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.

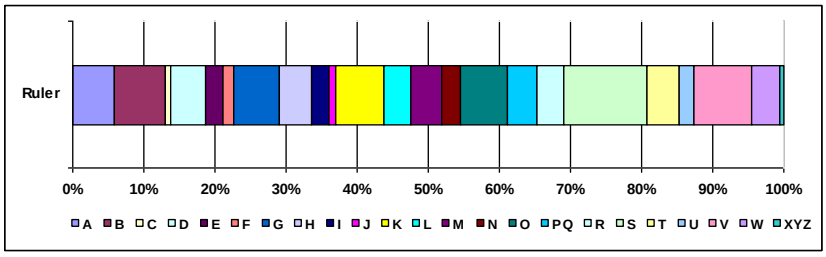

Figure 4: A lexicographic ruler for Afrikaans

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.

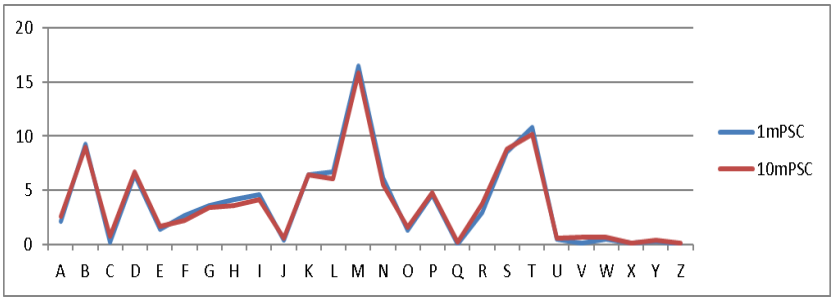

Figure 6: A ruler graph for 1m PSC versus 10m PSC

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).