Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus

D.J. Prinsloo, *Department of African Languages, University of Pretoria, Pretoria, South Africa (danie.prinsloo@up.ac.za)*

Abstract: This article focuses on lesser-resourced languages for which only very limited corpora are available and how such relatively small and often unbalanced, raw corpora could be maximally utilized for lexicographic purposes to obtain similar results as for bigger corpora. Sepedi and Afrikaans will be studied in this regard. The aim is to determine to what extent enlarging a corpus from e.g. one to 10 million, and from 10 million to 100 million words enhances its potential for (a) macrostructure compilation, (b) sourcing information on the most important microstructural aspects and (c) the creation of lexicographic tools. It will be argued that valuable and even sufficient data for the compilation of a specific dictionary can be extracted from a relatively small corpus of approximately one million words but that "bigger" in some instances indeed means "better".

Keywords: CORPUS-BASED LEXICOGRAPHY, LESSER-RESOURCED LANGUAGES, LIMITED CORPORA, CORPUS TOOLS, LEXICOGRAPHIC TOOLS

Opsomming: Korpusgebaseerde leksikografie vir hulpbronbeperkte tale — die maksimalisering van die beperkte korpus. Die fokus in hierdie artikel is op hulpbronbeperkte tale waarvoor slegs baie beperkte korpusse beskikbaar is en hoe sodanige relatief klein en dikwels ongebalanseerde, rou korpusse maksimaal benut kan word vir leksikografiese doeleindes om soortgelyke resultate as van groter korpusse te verkry. Sepedi en Afrikaans, word in hierdie verband bestudeer. Die doel is om te bepaal tot watter mate die vergroting van 'n korpus van byvoorbeeld een na 10 miljoen, en van 10 miljoen na 100 miljoen woorde die potensiaal sal verhoog vir (a) makrostruktuur samestelling, (b) die inwin van inligting omtrent die belangrikste mikrostrukturele aspekte en (c) die ontwerp van leksikografiese hulpmiddels. Daar sal aangevoer word dat waardevolle en selfs voldoende data vir die samestelling van 'n spesifieke woordeboek onttrek kan word uit 'n relatief klein korpus van ongeveer een miljoen woorde maar dat "groter" wel in sekere omstandighede "beter" is.

Sleutelwoorde: KORPUSGEBASEERDE LEKSIKOGRAFIE, HULPBRONBEPERKTE TALE, BEPERKTE KORPUSSE, KORPUSGEREEDSKAP, LEKSIKOGRAFIESE HULPMIDDELS

Introduction

The days of a default corpus size of one million words such as the groundbreaking first computer-readable general text corpus, the *Brown Corpus of Stan-*

dard American English being regarded as an acceptable norm, are long gone. Currently corpora for major languages typically run into hundreds of millions and even billions of words, for example *Google Books* with 155 billion for American English, 45 billion for Spanish and 34 billion for British English, and are typically referred to as "big corpora".

In many cases sincere attempts at corpus designs and the compilation of balanced and representative corpora reflecting stratified speaker groups have been made, e.g. in the compilation of the *Brown* corpus. Different levels of corpus annotation and sophisticated corpus manipulation tools e.g. *Sketch Engine*, *Dante*, *Interactive language Toolbox*, *WordSmith Tools* and *AntConc* became the norm as an international standard and represent the typical scenario for major languages of the world.

This article, however, focuses on lesser-resourced languages for which only very limited corpora are available and how such relatively small and often unbalanced, raw corpora could be maximally utilized for lexicographic purposes to obtain similar results in the absence of large corpora. It presents empirical research for Sepedi. English and Afrikaans corpora are used as measurement instruments to determine the power of limited corpora for lexicographic purposes.

"Big corpus" is a relative term. For lesser-resourced languages with a limited number of printed material such as many of the African languages, a corpus of 10 million words can be regarded as a "big corpus". The aim is to determine to what extent enlarging a corpus from e.g. one to 10 million, and from 10 million to 100 million words enhances its potential for (a) macrostructure compilation, (b) sourcing information on the most important microstructural aspects and (c) the creation of lexicographic tools. It will be argued that valuable and even sufficient data for the compilation of a specific dictionary can be extracted from a relatively small corpus of approximately one million words. The question is how much energy should be invested for lexicographic purposes in the maximum utilization of a limited corpus for macrostructural and microstructural compilation versus increasing the corpus size. Macrostructural compilation mainly concerns the compilation of the lemmalist and microstructural aspects include sense distinction, collocations, idioms and examples of usage.

English, Afrikaans and Sepedi corpora

For the purpose of this study corpora for English, Afrikaans and Sepedi were used. For English the *Pretoria English Internet Corpus* (PEIC) consisting of 12 million words and a subsection of approximately one million words were used. These corpora will be referred to as the 10m PEIC and 1m PEIC respectively. For Afrikaans a small section of the *Media 24* archive for the newspaper *Beeld* consisting of 119 million words as well as two subsections consisting of approximately 10 million and one million words respectively were used and will be referred to as 100m MED 24, 10m MED 24 and 1m MED 24 respectively.

For Sepedi a 10 million-word corpus and a one million subsection thereof were used and will be referred to as 10m PSC and 1m PSC respectively. The corpora and subsections of the corpora are schematically indicated and their exact sizes are given in figure 1:

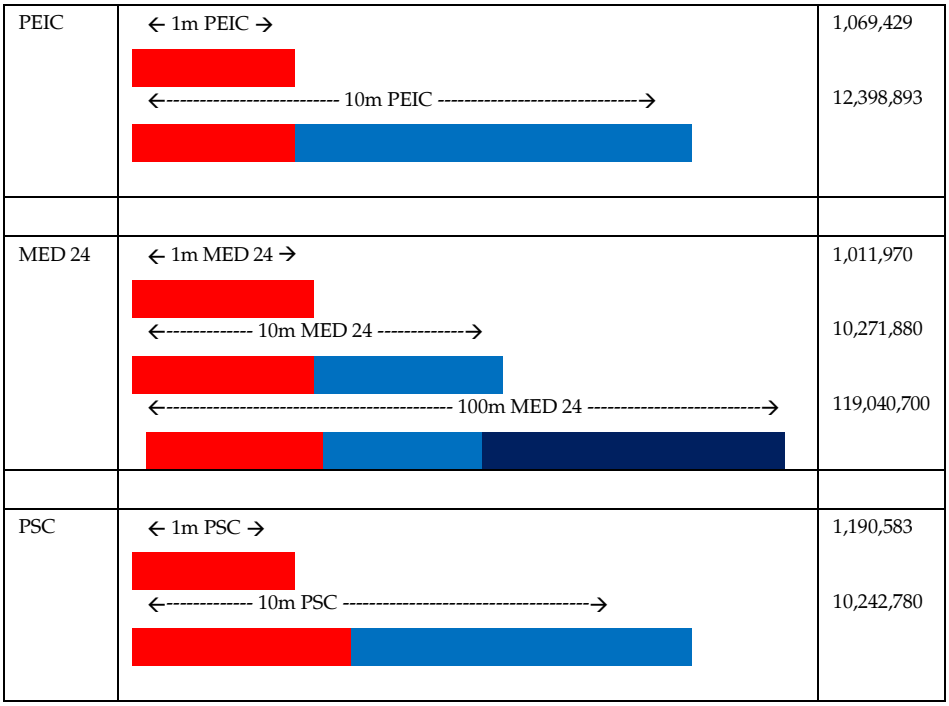

Figure 1: Corpora and sub-corpora used for English, Afrikaans and Sepedi

Macrostructure

In Africa publishers normally restrict dictionaries to a very limited number of pages. 5000 articles are often the norm and by necessity put the focus on commonly used words for inclusion in the dictionary. This study thus assumes that the basic/common words of a language are most likely to be looked for especially by learners of the language in such a small dictionary. These are the frequently used words typically marked by means of e.g. a star-rated system, filled diamonds, and/or by a different colour in dictionaries such as the *Macmillan English Dictionary* (MED), and *Collins COBUILD English Dictionary* (COBUILD), e.g. **car** … *** (MED) and **car** … **cars** ♦♦♦♦♦ (COBUILD). MED states that a word marked with three stars is one of the most basic words in English. COBUILD, as indicated in table 1, states that the 1,900 most frequently used words in the language, marked with four or five filled diamonds represent 75% of all written

and spoken words in English and that the top 14,700 words account for 95% of English words.

Number of filled diamonds	Lemmas per category	Totals	% of all written and spoken English
5	700		
4	1200		
(Total 5 + 4)		1900	75
3	1500		
2	3200		
1	8100		
(Total 3 + 2 + 1)		12800	20
(Total 5 + 4 + 3 + 2 + 1)		14700	95

Table 1: Summary of frequency band values in COBUILD (p. xiii)

On the macrostructural level an evaluation was made of frequency lists compiled from the 1m PEIC and 10m PEIC for English, the 1m MED 24, the 10m MED 24 and the 100m MED 24 for Afrikaans, and the 1m PSC and 10m PSC for Sepedi. The most basic words in English indicated with three stars (***) in MED were used as a benchmark against the 1m PEIC and 10m PEIC English corpora. There are 2,275 three-starred words in MED. Of these words 2,203 occur in the 31,982-word frequency list culled from the 1m PEIC; thus an overlap of 96.8%. Since it is hardly feasible for a lexicographer to work through a frequency list of this size when compiling a lemmalist, a more realistic number of words were considered, i.e. 11,559 which occurred five times or more in the corpus. 2,061 three-starred words in MED remained, i.e. an overlap of 90.6%. This means that the lexicographer who only had a one million English corpus at his/her disposal, and willing to read through a list of 11,000 words would be in a position to capture 90.6% of the most basic English words. A 90%+ figure can surely be regarded as quite a significant achievement on such a small corpus.

This experiment was repeated for the entire 10m PEIC. Of the 2,275 three-starred words in MED, 2,272 (only three not: e-mail, long-term and no-one), and with the exception of *metre* with a frequency of 1, appear in the 10m PEIC. All of these 3-starred words have a frequency count higher than 10 and occur in the 118,202-word frequency list of the 10m PEIC; thus an overlap of 99.9%. Once again, a more realistic number of words were considered, i.e. 11,161, which occurred 65 times or more in the corpus. 2,191 three-starred words in MED remained. This means that the lexicographer who only had a 10 million English corpus at his/her disposal, and willing to read through a list of 11,000

words would be in a position to capture 96.3% of the most basic English words. Once again, a relatively small corpus of 10 million words enabled the lexicographer to capture the most basic words. It is also significant that a tenfold increase in the corpus size from one million to 10 million only resulted in a 5.7% increase in the three-starred words retained.

Consider table 2 as summary:

MED	**1m PEIC**	**10m PEIC**
2,275 (three-starred words)	2,203 MED *** in 1mPIC (overlap with MED ***): 2,061 = 90.6% (Lexicographer considers freq. >4) (11,559 words to consider)	2,272 MED *** in PEIC (overlap with MED ***): 2,191 = 96.3% (Lexicographer considers freq. >64) (11,161 words to consider)

Table 2: MED 3-starred words versus the 1m PEIC and the 10m PEIC

For the Afrikaans experiment the aim was to see to what extent increasing a one-million word corpus to 10 million and again to a 100-million word corpus would enhance the quality of the lemmalist in terms of the most basic words of Afrikaans.

In the absence of a benchmark for basic words such as the three-starred words for English, an alternative approach and criterion for comparison had to be found. This was done through comparison of top frequencies in the 1m MED 24 with those in the 10m MED 24 with 100m MED 24 in order to determine internal stability in terms of top frequencies, or formulated differently, to what extent the top frequencies differ when a corpus is enlarged from one to 10 to 100 million words. The ideal situation would be if the top frequencies were identical as schematically illustrated by the single centre dot in figure 2a. Figure 2b represents a situation where there is great overlap in terms of this top frequency core and figure 2c a possible situation where the top frequencies do not overlap.

a. Identical	b. Substantial overlap	c. No overlap
		

Figure 2: Possible scenarios of overlap in top frequencies

Consider table 3 where the top-ranking 100 words in terms of frequency in the 100n MED 24 are compared to the 1m MED 24 in columns 2 and 3. Columns 4 and 5 indicate the difference in ranks and the extent of the deviation respectively.

	100m	1m	Comp.	Diff.
DIE	1	2	-1	1
#	2	1	1	1
VAN	3	3	0	0
HET	4	4	0	0
IN	5	5	0	0
EN	6	7	-1	1
N	7	6	1	1
IS	8	8	0	0
NIE	9	9	0	0
WAT	10	10	0	0
TE	11	11	0	0
OP	12	14	-2	2
OM	13	12	1	1
MET	14	15	-1	1
SE	15	13	2	2
WORD	16	17	-1	1
SY	17	18	-1	1
VIR	18	16	2	2
HY	19	21	-2	2
DIT	20	19	1	1
DAT	21	20	1	1
SAL	22	24	-2	2
AS	23	23	0	0
AAN	24	22	2	2
WAS	25	25	0	0
MNR	26	50	-24	24
KAN	27	26	1	1
BY	28	27	1	1
DEUR	29	31	-2	2
DIÉ	30	30	0	0
OOR	31	28	3	3
OOK	32	33	-1	1
GESÊ	33	34	-1	1
HULLE	34	29	5	5
TOT	35	32	3	3
SUID	36	38	-2	2
NA	37	36	1	1
MAAR	38	37	1	1
OF	39	35	4	4
DAAR	40	41	-1	1
ONS	41	39	2	2
MOET	42	43	-1	1
JAAR	43	40	3	3
EK	44	44	0	0
HUL	45	42	3	3
TEEN	46	47	-1	1
AFRIKA	47	46	1	1
GAAN	48	45	3	3
UIT	49	52	-3	3
MEER	50	48	2	2
SÊ	51	49	2	2
NET	52	51	1	1
TWEE	53	53	0	0
NOG	54	54	0	0
TOE	55	56	-1	1
WEES	56	61	-5	5
GISTER	57	57	0	0
EEN	58	55	3	3
HAAR	59	58	1	1
MENSE	60	63	-3	3
HOM	61	59	2	2
ANDER	62	65	-3	3
BAIE	63	64	-1	1
NUWE	64	67	-3	3
SOOS	65	60	5	5
EERSTE	66	68	-2	2
AL	67	62	5	5
NOU	68	72	-4	4
ONDER	69	69	0	0
GROOT	70	70	0	0
VOLGENS	71	71	0	0
NÁ	72	75	-3	3
VERLEDE	73	66	7	7
BEGIN	74	73	1	1
DRIE	75	88	-13	13
MY	76	78	-2	2
WIL	77	82	-5	5
MAAK	78	79	-1	1
SO	79	80	-1	1
EGTER	80	92	-12	12
WEER	81	85	-4	4
SOWAT	82	81	1	1
AFRIKAANSE	83	90	-7	7
VOOR	84	94	-10	10
NADAT	85	74	11	11
REEDS	86	84	2	2
TUSSEN	87	83	4	4
OMDAT	88	86	2	2
LAAT	89	89	0	0
WAAR	90	76	14	14
MILJOEN	91	87	4	4
DE	92	105	-13	13
THE	93	77	16	16
GEEN	94	102	-8	8
PRETORIA	95	106	-11	11
KRY	96	96	0	0
KOM	97	97	0	0
VANJAAR	98	120	-22	22
LAND	99	100	-1	1
DOEN	100	91	-1	1
				Average 3.1 positions different

Table 3: Top 100 ranks in 100m MED 24 versus 1m MED 24

From this table the stability in terms of the top 100 frequencies in the one million corpus versus the 100 million corpus is illustrated. Only 4 items, e.g. 92. *de*, 94. *geen*, 95. *Pretoria* and 98. *vanjaar* in the top 100 ranks of the 100 million corpus do not appear in the top 100 ranks of the one million corpus. Furthermore the actual difference in the rank numbers is very small. So, for example, are the rank numbers for rank 3, i.e. *van*, 4 *het*, 5 *in*, 8 *is*, 9 *nie* and 10 *wat* identical in both corpora. For the top 100 ranks the average variation in rank positions is only 3.1%. For the compilation of a dictionary with approximately 5,000 lemmas in mind, a random cut-off point of the top ranks at approximately 7,700 ranks were made in all three corpora. The aim is to determine which words likely to be looked for by the target user will be missed if only a one million corpus was available instead of a 10 million corpus and only a one million corpus versus a 100 million corpus. 7,737 words occur in the one million Afrikaans corpus with a frequency of 11 and more. Compared with the closest match in terms of frequency, 7,734 words occur in the 10 million corpus with a fre-

quency of 100 and more and 7,733 in the 100 million corpus with a frequency of 1081 and more. The overlap between these selected sections of the 1m MED 24 corpus' frequency list and the 10m MED 24 corpus is 6,449, i.e. 83.4%. The overlap between these selected sections of the 1m MED 24 and the 100m MED 24 is 5,991, i.e. 77.5%. This means that 1,742 words, i.e. 22.5% of the selected top section of the 100 million corpus would not have been available for consideration if the lexicographer only had the one million corpus available and 1,285 words or 16,6% if a 10 million corpus was available.

1m MED 24	**10m MED 24**	**100m MED 24**
Top 7,737 ranks considered Frequency of 11 and more	Top 7,734 ranks considered Frequency of 100 and more	Top 7,733 ranks considered Frequency of 1081 and more
Overlap 1m MED 24 versus 10m MED 24: 6,449 = 83,4%		
Overlap 1m MED 24 versus 100m MED 24Million: 5,991 = 77,5%		

Table 4: Comparison of top frequencies in the 1m MED 24, 10m MED 24 and 100m MED 24

The question is how significant this presumed 22.5% "loss" is for the compilation of the lemmalist. Among the words occurring with a high frequency are *Kersfees* 'Christmas', *koningin* 'queen', *toesig* 'supervision', *eksamen* 'exam', *koor* 'choir', *volk* 'nation', *aardbewing* 'earthquake', *skandaal* 'scandal', *digter* 'poet', *opskrif* 'heading', *strook* 'strip', *tjek* 'cheque' and *gogga* 'bug'. The Afrikaans lexicographer would probably regard these words as likely to be looked for and that they deserve a place in the dictionary.

For Sepedi the same procedure was followed in order to determine to what extent increasing a one-million word Sepedi corpus to a 10-million word corpus would enhance the quality of the lemmalist, i.e. to see which words likely to be looked for by the target user will be missed if only the 1m PSC was available instead of the 10m PSC. Consequently, the top 7,646 ranks occurring 8 times or more in the 1m PSC were compared to the top 7622 ranks occurring 62 times or more in the 10m PSC. The overlap was 5,553 words, i.e. 72.8%. This means that 2,069 high frequency words in 10m PSC were missed by the 1m PSC.

1m PSC	**10m PSC**
Top 7,646	Top 7,622
With frequency 8 times or more	With frequency 62 times or more
Overlap 5,553 words = 72.8%	

Table 5: Comparison of the top frequencies in 1m PSC and 10m PSC

As for Afrikaans, words occurring with high frequency in 10m PSC but not in the top 7,646 of 1m PSC were considered. These words include *bjalobjalo* 'et cetera', *diteng* 'contents', *seyalemoya* 'radio', *metara* 'metre', *semolao* 'legal', *kamano* 'relationship', *Bathobaso* 'Black people' and *komiti* 'committee'. Once again it is likely that the Sepedi lexicographer would regard them as common words likely to be looked for and that they should be included in the dictionary.

Microstructure

On the microstructural level the evaluation focused on the value of information drawn from limited corpora in terms of meaning, sense distinction, examples of usage, collocations and proverbs/idioms.

Consider as a first example the randomly selected adjective *great* in *Sketch Engine* in figure 3.

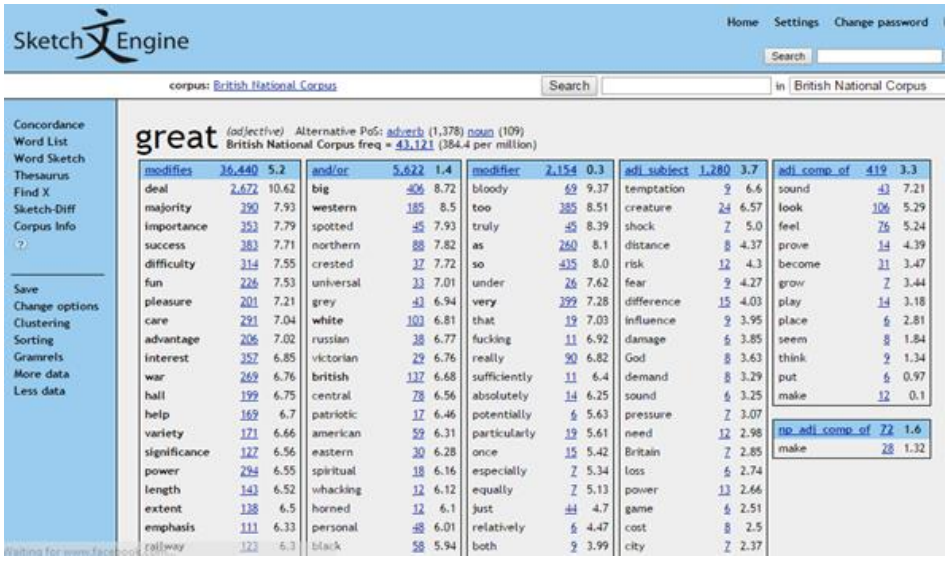

Figure 3: Collocations: *great* in Sketch Engine

The top 20 combinations of *great* + a noun in column 1 were then compared to the collocations for *great* given in MED, the 1m PEIC and the 10m PEIC as given in table 6. There were in total 1,709 occurrences of *great* in the 1m PEIC and 15,887 in the 10m PEIC.

	GREAT …	Sketch Engine	MED	1mPEIC (1,709)	10m PEIC (15,887)
1	great deal	✓	✓	22	58
2	great majority	✓	✓	12	51
3	great importance	✓		13	72
4	great success	✓		5	25
5	great difficulty	✓	✓	8	70
6	great fun	✓		0	12
7	great pleasure	✓	✓	9	51
8	great care	✓		1	52
9	great advantage	✓	✓	10	53
10	great interest	✓		2	33
11	great war	✓		3	35
12	great hall	✓	✓	4	56
13	great help	✓		0	8
14	great variety	✓		3	33
15	great significance	✓		1	2
16	great power	✓		3	67
17	great length	✓		3	20
18	great extent	✓		10	36
19	great emphasis	✓		0	3
20	great railway	✓		0	0

Table 6: Sketch Engine's *great* as modifier vs. MED, 1m PEIC and 10m PEIC

From table 6 column 4 it is clear that MED accounts for six of the 20 collocations, i.e. 30%. The 1m PEIC has examples of 16 (80%) and the 10m PEIC of 19 (95%). 80% for the 1m PEIC is significant for such a small corpus but a corpus should provide more evidence to the English lexicographer for common combinations such as *great fun*, *great care*, *great help* and *great significance*, etc. which are under-represented or missing in the 1m PEIC.

As a second example the senses of the verb *count* were studied in the 1m PEIC and the 10m PEIC. The senses distinguished in MED given in table 7 were used as a benchmark. As in the case of the frequency lists, it is not feasible for the lexicographer to read through thousands of concordance lines generated for a specific keyword in context – 100-300 lines could be regarded as a reasonable number to consider for detecting senses and to find typical collocations and authentic examples of use. The first deficiency encountered in the 1m PEIC was

an insufficient number of concordance lines. For *count* only 66 concordance lines were found in the 1m PEIC in contrast to 813 in the 10m PEIC. In the 10m PEIC a sufficient number of concordance lines were found for at least four out of five of the senses listed in table 7 but no or insufficient information for all senses, with the possible exception of the first sense *to calculate* in the 1m PEIC.

	Sense description	**1m PEIC**	**10m PEIC**
1	To calculate how many people or things there are in a group e.g. *all the votes have been counted*	3	27
2	Say numbers one after another in order e.g. *I can count up to ten in German*	1	5
3	To include someone or something in a calculation e.g. *sick pay is counted as income*		7
4	To think of someone or something as a particular thing e.g. *that counts as a lie*	1	11
5	To be important, or to have influence e.g. *what really counts is …*		1

Table 7: Verbal senses of *count* in MED compared to their occurrence in 1m PEIC and 10m PEIC

As a third example, consider three randomly selected Sepedi idioms in table 8: *monna ke nku (o llela) teng* 'a man is a sheep (he cries inside)', *bana ba tau (ga re jane)* 'children of a lion (we do not eat each other)' and *go sepela ke go bona* 'to travel is to see (become experienced)'.

Idiom	**1m PSC**	**10m PSC**
Monna ke nku …	11	127
Bana ba tau …	9	25
Go sepela ke go bona …	4	35

Table 8: Occurrence of idioms in 1m PSC versus 10m PSC

From table 8 it is clear that although in a limited number, these idioms do occur in a one million corpus but the lexicographer is more likely to detect them in a bigger corpus such as the 10m PEIC.

As for finding authentic examples of use, a one-million corpus proved to be quite significant for commonly used words of the language and as such could go a long way in supplementing the lexicographer's intuition when compiling a relatively small dictionary. Consider, for example, the potential for good examples even for the limited number of collocations *great success*, *great*

care and *great interest* in table 6 that can be found in the concordance lines from the 1m PEIC given in table 9.

troops that day was about twelve miles. This I regarded as a	**great**	success, and it removed from my mind the most serious
of his making his escape, that the Southern troops had had	**great**	success all day. Johnston forwarded the dispatch to Ri
opportunities should present themselves which would insure	**great**	success. General Meade was left in command of the few
destroy the railroad between Petersburg and Richmond, but no	**great**	success attended these latter efforts. He made no grea
entry into politics, a career he followed ever after with	**great**	success, and in which he died enjoying the friendship,
uniform and in prescribed order. Orders were prepared with	**great**	care and evidently with the view that they should be a
back to his grandfather. On the other side, my father took a	**great**	interest in the subject, and in his researches, he fou
change his position. While at Cairo I had watched with very	**great**	interest the operations of the Army of the Potomac, lo

Table 9: Concordance lines for *great success*, *great care* and *great interest* in 1m PEIC

Lexicographic tools

As for the creation of lexicographic tools, the aim was to determine whether a relatively small corpus of one million words can be utilized to create useful tools such as rulers, block systems, indicators of spreading-across-sources, etc. So, for example, the aim was to see whether, in the absence of larger corpora, a one-million word corpus would be sufficient to build a sensible guide for the lexicographer for balancing alphabetical stretches in the dictionary or whether larger corpora would contribute substantially to the refinement of such tools. Prinsloo and De Schryver (2002) introduced the concept of a measurement instrument for the relative length of alphabetical stretches in dictionaries and referred to it as a *lexicographic ruler*. Such a ruler guides the compiler of a dictionary to appropriately balanced alphabetical stretches in terms of overall length and the number of lemmas treated, i.e. not to over/under treat a specific alphabetic stretch in relation to the other alphabetic stretches. They indicate how, for example, a compiler could enthusiastically treat the first few alphabetic categories exhaustively but 'gets tired' towards the end of the alphabet. Formulated differently, a lexicographic ruler tells the compiler when alphabetic stretch 'A' has been sufficiently treated, i.e. when it is time to move on to 'B'. So, for example, Prinsloo and De Schryver (2003: 110) give a schematic illustration of a ruler for Afrikaans in figure 4.

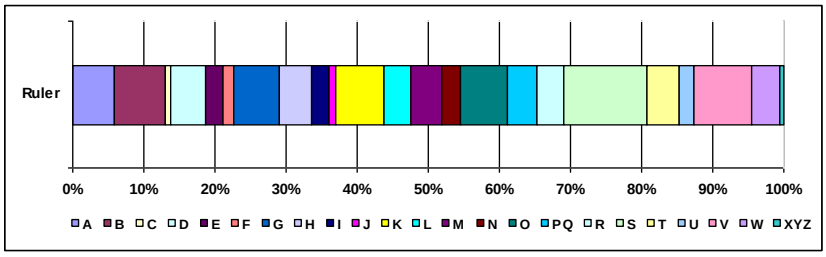

Figure 4: A lexicographic ruler for Afrikaans

This ruler indicates at a glance that e.g. B, K, O, S and V are relatively big stretches in Afrikaans whilst C, F, J, X, Y and Z are small. Figure 4 also gives a basic indication in terms of percentage of progress through the alphabetic stretches moving from A to Z. For example that M roughly represents the middle of the dictionary and that concluding S means reaching the 80% stage of compilation. They performed a formal breakdown of the ruler into percentages to guide dictionary compilation referred to as a block system. Consider, for example, the block system for Setswana in figure 5.

1	**ALAF**		21	**FELE**		41	**KOUS**		61	**MOTL**		81	**SELE**
2	**AROG**		22	**FOLO**		42	**LAEL**		62	**MPHE**		82	**SERA**
3	**BADI**		23	**GAGW**		43	**LEBO**		63	**NATE**		83	**SETO**
4	**BANN**		24	**GATS**		44	**LEKI**		64	**NGWA**		84	**SIMO**
5	**BATW**		25	**GOLO**		45	**LERI**		65	**NKUK**		85	**SUAS**
6	**BIRO**		26	**GWET**		46	**LETS**		66	**NTEM**		86	**TALE**
7	**BOGA**		27	**HUBE**		47	**LOKO**		67	**NTSH**		87	**THAA**
8	**BOLA**		28	**IJES**		48	**MAAD**		68	**NYOR**		88	**THIB**
9	**BONK**		29	**IKGO**		49	**MAHA**		69	**OOMA**		89	**THWE**
10	**BORU**		30	**INOL**		50	**MALE**		70	**PANT**		90	**TLAM**
11	**BOUT**		31	**IPUS**		51	**MARA**		71	**PHAK**		91	**TLHA**
12	**DAAM**		32	**ITIS**		52	**MATL**		72	**PHIM**		92	**TLHO**
13	**DIFA**		33	**ITSH**		53	**MEFA**		73	**PITL**		93	**TLWA**
14	**DIKG**		34	**JOKO**		54	**MESU**		74	**PUDU**		94	**TSAP**
15	**DINK**		35	**KANY**		55	**MMAL**		75	**RAMO**		95	**TSHE**
16	**DIRA**		36	**KERO**		56	**MMOL**		76	**RENG**		96	**TSHW**
17	**DITH**		37	**KGAR**		57	**MOFI**		77	**ROKG**		97	**TSUN**
18	**DITU**		38	**KGOM**		58	**MOKG**		78	**RURU**		98	**UBAU**
19	**EGEP**		39	**KHAN**		59	**MONG**		79	**SEBA**		99	**WABO**
20	**ETLH**		40	**KODU**		60	**MORW**		80	**SEHI**		100	**ZIMB**

Figure 5: A block system for Setswana

A useful practical application of a block system is to pace dictionary compilation in terms of time and resources. It suggests that the compiler should be at IN when 30% of time and resources have been spent, that MA roughly repre-

sents 50% of completion but that 15% of time and resources should be spent on M, and that SE is the 80% mark.

Rulers are calculated by determining the percentage of words in each alphabetic category from an alphabetic list of words culled from a corpus. This simply means how many words start with a, b, c, … z. The same data is used for calculating a block system but instead of the 26 letters of the alphabet, the list is broken down into 100 sections to each represent 1%.

The question here is whether a ruler compiled from a one-million word corpus could provide a reliable ruler when compared to a 10 million corpus. In table 10 the breakdown of words into alphabetical stretches of both the 1m PSC and the 10m PSC is given. Columns 3 and 5 reflect the percentage breakdown per alphabetical stretch in the 1m PSC versus the 10m PSC and the difference between these percentages is given in column 6.

	1m PSC	% 1m PSC	10m PSC	% 10m PSC	Difference
A	1164	2.13	6521	2.55	-0.41
B	5045	9.25	23123	9.02	0.22
C	98	0.18	1853	0.72	-0.54
D	3486	6.39	17241	6.73	-0.34
E	753	1.38	4271	1.67	-0.29
F	1475	2.70	5703	2.23	0.48
G	1945	3.57	8697	3.39	0.17
H	2275	4.17	9147	3.57	0.60
I	2475	4.54	10668	4.16	0.37
J	206	0.38	1311	0.51	-0.13
K	3519	6.45	16433	6.41	0.04
L	3657	6.70	15466	6.04	0.67
M	9005	16.51	40687	15.88	0.63
N	3357	6.15	14010	5.47	0.69
O	715	1.31	4032	1.57	-0.26
P	2484	4.55	12123	4.73	-0.18
Q	0	0.00	386	0.15	-0.15
R	1581	2.90	9663	3.77	-0.87
S	4629	8.49	22433	8.76	-0.27
T	5872	10.77	26155	10.21	0.56
U	270	0.50	1521	0.59	-0.10
V	68	0.12	1601	0.62	-0.50
W	247	0.45	1742	0.68	-0.23
X	45	0.08	324	0.13	-0.04
Y	154	0.28	901	0.35	-0.07
Z	20	0.04	215	0.08	-0.05

Table 10: Alphabetical stretches in 1m PSC compared to 10m PSC

The final column indicates that the difference between the rulers is very small with the difference in all stretches less than 1%. The similarity is visually illustrated in figure 6 where the two lines of the graph are very close to each other.

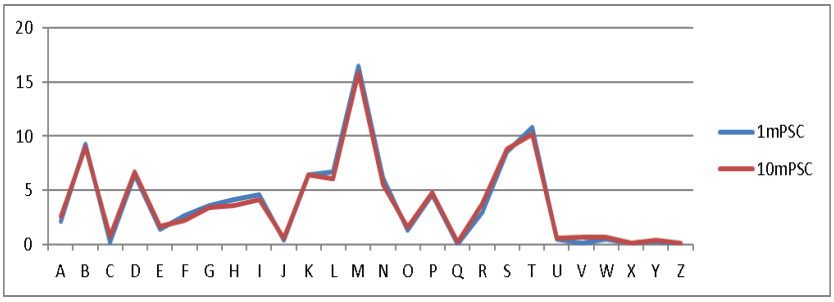

Figure 6: A ruler graph for 1m PSC versus 10m PSC

The same similarity is observed in the breakdown in the block systems calculated from the 1m PSC versus the 10m PSC in table 11.

%	1m PSC	10m PSC
1	ALO	AKO
2	ATH	ARE
3	BAH	BAF
4	BAR	BAP
5	BEL	BEF
6	BLO	BJE
7	**BOH**	**BOH**
8	BOL	BOL
9	BON	BOM
10	BOS	BOR
11	BUA	BOT
12	DIB	COM
13	DIK	DIB
14	DIM	DIK
15	DIP	DIN
16	DIT	DIP
17	**DIT**	**DIT**
18	EDI	DIU
19	ERI	DUT
20	FEE	ENK
21	FIH	FAR
22	FUL	FIH
23	GAM	GAB
24	GIL	GAM
25	GON	GOB
26	HLA	GRA
27	HLA	HLA
28	HLO	HLA
29	HOM	HOL
30	**IDI**	**IDI**
31	IKI	IKG
32	IPH	IPA
33	ITH	ITH
34	ITS	IWE
35	KAM	KAN
36	KGA	KGA
37	KGE	KGO
38	KGO	KGO
39	KHU	KIL
40	KON	KOT
41	KWE	LAB
42	LEA	LEB
43	LEF	LEH
44	LEK	LEN
45	LEP	LET
46	LET	LLA
47	LOG	MAA
48	MAB	MAF
49	MAG	MAJ
50	MAI	MAM
51	MAL	MAR
52	MAR	MAT
53	MAS	MED
54	MAZ	MEP
55	MEL	MMA
56	MET	MMO
57	MME	MOG
58	MOD	MOK
59	MOH	MOM
60	MOL	MOR
61	MON	MOT
62	MOS	MPH
63	MOT	NAG
64	MPO	NGW
65	NEE	NKG
66	NIK	NTA
67	NKU	NTS
68	NTE	NYA
69	NTS	OLO
70	NYS	PAF
71	OKS	PET
72	PAL	PHE
73	PHA	PHU
74	PHE	POT
75	PHU	RAG
76	PŠH	REI
77	RAP	ROB
78	RIP	ROT
79	RUR	ŠAR
80	SEB	SEE
81	SEG	SEJ
82	SEK	SEL
83	SEN	SER
84	SER	SET
85	SET	SIS
86	SOB	SOU
87	ŠUT	SWA
88	TAL	TAU
89	**THA**	**THA**
90	THE	THI
91	THU	TIA
92	TIT	TLE
93	TLH	TLW
94	TOM	TSE
95	TSE	TŠH
96	TŠH	TSI
97	TŠI	TSW
98	TŠW	UTI
99	UTS	WEB
100	**ZUL**	**ZUL**

Table 11: Sepedi block systems: 1m PSC versus 10m PSC

So, for example, both block systems indicate that the compiler should be at the sub-stretch ID after 30% of the available time and resources for the project, at MA after 50%, SE after 80%, etc. All of the other comparative blocks are alphabetically very close to each other.

Conclusion

In this article it has been argued that raw corpora built only from written data, although not reflecting an ideal situation, can substantially assist the lexicographer in the compilation of especially small bilingual and monolingual dictionaries.

On the macrostructural level a corpus of one million words is useful to pinpoint the most commonly used words in the language and would be a useful tool for the lexicographer tasked with the compilation of a relatively small dictionary of approximately 5,000 lemmas. Additional common words will however have to be found. Consider in this regard high-ranking words in the 100m MED 24 mentioned which were not found in the 1m MED 24. The lexicographer will have to find such words through other means, e.g. introspection, field work and reading and marking. If a one million corpus is extended to 10 million words the offering of commonly used words in the top frequency ranks becomes more reliable and represents a gradual enhancement. If the corpus is further extended to a 100 million words, the frequently used words provide a reliable account of the commonly used words in the language and little additional collection is required from the lexicographer for a small dictionary.

As far as microstructural elements are concerned, it is clear that a one million corpus is useful in determining the basic senses of a word as well as typical examples of usage of these basic senses. Such a corpus would typically include a limited number of idioms. Increasing the corpus to 10 million words gradually improves the situation in the sense that more senses are detected, more idioms can be found and more evidence on the use and meaning of such words and idioms is available.

As for lexicographic tools, the results clearly indicate that reliable lexicographic rulers and block systems could be compiled from a corpus as small as one million words. In this case enlarging the corpus to 10 million did not substantially enhance the quality/accuracy of the tool.

In conclusion it could be recommended that the lexicographer should carefully analyse the situation for each specific language. If no written sources are available (s)he should attempt to compile, say, a one-million token corpus of the spoken language. If a limited number of written sources are available, (s)he should try to compile a 10 million corpus and if sources are available in abundance, especially in electronic format, a 100 million corpus will be extremely valuable.

Acknowledgement

This research is (a) conducted within the SeLA project (Scientific e-Lexicography for Africa), supported by a grant from the German Ministry for Education and Research, administered by the DAAD and (b) supported in part by the National Research Foundation of South Africa (Grant specific unique reference number (UID) 85763). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are those of the author, and that the NRF accepts no liability whatsoever in this regard.

References

AntConc: http://www.laurenceanthony.net/software/antconc/ (Consulted 25 June 2015).

Brown Corpus of Standard American English: http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html (Consulted 25 June 2015).

COBUILD: Sinclair, J. (Ed.). 1995. *Collins COBUILD English Dictionary*. Second Edition. London: HarperCollins.

Dante: http://www.webdante.com/ (Consulted 25 June 2015).

Google Books: http://googlebooks.byu.edu/ (Consulted 25 June 2015).

Interactive Language Toolbox: https://ilt.kuleuven.be/inlato/ (Consulted 25 June 2015).

MED: Rundell, M. 2007. *Macmillan English Dictionary for Advanced Learners*. Second Edition 2007. Oxford: Macmillan.

MEDIA 24: Subsection of the archive for the newspaper *Beeld* http://argief.beeld.com/cgi-bin/beeld.cgi (Extract made available by Pharos/Media 24).

PSC: Pretoria Sepedi Corpus compiled at the University of Pretoria.

PEIC: Gauton, Rachélle: The University of Pretoria English Internet Corpus.

Prinsloo, D.J. and G.-M. de Schryver. 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, A. and 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, University of Copenhagen.

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 [Effective Progress with the *Woordeboek van die Afrikaanse Taal* as Measured in Terms of a Multidimensional Ruler]. Botha, W. (Ed.). 2003. *'n Man wat beur. Huldigingsbundel vir Dirk van Schalkwyk*: 106-126. Stellenbosch: Buro van die WAT.

Sketch Engine: http://www.sketchengine.co.uk/ (Consulted 10 January 2015).

WordSmith Tools: http://www.lexically.net/wordsmith/index.html (Consulted 10 January 2015).