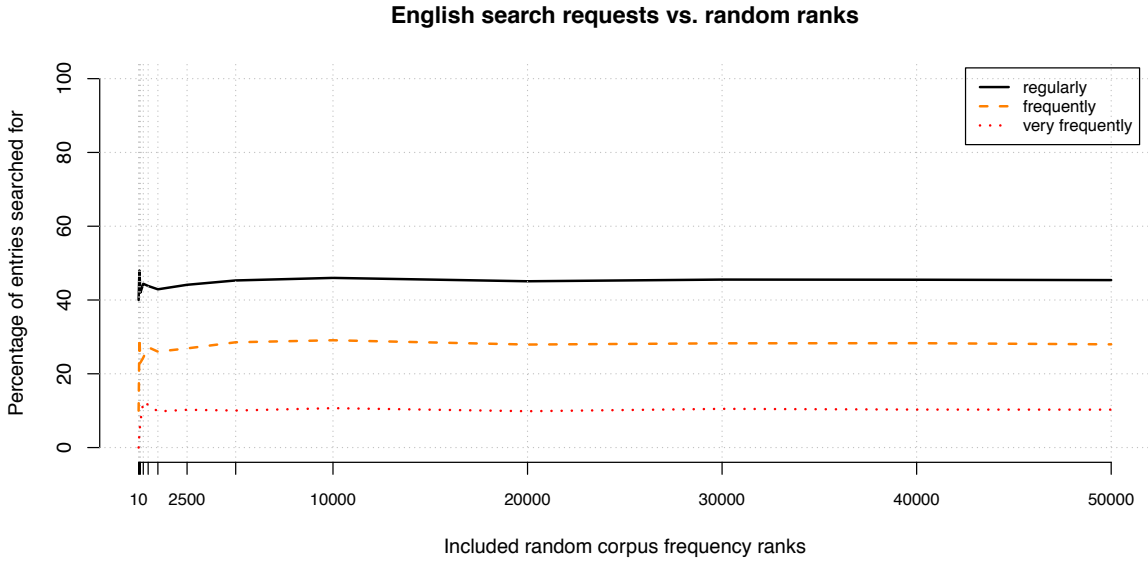

FIGURE 9. Relationship between the number of items *randomly drawn* from the English wordform list and the percentage of entries that are searched for regularly, frequently and very frequently in English.

The results for virtual dictionaries where the first 5,000 or 10,000 entries from the top of the English frequency list have been excluded (**Figure 10**) look similar to the results from the Swahili search in **Figure 6**. The dictionaries based on corpus frequency perform better in regard to search term coverage – both for entries frequently and very frequently searched for. For example, if we select the items with corpus frequency ranks 5,001 to 10,000, 76.2% of the entries are frequently searched for (left-most bar). If we randomly sample 5,000 entries among all entries with a rank higher than 5,000, only 25.0% of the entries are searched for frequently (second bar from the left). The same qualitative differences are found when we exclude the top 10,000 ranks from the corpus frequency list and when we look at the entries that are searched for very frequently.

eISSN: 2550-2131
ISSN: 1675-8021

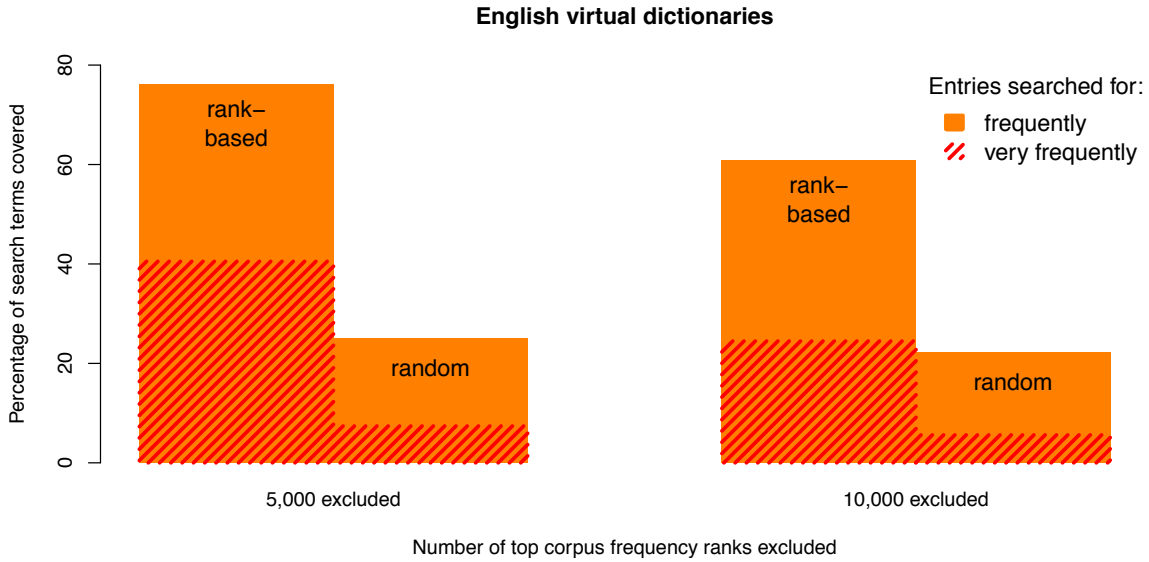

FIGURE 10. Ratio of entries frequently (orange) and very frequently (red shades) searched for in virtual dictionaries consisting of 5,000 (bars on left) or 10,000 (bars on right) English entries after the top 5,000 (left) or 10,000 (right) items have been removed from the corpus frequency list. The rank-based dictionaries include the next 5,000/10,000 items from the frequency list, while the random dictionaries have been randomly sampled from the rest of the frequency list

HEURISTIC FOR THE COMBINED SEARCHES

Halfway during the first decade in the lifetime of the online Swahili-English dictionary, the two dedicated search boxes, one per language, were replaced with a single combined search box (see the Section **The Swahili-English online dictionary**). The main reason for this change was to simplify and streamline matters for the dictionary users. The logs had in any case indicated that users paid little attention to the two boxes, as they would typically search for both Swahili and English using just any of the boxes. With a single search box, we simply showed results from the 'relevant' side. For those rare cases where a search term could be both Swahili and English (e.g., *kite* = Sw. 'passion, grief, torment, …', but also Eng. 'bird of prey which catches and eats small animals') we would then show all the relevant entries (in this case *kite*, *mwewe* and *kipanga*/*vipanga*). For our look-up analysis this meant that we had to choose how to go about deciding on the intended language when an item could be both Swahili and English. We pointed out that we opted to use a simple heuristic to do so: an item that is found in both corpus frequency lists is assigned to the language for which the normalized frequency for that item is the highest.

As part of the analysis for this article, we decided to also study how well this strategy worked. The study was undertaken for the overlap between all the types in our 22m Swahili corpus and the top 200,000 types in the full enTenTen12 corpus for English. For these two lists, there was an overlap of 17,890 types. This overlap was tagged manually for language, see Addendum 3, and may now serve as a Gold Standard.

If we now check this manually annotated list in terms of how often the selection based on frequency (the simple heuristic) is the correct one, the results are as follows:

- 11,911 of the 12,619 English entries are assigned correctly by the frequency heuristic (94.4%);
- 489 of the 492 Swahili entries are assigned correctly by the frequency heuristic (99.4%);
- out of the 4,779 'other' entries (i.e., proper names (46.1%), abbreviations (23.0%), place names (12.6%), words in other languages (8.5%), typos, junk and Roman numbers (6.3%), and words that may be both Swahili and English (3.5%)), the frequency heuristic assigned 1,698 to English and 3,081 to Swahili;
- altogether, 69.3% of all entries are assigned correctly by the frequency heuristic;

- if we leave the 'other' items out, 94.6% of all assignments by the frequency heuristic are correct.

Clearly, there is an excellent match. Given that going through a long list of overlaps is a costly (as manual) procedure, and given that we have determined that the cheaper (as automated) strategy also works, future researchers know that they don't have to go through such lists anymore.

This Gold Standard also gives us an insight into the value of the use of two separate search boxes, one per language (i.e., one per 'dictionary side'), versus the single combined search box. Altogether, 96% of the 'overlap entries' were searched for (in any box), 80% in the Swahili box, 86% in the English box, and 93% in the combined box. Focusing on the Swahili search box first, just 7% of the Swahili items were looked up solely in that box, while as many as 92% were searched for in *both* the Swahili and the English search boxes. A similar situation is found for the English search box: just 13% of the English items were looked up solely in that box, while as many as 77% were looked up in *both* the English and Swahili search boxes. While none of the Swahili items was looked up in the English search box only, still 3% of the English items were looked up in the Swahili box only. This difference is likely directly related to the design of the graphical user interface during the first half of the first decade in the lifetime of the online Swahili-English dictionary. During that time, the Swahili search box was presented first (on top), followed by the English search box (underneath), and this no matter whether the language of the interface was Swahili or English (for screenshots, see De Schryver et al. 2006, p. 81 resp. p. 82). So, dictionary users simply opted for the most readily available box, being the one at the top of the page.

Looking at the top 50 of the English items looked up in the Swahili box only, one can only wonder why anyone would think these are Swahili words:

> servers, unions, formats, rankings, extends, resorts, commissions, amino, weighed, generators, granting, turbines, playback, shedding, asserted, inhibitors, investigative, expressly, neural, registers, yielded, computerized, peroxide, dipped, externally, articulated, hubs, twenties, budgetary, visibly, generalized, grills, reversing, rotor, behavioural, spreadsheets, anti-virus, accommodated, reckoned, bots, recalling, propelled, tinted, pituitary, psi, bikers, loaned, groupings, phosphate, cropping

In short: Even though intuitively the right and even 'correct' thing to do for a bilingual dictionary, actual online dictionary usage behaviour reveals that users do not care or pay attention to 'details' such as the side of a dictionary to be consulted. A single search box is therefore not only a concession to this deplorable dictionary usage behaviour, it is also a necessity if one wishes to return relevant dictionary articles. With two boxes, any searches in 'the wrong box' will not return results (except for the few words that truly overlap between the languages), but with a combined box one simply returns all relevant material, no matter the direction of the dictionary. Metalexicographically, this 'media-related feature' (see Tan & Woods, 2008) may be seen as the digital equivalent of the 'single amalgamated central list' proposed by Martin and Gouws (2000, p. 786) which was implemented in the paper dictionary ANNA (Martin 2011).

Amalgamated bilingual dictionaries entail of course far more than simply joining the macrostructures of the two sides of a bilingual dictionary. In actual fact, the new dictionary model was designed with closely-related languages in mind, whereby similarities and differences between the languages are highlighted (see also Martin, 2012a). This is the case for ANNA, which deals with Afrikaans and Dutch; while several proposals have also been made to compile such contrastive dictionaries for Bantu languages that are close to one another (Martin, 2012b, p. 413; Prinsloo, 2014). The simple and straightforward interweaving of the two macrostructures of a bilingual dictionary has also been implemented in early

desktop dictionaries; it is for instance one of the view and search options in the first edition of the English-French CD-ROM dictionary by Collins-Robert (2003).

CONCLUSIONS AND DISCUSSION

Using, on the one hand, words actually looked up in a decade's worth of logs for Swahili, respectively English, in a real-world online Swahili-English dictionary and, on the other hand, actual word occurrence frequencies for Swahili, respectively English, as found in large corpora for these languages, we proposed two sets of simulations to settle the debate as to whether or not there is any correlation between what words people actually search for versus what people actually speak and write as reflected in corpora. Both our testing approaches revealed a clear positive relationship between corpus frequency and search frequency: items that occur more often in a corpus *are* looked up more often in a dictionary, and, even more significant, items that appear less often in a corpus *are* also looked up less often in a dictionary after all. This effect is evident in both Swahili and English.

Our first approach was an entry look-up simulation, the idea being to include incrementally more items in a dictionary from the top of a frequency list, and to note how many of them (as evidenced in our ten years of actual online dictionary logs) are looked up regularly, frequently, and very frequently. This analysis demonstrated that as one digs deeper into the frequency list, apparently ever more of the not-so-popular items are also looked up, and the percentages of all three categories gradually fall. This is evident in **Figure 4** for Swahili, and in **Figure 8** for English. When the same exercise is repeated but with dictionary entries included randomly, rather than based on their top ranks in a corpus, the proportion of items looked up regularly, frequently, and very frequently remains virtually constant, and at all times these proportions are also much lower than for entries taken off the top of the frequency lists. Reformulated: There is no effect whatsoever of dictionary size on look-up success when one is just pulling random dictionary samples of varying sizes from a corpus. This is evident in **Figure 5** for Swahili, and in **Figure 9** for English.

With regard to the issue of the point at which corpus frequency is no longer helpful, hypothesized in De Schryver et al. (2006, p. 78) to be around 3,000 for Swahili and 5,000 for English, our second batch of simulated tests, consisting in discarding the top 5,000 or top 10,000 items from the dictionary and working with the remainder of the frequency list only, demonstrated that corpus frequency continues to exert a positive effect beyond these putative threshold values. As shown in **Figure 6** for Swahili and **Figure 10** for English, the rank-based dictionaries (based on the items immediately following ranks 5,000 or 10,000 from the top of the truncated frequency list) exhibit clearly better coverage than the random dictionaries (based on 5,000 or 10,000 items sampled from anywhere in the truncated frequency list).[9] Reformulated: In order to boost look-up success for words looked up less frequently in a dictionary, the best way to select those lesser-frequent entries remains to base that selection on corpus frequencies.

Although the various results for Swahili and English turn out to be comparable – to a point even rather similar – they are not identical. Take the first set of simulations (**Figure 4** and **Figure 5** for Swahili vs. **Figure 8** and **Figure 9** for English). The fact that look-up success rates remain constant in dictionaries for which the lemmas have been randomly selected is a truly stunning outcome from our analyses, as we are provided with a baseline: in Swahili this baseline is 30% for regularly, 17% for frequently, and 5% for very frequently searched-for entries, in English it is 45% for regularly, 28% for frequently, and 10% for very frequently searched-for entries. In other words, these baselines are language-dependent; in our case being better for English than for Swahili. We claim that this difference is a direct

result of the different morphological structure of words in Swahili (rather complex) compared to English (very straightforward). These values also give us the absolute minimal look-up success rates for online dictionaries for these languages; it is literally simply impossible to do 'worse'. Indeed, when one bases the selection of the lemmas on corpus frequencies, one improves the success rates from what is seen in **Figure 5** to what is seen in **Figure 4** for Swahili, and from what is seen in **Figure 9** to what is seen in **Figure 8** for English. Comparisons of these graphs also indicate that while the 'improvements' are substantial for the high and mid-frequent corpus items, the bigger a dictionary becomes, the smaller the improvements to the look-up success rates. For very huge dictionaries, one may assume one reaches a point where there are no serious further improvements to speak of.

This relates to a point made in Müller-Spitzer et al. (2015, pp. 13-14), with reference to the second type of simulations (**Figure 6** and **Figure 10** in our case):

> 'One could now wonder how many frequency ranks have to be excluded until frequency really does not matter anymore. We would argue that this question cannot be answered given the available corpus data. Due to the Zipfian pattern of frequency distributions, corpora get less and less sensitive to frequency differences in lower frequency ranges. Therefore, as soon as we enter very low regions of the frequency band, observed frequency differences get too small to show any effects on look-up frequency. Note that this does not have to be due to the fact that there really are no effects anymore – our available corpus data is simply not sensitive enough to capture them.'

Müller-Spitzer et al. (2015) made these points based on data for German; we can make the same points based on our data for Swahili and English.

Lastly, taking a bird's-eye view of **Figure 3** through **Figure 10**, the relationships seem to be strikingly similar despite the dramatic differences in the morphological type of the language. In other words, these languages from two different language families behave quite alike in terms of corpus frequencies predicting look-up frequencies, a predictive power that may very well be language universal. This, clearly, is an important finding.

ACKNOWLEDGEMENTS

Heartfelt thanks are due to David Joffe of TshwaneDJe HLT for collecting and providing us with the online Swahili-English dictionary log files, to Jutta De Nul of Ghent University for her help with tagging the overlap between the Swahili and English corpus data, and to Ondřej Matuška of Lexical Computing for providing us with a sample of the enTenTen12 corpus.

ENDNOTES

[1] This dictionary was conceived as a research tool, so the changes over the years are not accidental, but inspired by research questions. In-depth longitudinal studies of the log files that take account of these changes will be reported on in forthcoming articles.

[2] From these, potential research questions that will be looked into in subsequent studies include: What is the effect of the localization language?, When do users click on cross-references?, Does the number of hits returned have any effect on the use of the dictionary?, Does the time of the day (per region) result in different types of searches?, etc.

[3] The General Data Protection Regulation (GDPR) is a legal framework that sets guidelines for the collection and processing of personal information of individuals within the European Union (EU).

[4] See https://www.sketchengine.eu/ententen-english-corpus/.

eISSN: 2550-2131
ISSN: 1675-8021

[5] Although searches for word stems, which by definition start with a hyphen, are also an option in the online Swahili dictionary, we are interested in searches of full orthographic words, which may be compared directly with the types in the unlemmatized frequency list derived from the 22m Swahili corpus.
[6] Bar q and z, Swahili uses all the other letters of the Latin alphabet, plus the apostrophe ', such as in *ng'ombe* 'cow'. The apostrophe can only appear after the letter sequence 'ng', and not word-initially nor in the word-final position.
[7] This was done by partitioning the database into three roughly equally-sized 'search groups', and three roughly equally-sized 'frequency groups', separately for each language side.
[8] Note that we did not just sample once, but 1,000 times. The figures reported for the random dictionaries are the mean percentages of these 1,000 sampling runs.
[9] This replicates Koplenig et al.'s (2014) results quite closely, but for other languages and dictionaries.

REFERENCES

Abate, F. (1985). Dictionaries past & future: Issues and prospects. *Dictionaries: Journal of the Dictionary Society of North America. 7*, 270–283.

Bergenholtz, H. & Johnson, M. (2005). Log files as a tool for improving internet dictionaries. *Hermes. 34*, 117–141.

Collins-Robert (2003). *The Unabridged Collins-Robert Electronic French Dictionary* (CD-ROM desktop dictionary, including the Collins-Robert Unabridged French Dictionary and the Collins-Robert Comprehensive French Dictionary). Paris: Dictionnaires Le Robert / VUEF.

Crystal, D. (1986). The ideal dictionary, lexicographer and user. In R. F. Ilson (Ed.), *Lexicography: An emerging international profession* (pp. 72–81). Manchester: Manchester University Press.

De Schryver, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography. 16*(2), 143–199.

De Schryver, G.-M. (2018). Towards a new type of dictionary for Swahili. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, 17-21 July 2018, Ljubljana, Book of Abstracts* (pp. 98–100). Ljubljana: Faculty of Arts, Ljubljana University Press.

De Schryver, G.-M. & Joffe, D. (2004). On how electronic dictionaries are really used. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004* (pp. 187–196). Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

De Schryver, G.-M., Joffe, D., Joffe, P. & Hillewaert, S. (2006). Do dictionary users really look up frequent words? – On the overestimation of the value of corpus-based lexicography. *Lexikos. 16*, 67–83.

Hillewaert, S. & De Schryver, G.-M. (2004). *Online Kiswahili (Swahili) – English Dictionary*. https://www.goswahili.org/dictionary/.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004* (pp. 105–116). Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Knowles, F. E. (1983). Towards the machine dictionary: 'mechanical' dictionaries. In R. R. K. Hartmann (Ed.), *Lexicography: Principles and Practice* (pp. 181–197). London: Academic Press.

Koplenig, A., Meyer, P. & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In C. Müller-Spitzer (Ed.), *Using online dictionaries* (pp. 229–249). Berlin: Walter de Gruyter.

Lemnitzer, L. (2001). Das Internet als Medium für die Wörterbuchbenutzungsforschung. In I. Lemberg, B. Schröder & A. Storrer (Eds.), *Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher* (pp. 247–254). Tübingen: Niemeyer.

Lew, R. & De Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography. 27*(4), 341–359.

Lorentzen, H. & Theilgaard, L. (2012). Online dictionaries – how do users find them and what do they do once they have? In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 654–660). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Martin, W. (2011). *Pharos Groot Woordeboek. Afrikaans en Nederlands (Prisma Groot Woordenboek Afrikaans en Nederlands)*. Cape Town: Pharos.

Martin, W. (2012a). Amalgamated bilingual dictionaries. In R. Genis, E. de Haard, J. Kalsbeek, E. Keizer & J. Stelleman (Eds.), *Between West and East: Festschrift for Wim Honselaar, on the Occasion of his 65th Birthday* (pp. 437–449). Amsterdam: Pegasus.

Martin, W. (2012b). ANNA: A dictionary with a name (and what lies behind it). *Lexikos. 22*, 406–426.

Martin, W. & Gouws, R. H. (2000). A new dictionary model for closely related languages: The Dutch–Afrikaans Dictionary Project as a case-in-point. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (Eds.), *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany* (pp. 783–792). Stuttgart: Institut for maschinelle Sprachverarbeitung, Universität Stuttgart.

Mohamed, A. A. (2009). *Kiswahili for Foreigners* (3rd revised edition). Zanzibar: Goodluck Publishers.

Müller-Spitzer, C., Wolfer, S. & Koplenig, A. (2015). Observing online dictionary users: Studies using Wiktionary log files. *International Journal of Lexicography. 28*(1), 1–26.

Prinsloo, D. J. (2014). Lexicographic treatment of kinship terms in an English/Sepedi-Setswana-Sesotho dictionary with an amalgamated lemmalist. *Lexikos. 24*, 272–290.

Schoonheim, T., Tiberius, C., Niestadt, J. & Tempelaars, R. (2012). Dictionary use and language games: Getting to know the dictionary as part of the game. In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 974–979). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Sinclair, J. (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.

Tan, K. H. & Woods, P. C. (2008). Media-related or generic-related features in electronic dictionaries: learners' perception and preferences. *GEMA Online® Journal of Language Studies. 8*(2), 1–17.

Trap-Jensen, L. (2014). Korpus eller brugerne – hvem får det sidste ord? In M. H. Andersen, J. N. Jensen & P. Jarvad (Eds.), *Neologismer. Dansk Sprognævns 2. seminar om nye ord. København 5.-6. november 2013* (pp. 129–144). Copenhagen: Dansk Sprognævn.

Trap-Jensen, L., Lorentzen, H. & Sørensen, N. H. (2014). An odd couple – Corpus frequency and look-up frequency: What relationship? *Slovenščina 2.0. 2*(2), 94–113.

Verlinde, S. & Binon, J. (2010). Monitoring dictionary use in the electronic age. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1144–1151). Leeuwarden: Fryske Akademy.

ADDENDUM 1: NUMBER OF SWAHILI LEMMAS IN THE ONLINE SWAHILI-ENGLISH DICTIONARY DURING THE FIRST TEN YEARS

Update	Total number of lemmas
2004-05-09	1,632
2004-05-24	1,908
2004-06-07	2,203
2004-07-02	2,616
2004-09-16	3,289
2006-03-12	6,475
2006-07-23	6,475
2006-07-29	6,475
2006-10-25	15,526
2006-10-28	15,525
2009-05-22	15,518

ADDENDUM 2: ANALYSIS OF A RANDOM SAMPLE OF 100 SWAHILI DICTIONARY SEARCHES NOT FOUND IN THE SWAHILI CORPUS

[see Excel sheet in Supplementary Materials]

34284-106999-3-SP.xlsx

ADDENDUM 3: STUDY OF THE OVERLAP BETWEEN THE 22M SWAHILI CORPUS AND THE FULL ENTENTEN12 CORPUS FOR ENGLISH

[see Excel sheet in Supplementary Materials]

34284-106990-3-SP.xlsx

eISSN: 2550-2131
ISSN: 1675-8021

ABOUT THE AUTHORS

Gilles-Maurice de Schryver is the President of the *European Association for Lexicography* (EURALEX), and a two-term past President of the *African Association for Lexicography* (AFRILEX). He holds an MSc in microelectronic engineering, as well as an MA and PhD in African languages and cultures. Currently a research professor at the Centre for Bantu Studies at Ghent University, in Ghent (Belgium), he has (co-)authored about 300 books, book chapters, journal articles and conference papers on lexicography.

Sascha Wolfer is a researcher in the Department for Lexical Studies at the Leibniz Institute for the German Language (IDS) in Mannheim (Germany). He holds an MA in German linguistics, cognitive science and political science as well as a PhD in cognitive science. He is a member of the editorial board of the *International Journal of Lexicography*. His current research focuses on quantitative approaches in corpus linguistics, psycholinguistics, text comprehension, and lexicography.

Robert Lew is a professor at the Faculty of English, Adam Mickiewicz University, Poznań (Poland). His research interests centre around dictionary use. He has worked as a practical lexicographer for major dictionary publishers and is the Editor of the *International Journal of Lexicography*.

eISSN: 2550-2131
ISSN: 1675-8021