http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Corpus-driven Bantu Lexicography Part 2: Lemmatisation and Rulers for Lusoga

Gilles-Maurice de Schryver, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of African Languages, University of Pretoria, Pretoria, South Africa (gillesmaurice.deschryver@UGent.be)*
and
Minah Nabirye, *BantUGent, Department of Languages and Cultures, Ghent University, Ghent, Belgium; and Department of Teacher Education and Development Studies, Kyambogo University, Kampala, Uganda (minah.nabirye@UGent.be)*

Abstract: This article is the second in a trilogy that deals with corpus-driven Bantu lexicography, which is illustrated for Lusoga. The focus here is on the macrostructure and in particular on the building of a lemmatised frequency list directly within a dictionary-writing system. The programming code for the parts of the lemmatisation that may be automated is included as addenda. A second focus is on the embedded part-of-speech and alphabetical rulers, for which it is shown how these may be used to plan the actual compilation of the dictionary entries.

Keywords: BANTU, LUSOGA, CORPUS LEXICOGRAPHY, LEMMATISATION, LEMMATISED FREQUENCY LIST, PART-OF-SPEECH RULER, ALPHABETICAL RULER, MULTIDIMENSIONAL LEXICOGRAPHIC RULER, DICTIONARY PLANNING, DICTIONARY-WRITING SYSTEM, TLEX, TSHWANELEX

Obufunze: Omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu. Ekitundu 2: Okugelaagelania eigambowaziso n'enta dha namugelo waalyo mu walifu w'Olusoga. Olupapula luno n'olwo'kubili mu nteeko y'okulaga omusomo gw'omutengeso gw'eitu ogukozesebwa mu namawanika w'ennimi dha Bantu ogulaga omulimu ogw'akolebwa ku Lusoga. Mu lupapula luno eisila liteebwa ku muteeko gw'omutindiigo okusingila ilala ku kuzimba olukalala lwa namungi w'ebigambowazo mu muteeko ogukozesebwa okuwandiika amawanika. Namugelo w'okutegekuza ebitundu by'okugambowaza ebisobola okuba mu mbeela ya kaneetindiigo bilagibwa mu kikugilo. Eisila ely'okubili lili ku mbu dh'ebigambo edh'ennimbyo n'engeli ye dhilagibwa mu nsengeka ya walifu ng'olupapula luno kwe lusinziila okuwa endowooza ekoba nti ebintu bino ebibili bisobola okukozesebwa okutaawo omusingi gw'okwingiza ebigambo mu iwanika.

Ebigambo ebikulu: BANTU, LUSOGA, EITU LY'ANAMAWAIKA, OKUGAMBOWAZA, OLUKALALA LWA NAMUNGI W'EBIGAMBOWAZO, ENNEYOLEKA Y'EMBU, ENNEYOLEKA YA WALIFU, OMUTENGO GW'ENNEYOLEKA YA NAMAWANIKA, ENTEGEKA Y'EIWANIKA, ENGELI EDHIKOZESEBWA OKUWANDIIKA AMAWANIKA, TLEX, TSHWANELEX

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

1. Goal of the present study

This article is concerned with the use of corpora to successfully kickstart Bantu-language dictionary projects. Considering the traditional lexicographic distinction between the macrostructural and the microstructural level, this therefore means that the present study will focus on the design of the macrostructure of a Bantu-language dictionary, for which Lusoga will serve as an example. The major reference for any corpus-based macrostructural issues in Bantu lexicography is de Schryver and Prinsloo (2000). A year later, de Schryver and Prinsloo (2001) looked at the difference between intuition-based and corpus-based designs of various lemma-sign lists, as found in and for Northern Sotho dictionaries. While a single study on how to draw up a dictionary's macrostructure may suffice for a disjunctively-written Bantu language like Northern Sotho, much more guidance is certainly needed for the conjunctively-written ones.[1] To date, there seems to be just one such published study, for Southern Ndebele (de Schryver 2003). In our case study for Lusoga below, which is based on Nabirye (2016), we will further develop the proposals from the 2003 study, and will in effect offer a hands-on method which may be performed directly within a dictionary-writing system. The programming code needed for the actual lumping of all the members of each single lemma, as well as for the summations of the underlying corpus frequencies, and the calculation of the frequency bands, will be presented as addenda.

As a supplementary objective, we will want to uncover the relationships between lemmatised frequency lists of conjunctive Bantu languages, and their unlemmatised counterparts. While lemmatised and unlemmatised frequency lists may be near-identical for a disjunctive Bantu language like Northern Sotho (Prinsloo and de Schryver 2007), this is certainly not the case for a conjunctive one like Lusoga. This part of the study will inevitably also require a consideration of two types of rulers: 'part-of-speech rulers' and 'alphabetical rulers' (aka 'multidimensional lexicographic rulers') (de Schryver 2013). In order to put our results in perspective, comparisons will furthermore be made with comparable data freshly drawn from the *Oxford Bilingual School Dictionary: Zulu and English* (de Schryver 2010a).

2. Automated vs. manual, and semi-manual lemmatisation

How does one begin analysing a corpus with the aim of compiling a dictionary of the language covered by that corpus? Modern dictionary-makers will want to start from a lemmatised frequency list derived from that corpus, with which they can set out to build the macrostructure of their dictionaries. A good entry point for the concept of lemmatisation in the field of computational and corpus linguistics remains Kilgarriff's:

> By 'lemmatised', we mean two things. First, for verbal *aim*, the count will consider all instances of *aim*, *aims*, *aiming*, *aimed*; and second, it will exclude all non-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

> verbal instances, so nominal *aim* and *aims* will not be counted. The count will be of verbal instances only of any of the four forms.
> (Kilgarriff 1997: 139)

In other words, the idea is to take a list of orthographic words, each with their type frequency as counted in a corpus, and to turn that list into its lemmatised counterpart, now with summed frequencies and a part of speech for each lemma. The result is a so-called 'lemmatised frequency list'.

While automatic lemmatisers capable of processing raw corpus data may be available for several of the world's major languages, no such software has of course been written for Lusoga. Actually, for the Bantu languages as a whole, only Swahili has been provided with working tools for this task, by Hurskainen (1992, 2016) who uses a rule-driven approach, and by the AfLaT team (De Pauw et al. 2006) who use a data-driven approach. The AfLaT team also developed small data-driven part-of-speech taggers for Northern Sotho, Zulu and Cilubà (De Pauw et al. 2012), while a team at the University of South Africa (UNISA) built broad-coverage finite-state morphological analysers for Xhosa, Swati and Southern Ndebele (Bosch et al. 2008) by adapting an existing prototype morphological analyser for Zulu (Bosch and Pretorius 2003, 2004).

In his MA, de Schryver (1999: 118-129) proposed a low-key, fully manual approach to the lemmatisation task of a Bantu language, which he successfully applied to Cilubà for the compilation of a set of bilingual Cilubà-Dutch dictionaries (de Schryver and Kabuta 1997, 1998). His basic assumption was that there is no need to lemmatise an entire corpus, as only the frequent orthographic word forms are needed as lemma signs in a general-language dictionary. Taking into account the Zipfian distribution of corpus frequencies (Zipf 1935, Kilgarriff 1997: 136-137), it is indeed clear that the lemmatised forms of low-frequency orthographic words and hapaxes hardly make a dent in what is frequent. De Schryver explained his approach as follows, after having used WordSmith Tools (Scott 1996–2018) to calculate the frequency of all the orthographic words in a 300 000-word corpus of Cilubà:

> [...] we simply went through the first 1,000 items of the [WordSmith Tools output, ranked in descending frequency order] and lemmatised 'by hand.' For nouns this meant that, when we encountered a singular form, we added the frequency of the plural form (or vice versa), where relevant. For verbs this meant that we kept track of those verbs we had already encountered and added the frequency of every single 'conjugated form' we encountered subsequently. Also, for very frequent verbs we brought together the frequencies of the entire paradigm. In addition to this 'true lemmatisation' we joined divergent orthographies — and this for all possible parts of speech.
> (de Schryver 1999: 125)

To move from a lemmatised frequency list to the actual macrostructure, de Schryver (1999: 127-128) further stipulated that candidate lemma signs should occur 'in a sufficient variety of sources' (Sinclair 1995: ix), or as put by Knowles:

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

> [...] a word must occur evenly in a large number of the stratified sub-samples rather than excessively often in a small number of them, given that these two very different cases could show identical 'total-corpus' frequencies.
> (Knowles 1983: 188)

Finally, and in imitation of Kilgarriff (1997), de Schryver (1999: 150-152) also marked the frequent lemma signs in his dictionary, using three frequency bands which had been directly derived from the top ranks as seen in his lemmatised frequency list.

In de Schryver (2003) a suggestion was made to enlist the power of spreadsheet software for the same task, where it was illustrated for Southern Ndebele. In the latter article, a four-step methodology was introduced to go from a raw corpus (i.e., a corpus without any linguistic annotations) to a lemmatised frequency list (i.e., the list of candidate dictionary citation forms together with summed frequencies, ordered from most to lesser frequent). The steps themselves have been summarised as follows:

> In Step 1 top-frequency words are extracted from a corpus of running text. This step can be performed with versatile corpus query software such as WordSmith Tools. In Step 2 the dictionary-citation forms are isolated from each of the top-frequency items; in Step 3 the dictionary-citation forms that are equal as well as their corresponding frequencies are brought together; and in Step 4 frequency bands are added to the lemma-sign list. Steps 2 to 4 can easily be performed with spreadsheet software such as Microsoft Excel.
> (de Schryver 2003: 22-23)

Observe that in this four-step methodology, parts of speech were not taken into account, as they should have been. This 'error'[2] has been corrected in the method to be explained now.

Over the subsequent years, the use of spreadsheet software morphed into using the dictionary application TshwaneLex (TLex) (Joffe and de Schryver 2002–18) to undertake Steps 2 to 4. When using TLex to lemmatise corpus data, orthographic words together with their frequencies and their spread across the corpus texts constitute the input, while the output consists of the lemma signs, with frequencies, parts of speech, ranks and frequency bands, and, optionally, main meanings. In effect, the Bantu to English sides of the school dictionaries for Northern Sotho, Zulu and Xhosa published by Oxford University Press Southern Africa (OUPSA) (de Schryver 2007, 2010a, de Schryver and Reynolds 2014) have all used TLex to draw up the macrostructure along these lines.[3]

Even though an in-depth analysis was undertaken of the compilation of the OUPSA Zulu school dictionary, the creation of its macrostructure was not discussed as part of that analysis: 'Detailing how the Zulu lemma list was created would need at least one other paper-length treatment' (de Schryver 2010b: 166). By explaining how Steps 2 to 4 may be performed within TLex in the present article (as will be done in §3 below), we will (finally) have begun dealing with this issue in the scientific literature of our discipline.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

3. From corpus to lemmatised frequency list

As was seen in Part 1 of the present series of three articles, a Lusoga corpus of 1.7 million words (tokens) contains approximately 200 000 orthographically different words (types), and it is the latter that need to be lemmatised. Two hundred thousand words are still too many to look at manually, so, as a proxy, the idea is again to work with the top-frequent orthographic words only, and thus also to lemmatise only that top section. In practical terms one chooses a cut-off frequency, and focuses on all the types with a frequency at and above that threshold. We decided to work through about 10 000 types, which corresponded to a cut-off frequency of 12 in the 1.7m Lusoga corpus.

By lemmatising the top 10 000 orthographic words in a Lusoga corpus, all the common 'words' of the language will be known: each will have been given a part-of-speech tag, as well as a relative frequency (and in the approach that will be suggested, also a brief meaning). The term *word* was placed between quotes, as we are referring here to the component known to computational linguists as the *lemma*, to dictionary-makers as the *dictionary citation form*, to metalexicographers as the *lemma sign*, and to Bantuists most likely as the *stem*.

The full 1.7m Lusoga corpus was loaded into WordSmith Tools, and with its *WordList* tool a wordlist of all the orthographic words in the corpus, together with their respective frequencies and the number of files each orthographic word occurs in, was generated. This information was imported into TLex, using its *Import* function. The approach from then onwards was to go down the frequency list in TLex, down to frequency 12, and to add for each orthographic word the following: the lemmatised form, the part of speech, and a brief meaning — all in dedicated slots in the dictionary-writing system. Differences in orthography were taken care of on the fly, as a uniform spelling was pursued in the slot for the lemma. See Figure 1 for a screenshot of the first step: the orthographic form from the corpus is in dark blue at the beginning of each entry; the lemmatised form follows in black and between square brackets; the part of speech is in pink and italics; the brief meaning(s) of the lemma is/are in green; the frequency of the orthographic form is in red and italics preceded by 'freq.'; the rank is in light blue and preceded by 'rank'; and the number of files in which the orthographic form was found is in black preceded by a hashtag and the word 'texts'.

As we proceeded down the frequency list,[4] the *fanouts* tool of TLex enabled us to preview those unlemmatised forms that would eventually be brought together under a single lemma. In the DTD (i.e., Document Type Definition (Joffe and de Schryver 2005)) one may actually choose which field to use for that, typically the field for the TEs (i.e., the translation equivalents), but at times using the lemma field for fanouts is also handy. The latter is done in Figure 2. Regardless of which one is used for fanouts, during actual lemmatisation the software will need to take the lemma *in combination with* the part of speech into account.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

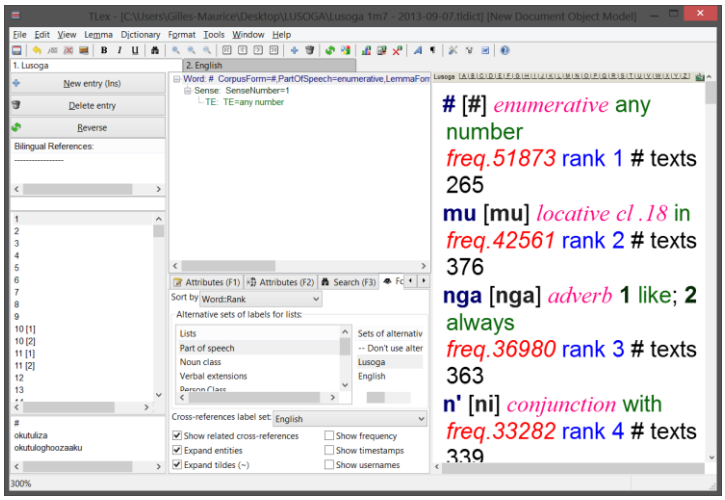

Figure 1: Lemmatising the 1.7m Lusoga corpus in TLex: going down the unlemmatised frequency list

In Figure 2 we went back to the infinitive form for the verb 'to come'. All other entries where we added **-idha** as a lemma are automatically brought together by the fanouts tool. They are all verbs, and they will indeed all be merged into a single **-idha**, and their respective frequencies will all be summed.

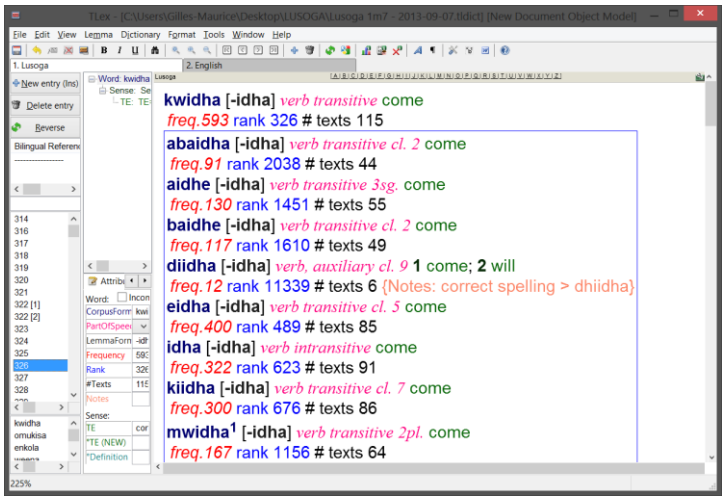

Figure 2: Lemmatising the 1.7m Lusoga corpus in TLex: the fanouts tool brings all the entries with the same lemma together

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Contrast this with the material seen in Figure 3, where the orthographic forms with **-kazi** as the lemma are brought together. Given that there are both nominal and adjectival forms, these two word classes will need to be kept separate from one another when the material is eventually merged.

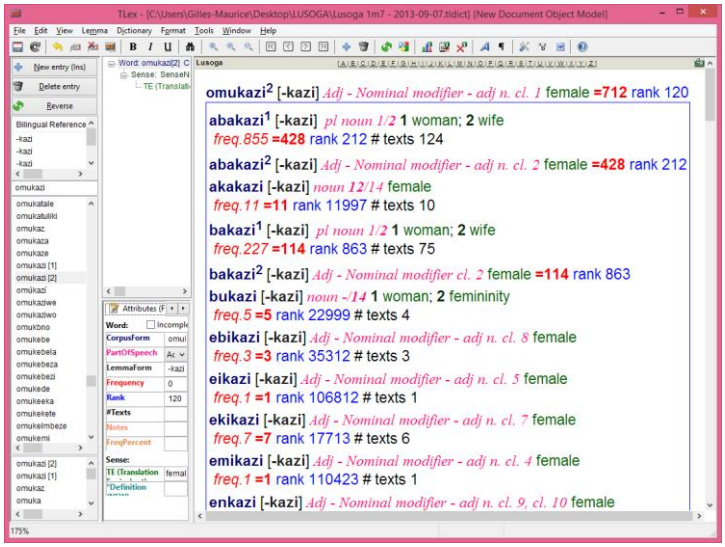

Figure 3: Lemmatising the 1.7m Lusoga corpus in TLex: the combination 'lemma & part of speech' will eventually be used to bring related forms together

Figure 2 illustrates that notes could additionally be attached to any entry; seen in orange and between curly brackets. Figure 3 illustrates another aspect, namely that for closed-class sets such as pronouns and adjectives, all the forms were considered in which the respective stems occurred in the 1.7m Lusoga corpus, and not only those with a frequency of at least 12. This could simply be achieved by doing field-specific searches across the entire TLex database, given that the full wordlist had been imported. This change in approach meant that the frequencies of the resulting lemma signs of these closed-class items were slightly raised. This was a trade-off, but with the advantage that the full picture became available for each of these closed-class items.[5]

Implicit in Figure 3, given the raised homonym numbers, is the fact that many entries had to be split up in two or more parts, typically because they could be assigned to different parts of speech, and/or because they had unrelated translation equivalents. Such entries were duplicated, and their frequencies were redistributed based on a quick and rough corpus sample.[6] In Figure 3, **omukazi**[1] (not shown) is the noun 'woman; wife'.

This lemmatisation phase took us about one month. A total of 10 318 items were eventually tagged,[7] which corresponds to just over 5% of the types in the

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

1.7m Lusoga corpus, but it also corresponds to well over 80% of the tokens. Eighty percent of the word forms in the 1.7m Lusoga corpus were accordingly seen by only looking at 5% of it.

Three Lua scripts were then written which run in TLex to actually perform the lemmatisation: (i) to bring the 'lemma – part-of-speech' pairs together, see Addendum 1; (ii) to sum the frequencies of all the members of each of these pairs and to calculate the new ranks, see Addendum 2; and (iii) to use the latter ranks to group the lemma signs into frequency bands, see Addendum 3. A random section of the outcome, ranks 500 to 510, is summarised in Table 1.

Table 1: Lemmatised frequency list for Lusoga, ranks 500-510, derived from the top 10 000 types in the 1.7m Lusoga corpus

Lemma	Part of speech	Meaning	Freq.	Rank	Freq. band
-lim-	*verb*	dig; farm	296	500	①
-goloza	*noun* ***5***/6	county	295	501	②
-ikiliza	*noun* ***1***/2	believer; saint	295	502	②
nkani	*connective*	at least	295	503	②
ee	*ideophone*	wonder	293	504	②
-lundi	*pl noun 3*/***4***	instances	293	505	②
-idhukil-	*verb*	remember; recall	292	506	②
-taama	*noun* ***9***/*10*	sheep	291	507	②
-teekw-	*verb, modal*	must	290	508	②
nguli	*connective*	if	288	509	②
-wanika	*noun* ***5***/*6*	treasury; mortuary; dictionary	286	510	②

Regarding these three Lua scripts, it is important to point out that they may be re-run at any time, with changing data, even (also!) during actual dictionary compilation, down to the very last day of preparing an actual dictionary. Specifically with regard to the third Lua script, the one which adds the frequency bands, it is moreover trivial to change the values, which are set here to mark the top 500 lemma signs with ①, the next 500 with ②, the third 500 with ③, and no symbol for the remainder.

Table 1, which summarises data (al)ready in TLex, can also be seen as the start-pack of a (bilingual) Lusoga dictionary. This, of course, is no coincidence.

To develop the potential of this material further, the next two sections (§4 and §5) are structured in the same way, based on the fact that the lemmatised frequency list that was built directly with and into TLex embeds both part-of-speech data as well as alphabetical information: first, a type of ruler is introduced theoretically; then, a practical one is built for Lusoga; followed by a comparison with an equivalent Zulu ruler; ending with the use of such a ruler in the planning of the actual compilation of a future (bilingual) Lusoga dictionary.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

4. From lemmatised frequency list to part-of-speech distributions

4.1 Part-of-speech rulers

As shown by de Schryver (2013), the relative size of each word class does not constitute a fixed percentage across corpora of the same language. Intuitively, it is clear that a large general-language corpus will proportionally contain more nouns and verbs than a smaller one (Hanks 2001). The trend, it turns out, is asymptotic, and from a few thousand items onwards one gets a good idea of the *direction* of the distribution of the various word classes. This may be illustrated with data taken from the unlemmatised version of the 100m *British National Corpus* (BNC 1994–2018), as shown in Figure 4.

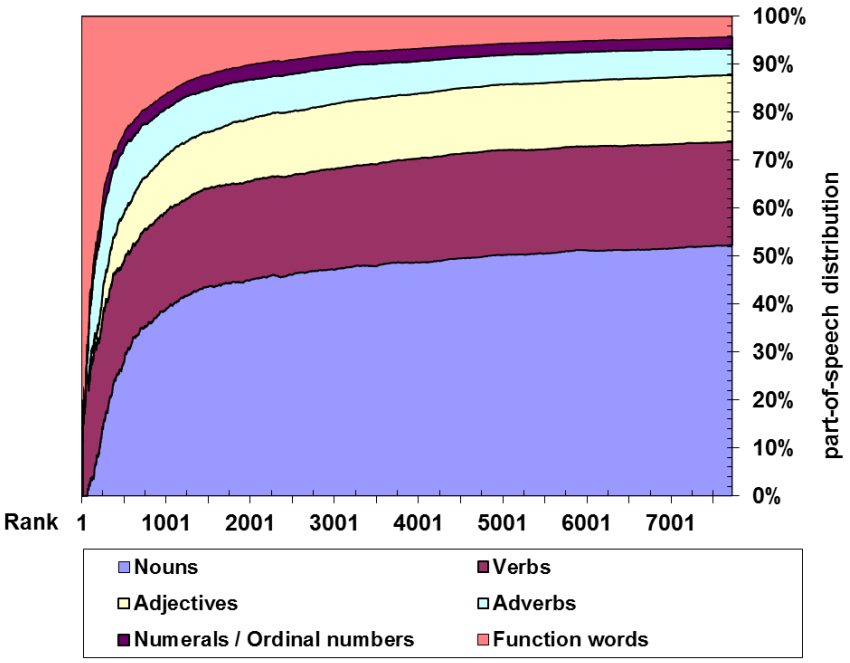

Figure 4: Part-of-speech distribution of the top 7 000+ types in the unlemmatised 100m *British National Corpus* [taken from de Schryver (2013: 1387)]

With regard to the data in Figure 4, de Schryver argues:

> One may clearly deduce from this graph that function words and verbs dominate the top-frequent ranks in an English corpus. The percentage of nouns grows steadily as one goes down the frequency list. At the 1,000+ mark the overall percentage of nouns already stands at 40 %, that of the verbs at 20 %, while the

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

> function words shrank to 16 % of the total (whereas these still represented roughly two thirds at the 100 mark). [...] The allocation to the nouns at the 7,000+ mark [...] stands at 52 %, that to the verbs grew to 22 %, while the function words shrank to a mere 4% of the total. These graphs can be extended down to any rank, while the same type of calculations can of course also be performed on lemmatized frequency lists, with similar results.
> (de Schryver 2013: 1386-1388)

What is important to remember from this is that there are as many part-of-speech rulers as there are numbers of lemma signs in a dictionary; each dictionary has a different distribution. Indeed, looking up from any rank in a graph like Figure 4, one obtains a different part-of-speech ruler.

4.2 Towards a part-of-speech ruler for Lusoga

The distribution of the main parts of speech in the lemmatised frequency list derived from the top section of the 1.7m Lusoga corpus is shown in Table 2 and Figure 5.

Table 2: Statistics for the distribution of the parts of speech in the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

Rank	Part of speech	Lemmatised	% = POS-ruler
1	noun	2 440	**57.41%**
2	verb	1 113	**26.19%**
3	pronoun	156	**3.67%**
4	quantifier	143	**3.36%**
5	adjective	117	**2.75%**
6	locative	75	**1.76%**
7	connective	68	**1.60%**
8	interjection	54	**1.27%**
9	ideophone	49	**1.15%**
10	adverb	35	**0.82%**
SUM		**4 250**	**100.00%**

As can be seen, the main part of speech of Lusoga is the noun, which accounts for 57% of all the lemma signs. The second most frequent part of speech is the verb, covering 26%. Nouns and verbs make up a staggering 83% of all the lemma signs in Lusoga. The third most frequent group are the various pronouns (4% of the total), followed by the quantifiers (3%), adjectives (3%) and locatives (2%). The remaining 5% is made up of connectives (2%), interjections (1%), ideophones (1%) and adverbs (1%). A comparison with the values seen in Figure 4 is tempting, but faces at least two problems.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

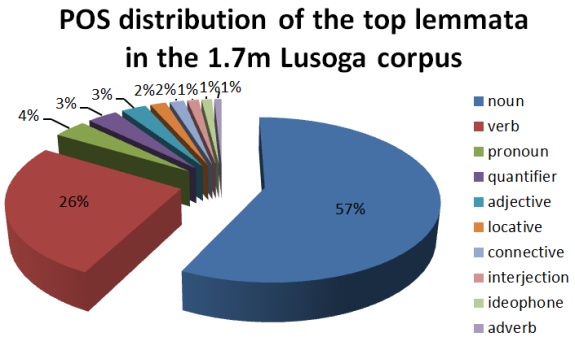

Figure 5: Pie chart showing the distribution of the parts of speech in the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

The first challenge is that the distributions across languages that belong to two very different language families are being compared. Even so, at the right-hand side of the graph seen in Figure 4, nouns and verbs already make up 74% of the total in English. The second challenge is that an unlemmatised distribution is compared to a lemmatised one. Indeed, as may be seen from Table 3, the original unlemmatised top-frequent 10 318 orthographic word forms (which includes some lower-frequent word forms from the closed-class parts of speech), as taken from the 1.7m Lusoga corpus, yielded a lemmatised frequency list of just 4 250 items.

Table 3: Statistics for the distribution of the parts of speech in the unlemmatised vs. lemmatised frequency lists derived from the top 10 000 types in the 1.7m Lusoga corpus

Part of speech	Unlemmatised	%	Lemmatised	%
verb	4 444	43.07%	1 113	26.19%
noun	3 622	35.10%	2 440	57.41%
adjective	1 105	10.71%	117	2.75%
pronoun	460	4.46%	156	3.67%
quantifier	231	2.24%	143	3.36%
locative	187	1.81%	75	1.76%
adverb	98	0.95%	35	0.82%
connective	68	0.66%	68	1.60%
interjection	54	0.52%	54	1.27%
ideophone	49	0.47%	49	1.15%
SUM	**10 318**	**100.00%**	**4 250**	**100.00%**

Expressed as a percentage of the total, three categories especially change their allocation drastically after lemmatisation. While verbs make up 43% of all the

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

top orthographic types in this Lusoga corpus, they only make up 26% after lemmatisation. Nouns do the reverse: they make up 35% of all the top orthographic types, but reach a massive 57% after lemmatisation. Adjectives go from nearly 11% down to about 3%. Unlemmatised and lemmatised part-of-speech distributions are thus different, as shown graphically in Figures 6 vs. 7.[8]

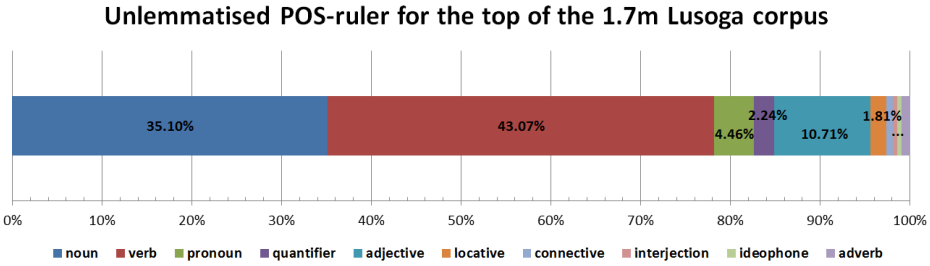

Figure 6: Part-of-speech ruler for the unlemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

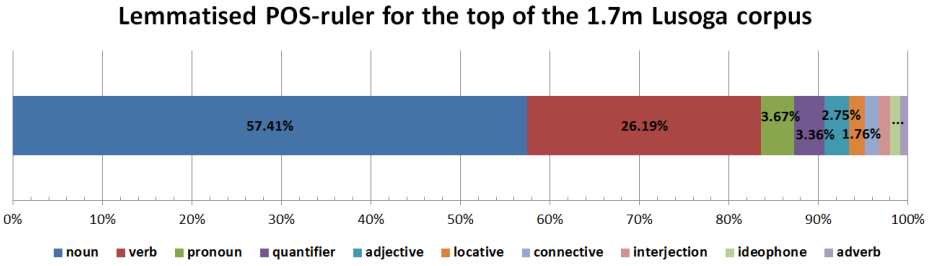

Figure 7: Part-of-speech ruler for the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

4.3 Contrasting part-of-speech rulers for Lusoga and Zulu

In order to judge whether the data seen in Table 2 and Figure 5 is plausible, it is instructive to compare the part-of-speech distribution for the Lusoga lemma signs with that for Zulu, as described in the corpus-based Zulu mini-grammar included in the *Oxford Bilingual School Dictionary: Zulu and English* (de Schryver 2010a: S13-S26) and summarised in Figure 8. On the Zulu to English side, this dictionary contains about 5 000 lemma signs (which were derived from the top section of a 7.5m general + 1m textbook Zulu corpus). This order of magnitude allows for comparisons with the 4 250 lemmatised forms which were obtained for Lusoga. While there are differences in the lemmatisation approach between the two languages, and even differences in categorising and naming the word classes, the overall picture seen for Zulu *may* be compared with that for

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Lusoga. At that point one realises that the two distributions are indeed rather similar, especially as regards nouns, with an allocation of 57% in Lusoga vs. 58% in Zulu. However, one does notice that there seems to be an exceptionally high number of verbs in Lusoga (26%) as compared to verbs in Zulu (16%).

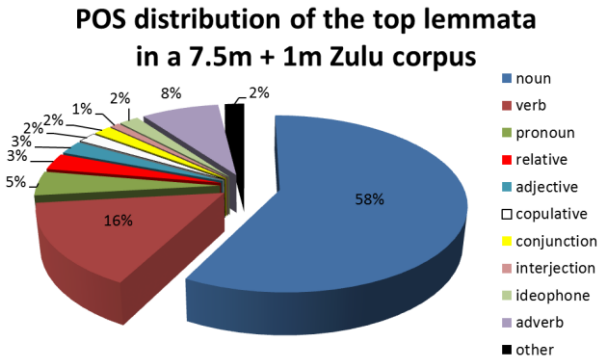

Figure 8: Part-of-speech distribution of the lemma signs in a corpus-based Zulu dictionary derived from the top types in a 7.5m general + 1m textbook Zulu corpus [adapted from de Schryver (2010a: S15)]

In these distributions, there are about ten main parts of speech ('main', as there are a number of sub-types as well) for both Lusoga and Zulu, but this could have been very different. The monolingual Zulu dictionary completed by the Zulu National Lexicography Unit (Mbatha 2006), for instance, uses just *four* parts of speech, following notions expounded in the PhD of Nkabinde (1975). Given the OUPSA Zulu school dictionary was meant to be as user-friendly as possible, such a drastic reduction of word classes was not entertained. The same holds for our decision regarding the word classes in Lusoga.

4.4 Using a part-of-speech ruler for Lusoga in dictionary planning

Using actual counts, Figures 6 and 7 can also be depicted as Figures 9 and 10 respectively. Of the two part-of-speech rulers, the lemmatised one is the most useful to support dictionary-making, hence Figure 10. The choice to lemmatise the top 10 000 orthographic words from the 1.7m Lusoga corpus was made in an attempt to arrive at a list of between 4 000 and 5 000 candidate lemma signs; we arrived at 4 250. If conceived in the way the OUPSA bilingual school dictionaries were conceived, then room must also be left for the inclusion of specialised vocabulary in the macrostructure, which is to be extracted from a separate, purpose-built specialised corpus. For Zulu, see de Schryver (2010b: 169), a concept based on the earlier de Schryver and Prinsloo (2003), where it was exemplified for Afrikaans. Basically, the Lusoga part-of-speech ruler seen in Figure 10 tells us that for a Lusoga dictionary of about 5 000 lemma signs, there

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

should/will be 2 440 nouns, 1 113 verbs, etc. down to 49 ideophones and 35 adverbs taken from the general language.

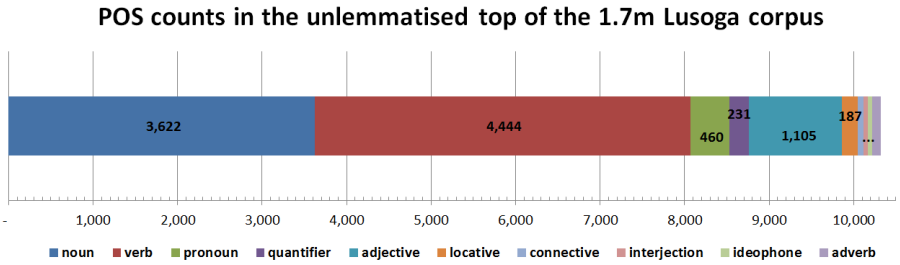

Figure 9: Counts per part of speech in the unlemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

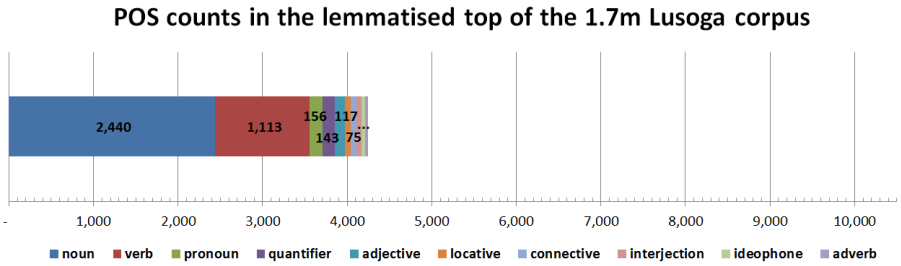

Figure 10: Counts per part of speech in the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

Knowing the (approximate) size of each word class in advance truly helps planning the actual dictionary work: equivalent and comparable chunks of the data may for instance be distributed to different team members, time extrapolations for the total work involved may be based on samples that were compiled for the different word classes, and dictionary-making itself may be organised and proceed 'by word class'. The latter has turned out to be an extremely important concept in Bantu lexicography, and may be spotted in the literature from article titles that refer to 'the lemmatisation of'-formula (de Schryver et al. 2004: 37). Taking Zulu as an example, the lemmatisation of nouns (Mpungose 1998, Prinsloo 2011), verbs (Prinsloo 2011), adjectives (de Schryver 2008b), pronouns (de Schryver 2008a, de Schryver and Wilkes 2008) and ideophones (de Schryver 2009), have all received attention in dedicated lexicographic studies, as have the treatment of terminological (Khumalo 2015) and cultural (Prinsloo and Bosch 2012) vocabulary.

Many problems in Bantu lexicography are part-of-speech dependent and need unique solutions that are different from one part of speech to the next.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Working through batches of a single word class during actual dictionary compilation therefore has ample advantages. In a dictionary-writing system like TLex, this is moreover fully supported: the part-of-speech tags that have been attached to the candidate lemma signs following lemmatisation (cf. §3) may first be used to isolate each word class as a group using the *Filter* tool, and that subset of the data may then be combined with any other filter parameters to allow for focused dictionary compilation.

5. From lemmatised frequency list to alphabetical distributions

5.1 Alphabetical rulers (aka 'multidimensional lexicographic rulers')

Some printed dictionaries have a thumb index per alphabetical category, either physically cut out in the pages or painted directly on the surface of the fore-edge, showing the progression of the different alphabetical categories, often in ladderised form. An alphabetical ruler is exactly that: an instrument which represents the relative allocation to each stretch of the alphabet. As a metalexicographical concept, such rulers were first introduced for Afrikaans (Prinsloo and de Schryver 2002a, 2003, de Schryver 2005, Prinsloo 2010, Taljard et al. 2017) and subsequently designed for all other official South African languages (de Schryver 2003, Prinsloo 2004, Prinsloo and de Schryver 2005, 2007).[9] Such rulers may be built from dictionary data, corpus data, or both. They may also be built to reflect the general language, or else a specific specialised domain of the language. In contrast to a part-of-speech ruler, an alphabetical ruler does not vary with corpus or dictionary sizes. The series of percentages per alphabetical stretch, for instance per alphabetical category, is very stable indeed, and the only difference one observes is between its lemmatised and unlemmatised versions.

Initially a 'measurement instrument', it quickly became clear that a ruler of this sort is also an 'evaluation instrument', as well as a 'prediction instrument', and ultimately even a 'management instrument' (de Schryver 2013). Given the many ways in which it can be used, such rulers have also been termed 'multidimensional lexicographic rulers'. Of the various uses, the one that interests us in the present contribution is as a prediction instrument, more specifically with the aim of predicting features of the compilation of a new Lusoga dictionary.

5.2 Towards an alphabetical ruler for Lusoga

From all the types in the full 1.7m Lusoga corpus as well as the unlemmatised and lemmatised frequency lists derived from the top 10 000 types (cf. §3), one can straightforwardly derive the data presented in Table 4. The three series of percentages represent general-language alphabetical rulers, and this in two unlemmatised environments and one lemmatised environment respectively.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Comparing the three distributions with one another, it is clear that there is a good correlation between the two unlemmatised ones, but no correlation between either of the unlemmatised distributions and the lemmatised one.[10]

Table 4: Statistics for the distribution of the alphabetical categories in the 1.7m Lusoga corpus as well as the unlemmatised and lemmatised frequency lists derived from the top 10 000 types

	Unlemmatised		Unlemmatised		Lemmatised	
Section	all corpus types	%	top corpus types	%	lemma signs from top	% = ABC-ruler
A	20 569	10.55%	1 152	11.16%	147	**3.46%**
B	25 030	12.83%	1 265	12.26%	368	**8.66%**
C	1 150	0.59%	5	0.05%	5	**0.12%**
D	3 089	1.58%	106	1.03%	83	**1.95%**
E	19 569	10.03%	1 354	13.12%	233	**5.48%**
F	643	0.33%	18	0.17%	78	**1.84%**
G	6 699	3.43%	260	2.52%	297	**6.99%**
H	830	0.43%	28	0.27%	24	**0.56%**
I	1 959	1.00%	187	1.81%	198	**4.66%**
J	309	0.16%	6	0.06%	5	**0.12%**
K	20 110	10.31%	1 116	10.82%	529	**12.45%**
L	4 462	2.29%	267	2.59%	338	**7.95%**
M	13 373	6.86%	933	9.04%	257	**6.05%**
N	14 425	7.40%	664	6.44%	277	**6.52%**
O	27 210	13.95%	1 720	16.67%	82	**1.93%**
P	1 126	0.58%	39	0.38%	84	**1.98%**
Q	36	0.02%	0	0.00%	0	**0.00%**
R	756	0.39%	3	0.03%	3	**0.07%**
S	2 032	1.04%	86	0.83%	374	**8.80%**
T	16 685	8.56%	453	4.39%	298	**7.01%**
U	415	0.21%	13	0.13%	14	**0.33%**
V	306	0.16%	10	0.10%	55	**1.29%**
W	4 028	2.07%	211	2.04%	202	**4.75%**
X	16	0.01%	0	0.00%	0	**0.00%**
Y	9 978	5.12%	411	3.98%	200	**4.71%**
Z	227	0.12%	11	0.11%	99	**2.33%**
SUM	**195 032**	**100.00%**	**10 318**	**100.00%**	**4 250**	**100.00%**

The only alphabetical ruler that is relevant to lexicographic work for a Bantu language is obviously the lemmatised one, except, perhaps, for those rare cases where full orthographic words are presented as lemma signs, including for all the verbs, as has been done for an experimental online Swahili dictionary

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

(Hillewaert and de Schryver 2004). Therefore, 'the' alphabetical ruler for Lusoga is as shown in Figure 11.[11]

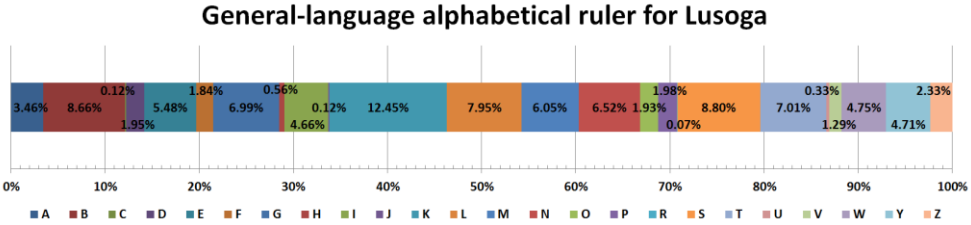

Figure 11: General-language alphabetical ruler based on the lemmatised frequency list derived from the top 10 000 types in the 1.7m Lusoga corpus

5.3 Contrasting alphabetical rulers for Lusoga and Zulu

The alphabetical ruler for Lusoga may be compared to the alphabetical ruler for Zulu that was used for the OUPSA Zulu school dictionary (de Schryver 2010a), shown in Figure 12.

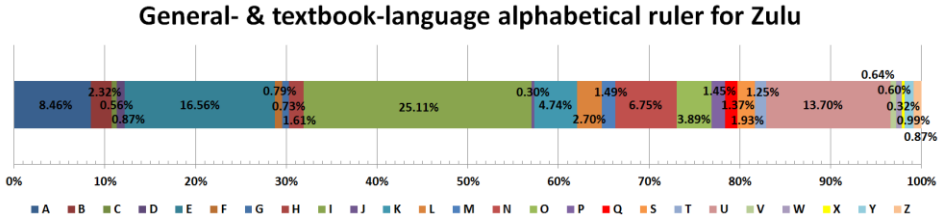

Figure 12: Alphabetical distribution of the lemma signs in a corpus-based Zulu dictionary derived from the top types in a 7.5m general corpus + 1m textbook Zulu corpus

As one may see, the two alphabetical rulers look very different indeed. This is because a decision was made in the Zulu dictionary to present full words for all parts of speech except verbs, on that account breaking with the stem tradition for this language. As a result of Zulu's pre-prefixes especially at nouns, the alphabetical categories A, I and U are massive, as is the alphabetical category E which contains the many locativised nouns for which the 'e-/o-...-ini locativisation strategy' was used (de Schryver and Gauton 2002).

Atypical alphabetical distributions such as the one seen in Figure 12 should remind every prospective compiler of a Bantu-language dictionary that careful thought should be put into who the envisaged target user group is. Reasoning back from the target user group, this then leads to a decision on pres-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

entation. Given that the Zulu dictionary was meant for school-going pupils, the goal was to present the material in as user-friendly a manner as possible, hence the decision to present words rather than stems for most parts of speech. Reasoning further back, from presentation to the actual lemmatisation required to achieve that presentation, one realises that there is always a direct link between target user group and lemmatisation approach, and vice versa. Relating this to the candidate Lusoga lemma-sign list means that the target user group envisaged is one that will be able to handle the lookup of word stems.

5.4 Using an alphabetical ruler for Lusoga in dictionary planning

Although the backbone of an alphabetical ruler is merely a single list of percentages totalling one hundred, it is a powerful instrument. From §5.2 it follows that the distribution of the number of (general-language) lemma signs per alphabetical category in Lusoga is not only according to the alphabetical ruler, but even the exact counts for each category are a given, and may be depicted as shown in Figure 13.

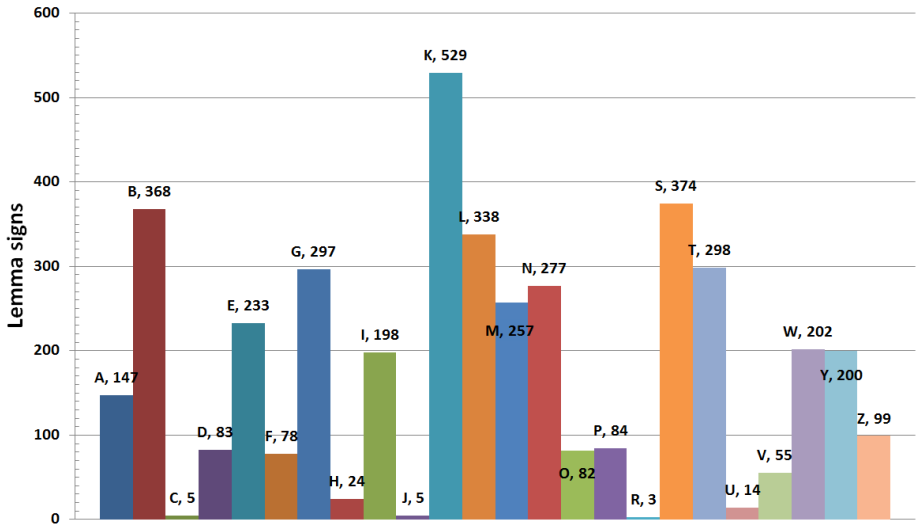

Figure 13: Distribution of the (general-language) lemma signs per alphabetical category in a planned Lusoga dictionary (sum: 4 250 lemma signs)

What is more, the actual lemma signs themselves are waiting in TLex, together with a brief preliminary meaning for each.

The alphabetical ruler may also be used to do some advance planning as far as dictionary size is concerned. Suppose a dictionary publisher envisages a central text for one side of the dictionary of 350 pages, then this ruler may

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

straightforwardly be used to predict the page allocation to each alphabetical category, as shown in Figure 14. Evidently, the presentation shown in Figure 14 is none other than the alphabetical ruler itself, hence Figure 11, now with a different *x*-axis.

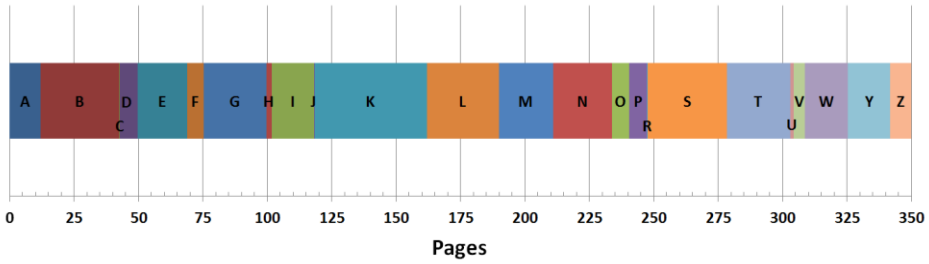

Figure 14: Distribution of the number of pages per alphabetical category in a planned Lusoga dictionary (aim: 350 pages for one side)

As a last example of the use of an alphabetical ruler as a prediction instrument, suppose the dictionary team wishes to work 'through the alphabet' (rather than, say, by word class), and that two years are available for the compilation of the central text, then Figure 15 predicts in which week which alphabetical category should be reached.

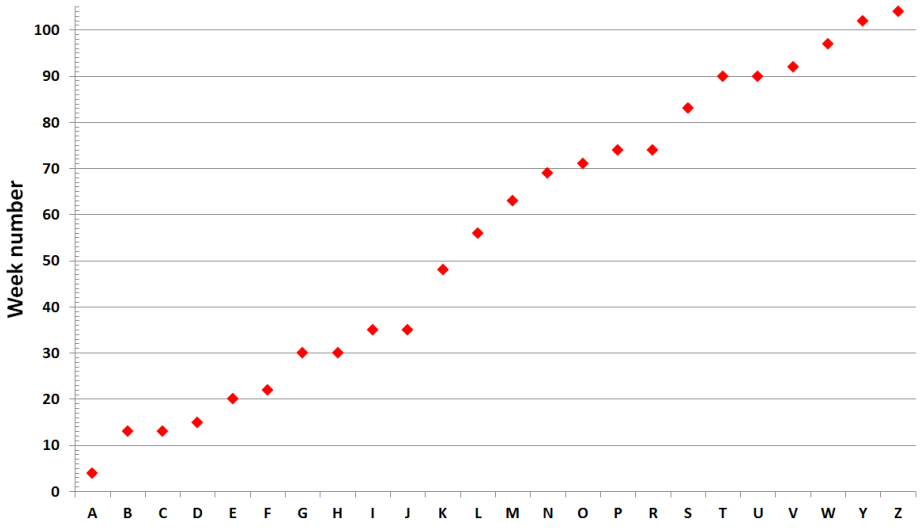

Figure 15: Projected progress through the alphabet for a planned Lusoga dictionary (aim: 2 years, or 104 weeks)

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

The underlying data for Figures 13 to 15 is shown in Table 5, but it should be clear that the alphabetical ruler may be used in any other creative way; for some of these, see the references in §5.1.

Table 5: Multidimensional predictions on lemma-sign, page and time levels for a planned Lusoga dictionary, using an alphabetical ruler for Lusoga

Section	ABC-ruler	Lemma signs	Pages	Reached in week	Days	...
A	3.46%	147	12.1	4	18.0	
B	8.66%	368	30.3	13	45.1	
C	0.12%	5	0.4	13	0.6	
D	1.95%	83	6.8	15	10.2	
E	5.48%	233	19.2	20	28.6	
F	1.84%	78	6.4	22	9.6	
G	6.99%	297	24.5	30	36.4	
H	0.56%	24	2.0	30	2.9	
I	4.66%	198	16.3	35	24.3	
J	0.12%	5	0.4	35	0.6	
K	12.45%	529	43.6	48	64.8	
L	7.95%	338	27.8	56	41.4	
M	6.05%	257	21.2	63	31.5	
N	6.52%	277	22.8	69	34.0	
O	1.93%	82	6.8	71	10.1	
P	1.98%	84	6.9	74	10.3	
R	0.07%	3	0.2	74	0.4	
S	8.80%	374	30.8	83	45.8	
T	7.01%	298	24.5	90	36.5	
U	0.33%	14	1.2	90	1.7	
V	1.29%	55	4.5	92	6.7	
W	4.75%	202	16.6	97	24.8	
Y	4.71%	200	16.5	102	24.5	
Z	2.33%	99	8.2	104	12.1	
SUM	**100.00%**	**4 250**	**350**		**521**	

6. Discussion

In this article we have illustrated how a lemmatised frequency list may be built directly within a dictionary-writing system like TLex, using as input plain orthographic words with occurrence frequencies as generated by corpus-query software like WordSmith Tools. These specific software programs are not crucial to the procedure, but they have been employed a number of times now and have proven their worth. Comparable programs will also do; what is important

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

to remember from the text is the necessary steps. The procedure is a mostly manual process, which needs to take the future target user group into account, and a process whereby all details are logged so that instant use may be made of two types of rulers: a part-of-speech ruler and an alphabetical ruler. A Lusoga corpus that was presented in the first of our three linked articles was processed to demonstrate the actual workings, and comparisons were also made with a completed Zulu dictionary project.

Honesty compels us to admit that the procedure described is the 'ideal' one, however. In actual practice, given that corpus data had to be analysed before it could be *explained* — and that the part-of-speech tagging and lemmatisation were merely the first steps of the analysis — even a seemingly basic task such as pinpointing the part(s) of speech of an orthographic word form was not that trivial. To start any analysis one needs a way to create order first, by grouping related material. But from the moment one starts to group material, one has already made a decision on how to analyse that material, as part-of-speech assignment is dependent on the framework or theory of the analysis. Conversely, without any advance decisions, one cannot begin to group and so can never get to any analysis. This chicken-and-egg conundrum was partly solved by falling back on received knowledge regarding the Bantu languages, as for instance summarised in handbooks such as that of Nurse and Philippson (2003) or the earlier ones of Guthrie (1948, 1953), Doke (1954) and Bryan (1959). Furthermore, as the analysis of the corpus material proceeded, we *did* go back to material that had already been completed in the TLex file, retagged some of the material, and reran the Lua scripts in order to generate an 'update' of the lemmatised frequency list.

Reformulated, even the mere act of labelling certain word forms as demonstratives or possessives, and considering these under the wider umbrella of pronouns, already crosses the line from analysis to explanation. That said, despite the received knowledge, we have tried to stick as much as possible to what we could observe in the corpus data, by also looking at the wider context and thus by avoiding limiting our look at words in isolation. With this we are now ready for the next step, the actual explanation of the material.

Acknowledgements

The research for this article was funded by the Special Research Fund of Ghent University. Thanks are due to the two anonymous referees.

Endnotes

1. For more on the difference between conjunctive and disjunctive writing systems in Bantu, see Prinsloo and de Schryver (2002b).

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

2. Whether or not this is an error actually depends on the lemmatisation strategy chosen. In Nguni lexicography, there is a 'stem tradition' (Ziervogel 1965, Van Wyk 1995), so if one also presents both nouns and verbs under the same stems (where relevant), then one could indeed lump their frequencies as well. Conversely, there is an argument to be made to keep the frequencies of different parts of speech separate, thereby leaving some presentation options open until actual dictionary compilation. In this regard, Prinsloo (1991), in the very-first exploratory study of the use of frequency counts for Bantu-language dictionary-making, did point out: 'It is very important to note that the interpretation of the output of a word frequency study is closely related to the lexicographical approach and the editorial policy from which the lexicographer embarked' (Prinsloo 1991: 59). The section from which this sentence is taken, 'Frequency studies in perspective' (Prinsloo 1991: 59-60), actually deals with lemmatisation options/decisions, even though Prinsloo does not use the term nor concept of lemmatisation.
3. Incidentally, the grammars included as middle matter in these dictionaries are furthermore the first corpus-based mini-grammars for any Bantu language, as described in de Schryver and Taljard (2007) for Northern Sotho, and de Schryver (2010b) for Zulu.
4. This is shown quite literally in Figure 1, where the data is sorted on the field 'Rank', so one truly moves from most frequent to least frequent. Another option is to use filters to extract the top-frequent section from the database, to work on in alphabetical order (or in any other, even random, order).
5. For more on the advantages, see for instance de Schryver et al. (2004), de Schryver (2008a, 2008b), de Schryver and Wilkes (2008) and de Schryver (2009).
6. When quick-and-rough frequencies were not provided, a Lua script (cf. further) would take care of this aspect, by automatically distributing the frequencies equally as a first approach (subject to correction later).
7. Junk was not tagged but deleted. Material with a poor spread across the sources was flagged as such, indicating that it may require a label.
8. The Pearson product moment correlation coefficient *r* between the unlemmatised and lemmatised part-of-speech distributions is 0.85.
9. The concept of an alphabetical ruler may be traced back to the 'block system of distribution of dictionary entries by initial letters' prepared for English by Edward L. Thorndike during the 1950s (Landau 2001: 360-362). Thorndike divided the alphabet into 105 blocks: 6 for A (A1: a-adk, A2: adl-alh, A3: ali-angk, ...), ... 1 for J (J50: j-jz), ... 3 for W (... , W104: wit-wz) and 1 for XYZ (XYZ105: x-zz). With approximately the same weight assigned to each of those blocks, this series supposedly reflects the 'distribution of lexical units throughout the alphabet'. See also Jackson (2002: 163-164), Moon (2004: 649-650) and Svensén (2009: 406).
10. The Pearson product moment correlation coefficient *r* between the two unlemmatised alphabetical distributions is an excellent 0.97; while it is just 0.56 between the full unlemmatised distribution and the lemmatised distribution, and 0.49 between the top unlemmatised distribution and the lemmatised distribution.
11. Observe that the letters c, j, q, r and x are not native to Lusoga, but may appear in borrowed abbreviations, place names and surnames, and the like.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

References

BNC. 1994–2018. British National Corpus. Available online at: http://www.natcorp.ox.ac.uk/.

Bosch, S.E. and L. Pretorius. 2003. Building a Computational Morphological Analyser/Generator for Zulu Using the Xerox Finite-State Tools. *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics*: 27-34. Budapest: ACL.

Bosch, S.E. and L. Pretorius. 2004. Software Tools for Morphological Tagging of Zulu Corpora and Lexicon Development. *Proceedings of the 4th International Language Resources and Evaluation Conference*: 1251-1254. Lisbon: ARTIPOL.

Bosch, S.E., L. Pretorius and A. Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies* 17(2): 66-88.

Bryan, M.A. 1959. *The Bantu Languages of Africa* (Handbook of African Languages 4). London: Oxford University Press (for the International African Institute).

De Pauw, G., G.-M. de Schryver and J. van de Loo. 2012. Resource-light Bantu Part-of-speech Tagging. De Pauw, G., G.-M. de Schryver, M.L. Forcada, K. Sarasola, F.M. Tyers and P.W. Wagacha (Eds). 2012. *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SaLTMiL 8 — AfLaT 2012)*: 85-92. Istanbul: European Language Resources Association.

De Pauw, G., G.-M. de Schryver and P.W. Wagacha. 2006. Data-driven Part-of-Speech Tagging of Kiswahili. Sojka, P., I. Kopeček and K. Pala (Eds). 2006. *Text, Speech and Dialogue, 9th International Conference, TSD 2006, Brno, Czech Republic, September 11–15, 2006, Proceedings* (Lecture Notes in Artificial Intelligence (LNAI), subseries of Lecture Notes in Computer Science (LNCS) 4188): 197-204. Berlin: Springer-Verlag.

De Pauw, G., G.-M. de Schryver and P.W. Wagacha. 2006–18. AfLaT — African Language Technology. Available online at: http://aflat.org/.

de Schryver, G.-M. 1999. *Bantu Lexicography and the Concept of* Simultaneous Feedback*, Some Preliminary Observations on the Introduction of a New Methodology for the Compilation of Dictionaries with Special Reference to a Bilingual Learner's Dictionary* Cilubà-Dutch. Unpublished M.A. dissertation. Ghent: Ghent University.

de Schryver, G.-M. 2003. Drawing up the Macrostructure of a Nguni Dictionary, with Special Reference to isiNdebele. *South African Journal of African Languages* 23(1): 11-25.

de Schryver, G.-M. 2005. Concurrent Over- and Under-treatment in Dictionaries — The *Woordeboek van die Afrikaanse Taal* as a Case in Point. *International Journal of Lexicography* 18(1): 47-75.

de Schryver, G.-M. 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English / Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford*. Cape Town: Oxford University Press Southern Africa.

de Schryver, G.-M. 2008a. The Lexicographic Treatment of Quantitative Pronouns in Zulu. *Lexikos* 18: 92-105.

de Schryver, G.-M. 2008b. A New Way to Lemmatize Adjectives in a User-friendly Zulu–English Dictionary. *Lexikos* 18: 63-91.

de Schryver, G.-M. 2009. The Lexicographic Treatment of Ideophones in Zulu. *Lexikos* 19: 34-54.

de Schryver, G.-M. 2010a. *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford*. Cape Town: Oxford University Press Southern Africa.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

de Schryver, G.-M. 2010b. Revolutionizing Bantu Lexicography — A Zulu Case Study. *Lexikos* 20: 161-201.

de Schryver, G.-M. 2013. Tools to Support the Design of a Macrostructure. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (Handbooks of Linguistics and Communication Science, HSK 5.4): 1384-1395. Berlin: Walter de Gruyter.

de Schryver, G.-M. and R. Gauton. 2002. The Zulu Locative Prefix ku- Revisited: A Corpus-based Approach. *Southern African Linguistics and Applied Language Studies* 20(4): 201-220.

de Schryver, G.-M. and N.S. Kabuta. 1997. *Lexicon Cilubà–Nederlands, Een circa 2500-lemma's-tellend strikt alfabetisch geordend vertalend aanleerderslexicon met decodeer-functie ten behoeve van studenten Afrikaanse Talen & Culturen aan de Universiteit Gent* (Recall Linguistics Series 1). Ghent: Recall.

de Schryver, G.-M. and N.S. Kabuta. 1998. *Beknopt woordenboek Cilubà–Nederlands & Kalombodi-mfùndilu kàà Cilubà (Spellingsgids Cilubà), Een op gebruiks-frequentie gebaseerd vertalend aanleerderslexicon met decodeerfunctie bestaande uit circa 3.000 strikt alfabetisch geordende lemma's & Mfùndilu wa myakù ìdì ìtàmba kumwèneka (De orthografie van de meest gangbare woorden)* (Recall Linguistics Series 12). Ghent: Recall.

de Schryver, G.-M. and D.J. Prinsloo. 2000. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The Macrostructure. *South African Journal of African Languages* 20(4): 291-309.

de Schryver, G.-M. and D.J. Prinsloo. 2001. Corpus-based Activities versus Intuition-based Compilations by Lexicographers, the Sepedi Lemma-sign List as a Case in Point. *Nordic Journal of African Studies* 10(3): 374-398.

de Schryver, G.-M. and D.J. Prinsloo. 2003. Compiling a Lemma-sign List for a Specific Target User Group: The Junior Dictionary as a Case in Point. *Dictionaries: Journal of The Dictionary Society of North America* 24: 28-58.

de Schryver, G.-M. and M. Reynolds. 2014. *Oxford Bilingual School Dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama sesikolo*. Cape Town: Oxford University Press Southern Africa.

de Schryver, G.-M. and E. Taljard. 2007. Compiling a Corpus-based Dictionary Grammar: An Example for Northern Sotho. *Lexikos* 17: 37-55.

de Schryver, G.-M., E. Taljard, M.P. Mogodi and S. Maepa. 2004. The Lexicographic Treatment of the Demonstrative Copulative in Sesotho sa Leboa — An Exercise in Multiple Cross-referencing. *Lexikos* 14: 35-66.

de Schryver, G.-M. and A. Wilkes. 2008. User-friendly Dictionaries for Zulu: An Exercise in Complexicography. Bernal, E. and J. DeCesaris (Eds). 2008. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)* (Sèrie Activitats 20): 827-836. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Doke, C.M. 1954. *The Southern Bantu Languages* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).

Guthrie, M. 1948. *The Classification of the Bantu Languages* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Guthrie, M. 1953. *The Bantu Languages of Western Equatorial Africa* (Handbook of African Languages). London: Oxford University Press (for the International African Institute).

Hanks, P. 2001. The Probable and the Possible: Lexicography in the Age of the Internet. Lee, S. (Ed.). 2001. *ASIALEX 2001 Proceedings, Asian Bilingualism and the Dictionary*: 1-15. Seoul: Center for Linguistic Informatics Development, Yonsei University.

Hillewaert, S. and G.-M. de Schryver. 2004. Online Kiswahili (Swahili) — English Dictionary. Available online at: http://africanlanguages.com/swahili/.

Hurskainen, A. 1992. A Two-level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122.

Hurskainen, A. 2016. Helsinki Corpus of Swahili 2.0 (HCS 2.0) Annotated Version. Available online at: http://urn.fi/urn:nbn:fi:lb-2016011301.

Jackson, H. 2002. *Lexicography: An Introduction*. London: Taylor & Francis Routledge.

Joffe, D. and G.-M. de Schryver. 2002–18. TLex Suite — Dictionary Compilation Software. Available online at: http://tshwanedje.com/tshwanelex/.

Joffe, D. and G.-M. de Schryver. 2005. Representing and Describing Words Flexibly with the Dictionary Application TshwaneLex. Ooi, V.B.Y., A. Pakir, I. Talib, L. Tan, P.K.W. Tan and Y.Y. Tan (Eds). 2005. *Words in Asian Cultural Contexts, Proceedings of the 4th Asialex Conference, 1–3 June 2005, M Hotel, Singapore*: 108-114. Singapore: Department of English Language and Literature & Asia Research Institute, National University of Singapore.

Khumalo, L. 2015. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary. *Lexikos* 25: 495-506.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.

Knowles, F. 1983. Towards the Machine Dictionary, 'Mechanical' Dictionaries. Hartmann, R.R.K. (Ed.). 1983. *Lexicography: Principles and Practice*: 181-193. London: Academic Press.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography (2nd edition)*. Cambridge: Cambridge University Press.

Mbatha, M.O. 2006. *Isichazamazwi sesiZulu*. Pietermaritzburg: New Dawn Publishers.

Moon, R. 2004. Cawdrey's A Table Alphabeticall: A Quantitative Approach. Williams, G. and S. Vessier (Eds). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 639-650. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Mpungose, M.H. 1998. Analysis of the Word-initial Segment with Reference to Lemmatising Zulu Nasal Nouns. *Lexikos* 8: 65-87.

Nabirye, M. 2016. *A Corpus-based Grammar of Lusoga*. Unpublished Ph.D. dissertation. Ghent: Ghent University.

Nkabinde, A.C. 1975. *A Revision of the Word Categories in Zulu*. Unpublished PhD dissertation. Pretoria: UNISA.

Nurse, D. and G. Philippson (Eds). 2003. *The Bantu Languages* (Language Family Series 4). London: Routledge.

Prinsloo, D.J. 1991. Towards Computer-assisted Word Frequency Studies in Northern Sotho. *South African Journal of African Languages* 11(2): 54-60.

Prinsloo, D.J. 2004. Revising Matumo's *Setswana–English–Setswana Dictionary*. *Lexikos* 14: 158-172.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Prinsloo, D.J. 2010. Die verifiëring, verfyning en toepassing van leksikografiese liniale vir Afrikaans. *Lexikos* 20: 390-409.

Prinsloo, D.J. 2011. A Critical Analysis of the Lemmatisation of Nouns and Verbs in isiZulu. *Lexikos* 21: 169-193.

Prinsloo, D.J. and S.E. Bosch. 2012. Kinship Terminology in English–Zulu / Northern Sotho Dictionaries — A Challenge for the Bantu Lexicographer. Fjeld, R.V. and J.M. Torjusen (Eds). 2012. *Proceedings of the 15th EURALEX International Congress, 7–11 August, 2012, Oslo*: 296-303. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Prinsloo, D.J. and G.-M. de Schryver. 2002a. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, A. and C. Povlsen (Eds). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*: 483-494. Copenhagen: Center for Sprogteknologi, Københavns Universitet.

Prinsloo, D.J. and G.-M. de Schryver. 2002b. Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages. *Nordic Journal of African Studies* 11(2): 249-265.

Prinsloo, D.J. and G.-M. de Schryver. 2003. Effektiewe vordering met die *Woordeboek van die Afrikaanse Taal* soos gemeet in terme van 'n multidimensionele Liniaal. Botha, W. (Ed.). 2003. *'n Man wat beur: Huldigingsbundel vir Dirk van Schalkwyk*: 106-126. Stellenbosch: Bureau of the WAT.

Prinsloo, D.J. and G.-M. de Schryver. 2005. Managing Eleven Parallel Corpora and the Extraction of Data in All Official South African Languages. Daelemans, W., T. du Plessis, C. Snyman and L. Teck (Eds). 2005. *Multilingualism and Electronic Language Management. Proceedings of the 4th International MIDP Colloquium, 22–23 September 2003, Bloemfontein, South Africa* (Studies in Language Policy in South Africa 4): 100-122. Pretoria: Van Schaik Publishers.

Prinsloo, D.J. and G.-M. de Schryver. 2007. Crafting a Multidimensional Ruler for the Compilation of Sesotho sa Leboa Dictionaries. Mojalefa, M.J. (Ed.). 2007. *Rabadia Ratšhatšha: Studies in African Language Literature, Linguistics, Translation and Lexicography*: 177-201. Stellenbosch: SUN PReSS.

Scott, M. 1996–2018. WordSmith Tools. Available online at: http://www.lexically.net/wordsmith/.

Sinclair, J.M. 1995. *Collins Cobuild English Dictionary*. London: HarperCollins.

Svensén, B. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

Taljard, E., D.J. Prinsloo and N. Bosman. 2017. Honderd jaar *Afrikaanse Woordelys en Spelreëls* — 'n oorsig en waardering. Deel 2: Die gebruiker in fokus. *Tydskrif vir Geesteswetenskappe* 57(2.1): 302-322.

Van Wyk, E.B. 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos* 5: 82-96.

Ziervogel, D. 1965. Die probleme van leksikografie in die Suid-Afrikaanse Bantoetale. *Taalfasette* 1(1): 45-53.

Zipf, G.K. 1935. *The Psycho-Biology of Language*. Boston, MA: Houghton Mifflin Co.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Addendum 1: Lua script: GenerateSecondSide.lua

```
-- 2013-09 Lusoga collapse LemmaForm into second 'side' of dictionary database
-- and add up frequencies
-- David Joffe

-- 'CONSTANTS' (script configuration - if e.g. attribute names change, update
-- script here)
CFG =
{
   ATTR_POS   = "PartOfSpeech",
   -- Part of speech
   ATTR_LEMMAFORM   = "LemmaForm",
   ATTR_FREQ   = "CalculatedFrequency",
   -- Recalculated frequency attribute (that incorporates homonym percentage).
   -- For reading the value from the corpus list.
   ATTR_FREQUENCY   = "Frequency",
   -- Actual "Frequency" attribute (not re-calculated one that incorporates
   -- percentage). For setting frequency on created entries.
   ATTR_INCOMPLETE   = "Incomplete",
   -- Section 0-based index with source list (i.e. corpus forms)
   SECTION_SRC   = 0,
   -- Section 0-based index for creating collapsed forms (e.g. "-ba")
   SECTION_DEST   = 1
}

local DOC=tApp():GetCurrentDoc();
if DOC==nil then return "";end

-- STATS
local nNumForms=0;
local nNumCreated=0;
local nNumExistingModified=0;
local SECTION=DOC:GetDictionary():GetLanguage( CFG.SECTION_SRC );
local SECTIONDEST=DOC:GetDictionary():GetLanguage( CFG.SECTION_DEST );
local i;
local data={}
for i=0,SECTION:GetNumEntries()-1,1 do
   local ENTRY=SECTION:GetEntry(i);
   local bDoEntry=false;
   local incomplete= tQuery(ENTRY,"/@"..CFG.ATTR_INCOMPLETE);
   if (incomplete=="") or (incomplete=="0") then
      bDoEntry = true;
   end
```

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458


```
   if (bDoEntry) then
      local pos       = tQuery(ENTRY,"/@"..CFG.ATTR_POS);
      local lemmaform   = tQuery(ENTRY,"/@"..CFG.ATTR_LEMMAFORM);
      local freqs     = tQuery(ENTRY,"/@"..CFG.ATTR_FREQ);
      -- Can return nil on empty string, so check for nil next and set
      -- to 0 in that case
      local freq = tonumber(freqs);
      if (freq==nil) then
         freq= 0;
      end

      tLuaLog("FORM:"..lemmaform
        .."(" .. ENTRY:GetLemmaSign()..") pos="..pos.." freq="..freq)

      -- Make a unique string that is the combination of LemmaForm and the
      -- partofspeech (e.g. "-ba_$$$_noun") .. this separator string must just
      -- be some string that doesn't occur in the actual data ever, but other
      -- than that it's arbitrary
      if ( data[ lemmaform .. "_$$$_" .. pos ] == nil ) then
         --data[ lemmaform .. "_$$$_" .. pos ] =
         --{ lemmaform, pos, tonumber(freq) };--ADD NEW
         data[ lemmaform .. "_$$$_" .. pos ] = { }
         data[ lemmaform .. "_$$$_" .. pos ][1] = lemmaform;
         data[ lemmaform .. "_$$$_" .. pos ][2] = pos;
         data[ lemmaform .. "_$$$_" .. pos ][3] = freq;
      else
         -- ADD UP FREQUENCIES (TO EXISTING) (note Lua arrays = 1-based index)
         data[ lemmaform .. "_$$$_" .. pos ][3] =
             data[ lemmaform .. "_$$$_" .. pos ][3] + tonumber(freq);
      end
      nNumForms = nNumForms + 1;
   end--bDoEntry
end

for key,value in pairs(data) do
   local lemmaform = value[1];
   local pos = value[2];
   local freq = value[3];
   tLuaLog("FINAL:"..lemmaform..","..pos..","..freq)

   -- See if there is an existing entry of this form and part of speech
   local ENTRY = nil;
   local CURRENT = SECTIONDEST:FindEntries( lemmaform );
   for i=0,CURRENT:size()-1,1 do
```

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458


```
      if (tQuery(CURRENT[i],"/@"..CFG.ATTR_POS) == pos) then
         ENTRY = CURRENT[i];
      end
   end

   -- If no existing entry, create a new one
   if ENTRY==nil then
      local NODE = DOC:AllocateElementByID(NODE_ENTRY,true);-- Alloc new entry
      ENTRY = tolua.cast(NODE, "tcEntry");
      ENTRY:SetLemmaSign( lemmaform );

      SECTIONDEST:InsertEntry(ENTRY);

      nNumCreated = nNumCreated+1;
   else
      nNumExistingModified = nNumExistingModified + 1;
   end

   -- Set frequency, POS etc.
   local ATTR_FREQ=ENTRY:GetElement():FindAttributeByName(CFG.ATTR_FREQUENCY);
   if (ATTR_FREQ~=nil) then
      ENTRY:SetAttributeI( ATTR_FREQ, freq );
   end

   local ATTR_POS = ENTRY:GetElement():FindAttributeByName(CFG.ATTR_POS);
   if (ATTR_POS~=nil) then
      ENTRY:SetAttributeDisplayByString( ATTR_POS, pos, false,
         "____prevent_unintentional_list_string_splitting___" );
   end
end
data=nil;
Evt_LemmasInserted:Trigger(nil, SECTIONDEST);--Update UI etc.
DOC:SetDirty();

local sRetMessage =
   "FORMS: ".. nNumForms ..
   " CREATED: " .. nNumCreated..
   " EXISTING_UPDATED: " .. nNumExistingModified
   ;
return sRetMessage;
```

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Addendum 2: Lua script: AssignRankBasedOnSortBy.lua

```
-- 2013-10 Assign numerical 'rank' based on sort order
-- (sort order defined by e.g. FIRST selecting F4 SortBy
-- 'Word::Frequency' in second section just before running this script)
-- David Joffe

-- 'CONSTANTS' (script configuration - if e.g. attribute names change,
-- update script here)
CFG =
{
   ATTR_RANK = "Rank", -- Rank
   SECTION = 1 -- Section 0-based index for generating ranks
}

local DOC=tApp():GetCurrentDoc();
if DOC==nil then return "No document"; end

-- STATS
local SECTION=DOC:GetDictionary():GetLanguage( CFG.SECTION );
if (SECTION==nil) then return "Invalid section index"; end

local SECTWND = tFrameWindow():GetLanguageWindow(SECTION);
if (SECTWND==nil) then
   return "No section window for section (try go out of expanded view mode)";
end

-- By default F4 SortBy puts highest frequency at bottom, so if so, invert rank
-- values set as we loop across entries (e.g. rank '1' would be the bottom entry
-- in the list if this is set to true)
local bInvertOrdering=true;

-- Iterate through (NB) the SECTION WINDOW entry list - so e.g. SortBy may be in
-- effect
local i;
local Attr=nil;
for i=0,SECTWND:GetNumLemmaListEntries()-1,1 do
   local ENTRY=SECTWND:GetLemmaListEntry(i);

   if Attr==nil then
      Attr = ENTRY:GetElement():FindAttributeByName( CFG.ATTR_RANK );
      if Attr==nil then return "Rank attribute not found"; end
   end
```

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458


```
   if bInvertOrdering then
      ENTRY:SetAttributeDisplayByString( Attr,
         SECTWND:GetNumLemmaListEntries() - i, false,
            "____prevent_unintentional_list_string_splitting___" );
   else
      ENTRY:SetAttributeDisplayByString( Attr,
         i+1, false,
            "____prevent_unintentional_list_string_splitting___" );
   end
end

-- Update user interface etc.
Evt_LemmasInserted:Trigger(nil, SECTION);
DOC:SetDirty();

return "";
```

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458

Addendum 3: Lua script: AssignFrequencyBandBasedOnRank.lua

```
--Frequency bands
local RANKS={
   {500,"1"},
   {1000,"2"},
   {1500,"3"},
   {5000,""},
   {999999999999999999,"LOW"}
};

--check that there is a global document object
--(i.e. make sure we actually have a dictionary open)
if g_pDoc == nil then
   --No document is loaded - exit
   return "Failed - No document loaded";
end

local LEMARRAY = {};
local LEMMAELEM = g_pDoc:GetDictionary():GetDTD():FindElementByName("Word");
local LEMMAFREQATTR = LEMMAELEM:FindAttributeByName("Frequency");
local LEMMAFREQBANDATTR = LEMMAELEM:FindAttributeByName("FrequencyBand");
local LEMMARANKATTR = LEMMAELEM:FindAttributeByName("Rank");

-- Change this 0 to 1 if doing a bilingual 2nd section (right half):
local LANG = g_pDoc:GetDictionary():GetLanguage(0);
local COUNT = 0;
for i=0,LANG:GetNumChildren()-1,1 do
   local LEMMA = LANG:GetChild(i);
   local FREQ = LEMMA:GetAttributeIIntValue(LEMMAFREQATTR);
   table.insert(LEMARRAY,{freq = FREQ,lem = LEMMA});
end

table.sort(LEMARRAY,
   function (a, b)
      return a["freq"] > b["freq"]
   end)

for j,k in pairs(LEMARRAY) do
   for v,d in pairs(RANKS) do
      if COUNT < d[1] then
         --First have to clear the original selection (in case we want an
         --empty value for any of the ranges) - SetAttribute*() does not
         --set the list value at all if passed a null string
```

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1458


```
            k["lem"]:SetAttributeListID(LEMMAFREQBANDATTR,0);
            k["lem"]:SetAttributeDisplayByString(LEMMAFREQBANDATTR,d[2]);
            --k["lem"]:SetAttributeDisplayByString(LEMMARANKATTR,COUNT+1);
            break;
         end
      end
      COUNT = COUNT + 1;
end

g_pDoc:SetDirty();

--script terminated without error
return "done";
```