Corpus-based lexicography for under-resourced languages – maximizing the limited corpus

D.J. Prinsloo

University of Pretoria, South Africa

danie.prinsloo@up.ac.za

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Under-resourced languages

- “Big corpora” is a relative term
- For lesser resourced languages 1m or 10m can be a “big corpus”
- Relatively small, unbalanced, un-annotated raw corpora
- Maximally utilized for lexicographic purposes
- To obtain similar results in the absence of large corpora versus corpus enlargement, corpus cleaning. Time best invested?
- African languages, such as the Bantu languages and Afrikaans, as a case in point

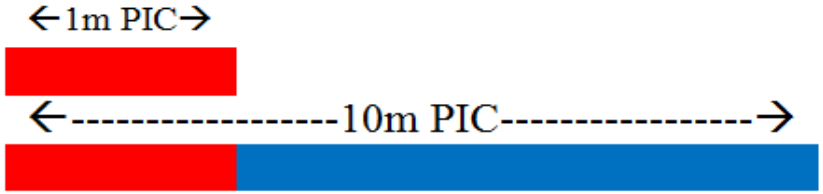
Aims

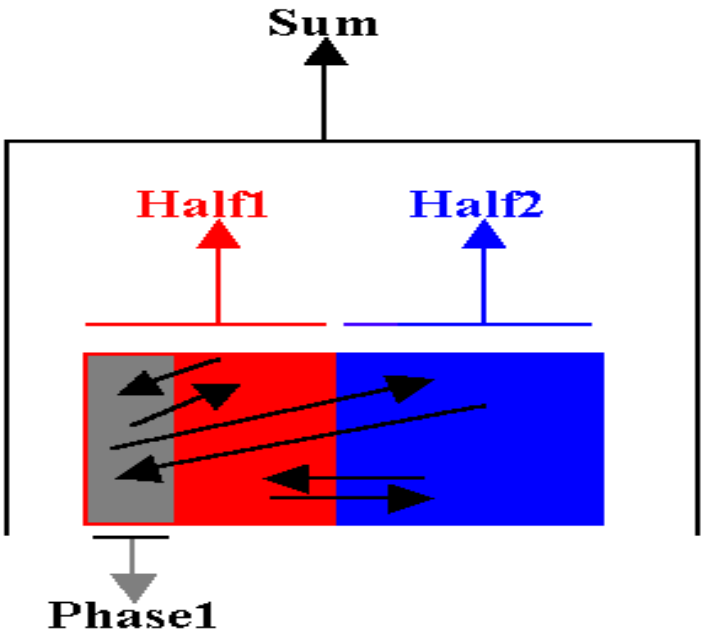
- Determine to what extent enlarging a corpus from e.g. 1 to 10 million, and from 10 million to 100 million tokens enhance its potential for
  - Macrostructure compilation,
  - Information on the most important microstructural aspects and
  - Creation of lexicographic tools.

Progress indicator:

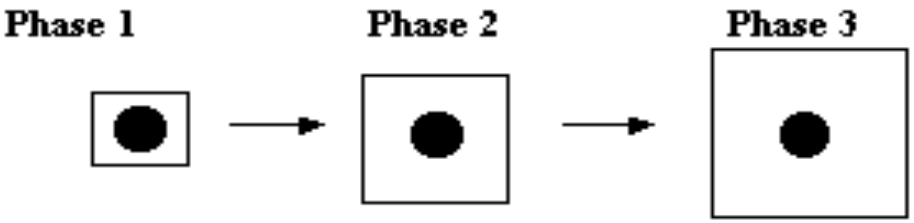
- **Macrostructure compilation**
- Microstructural aspects
- Lexicographic tools

Stages of comparison: English, Afrikaans and Sepedi

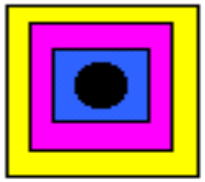




COMPARING DIFFERENT STAGES OF THE SAME CORPUS IN RESPECT OF THE CORE



Internal stability in terms of core vocabulary: Ideal



Internal stability in terms of core vocabulary: Extreme: no overlap in core vocabulary

Result of internal stability test for Sepedi: (90 % perfect)

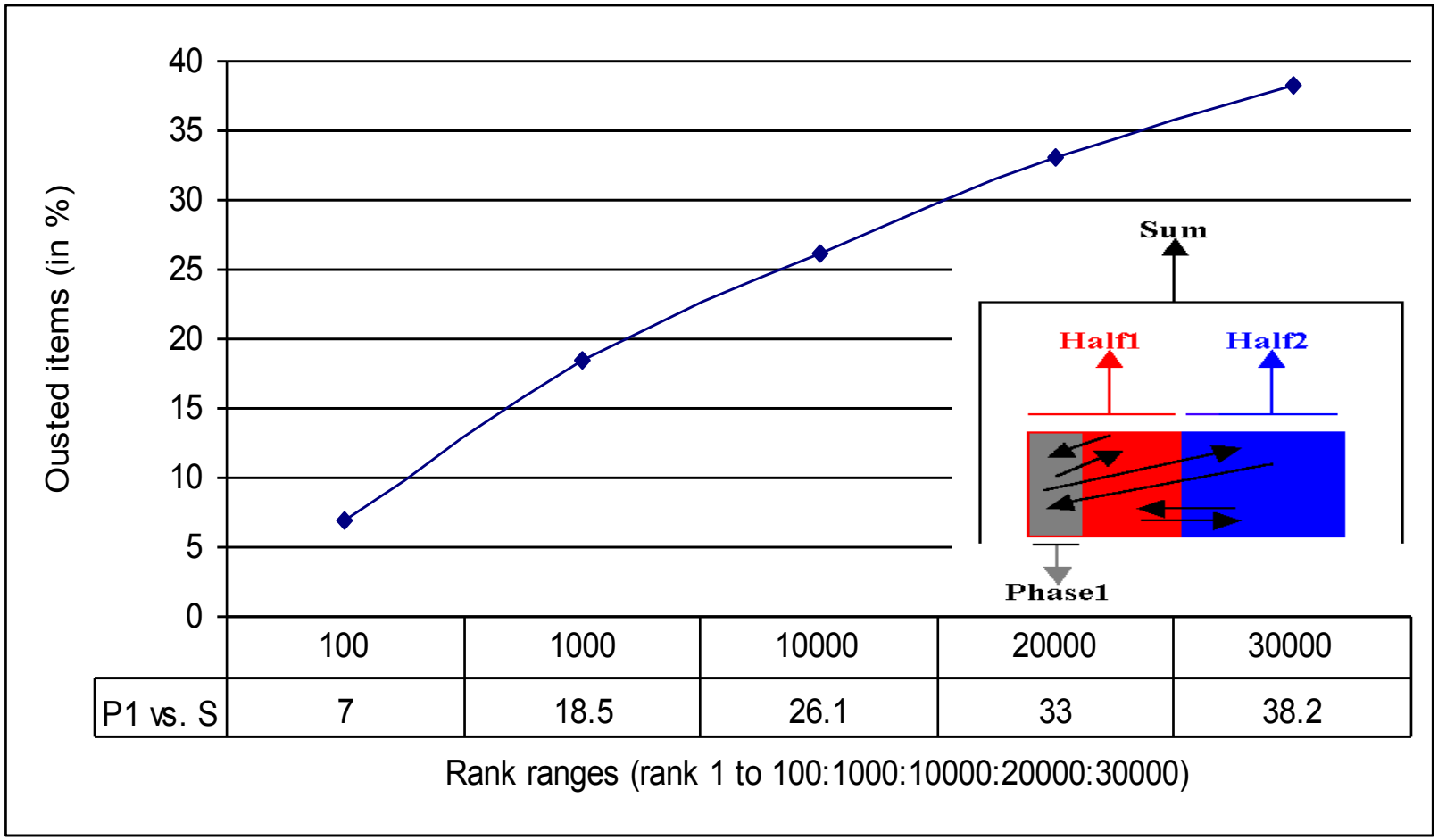
RANKS OF THE TOP 100 ITEMS WHEN COMPARING PSC-PHASE1 WITH PSC-SUM

Item	PSC-Phase1	PSC-Sum
bao	76	73
selo	77	97
kwa	78	76
mokgwa	79	89
tšona	80	64
yoo	**81**	**106**
morago	82	87
banna	83	91
woo	84	88
tle	85	100
gape	86	85
gago	87	48
thoma	88	96
no	89	84
ao	90	92
aowa	**91**	**114**
nyaka	**92**	**110**
bjang	**93**	**108**
bangwe	**94**	**102**

Summary of frequency band values in Cobuild

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

OUSTED ITEMS WHEN COMPARING RANK RANGES OF PSC-PHASE1 WITH PSC-SUM

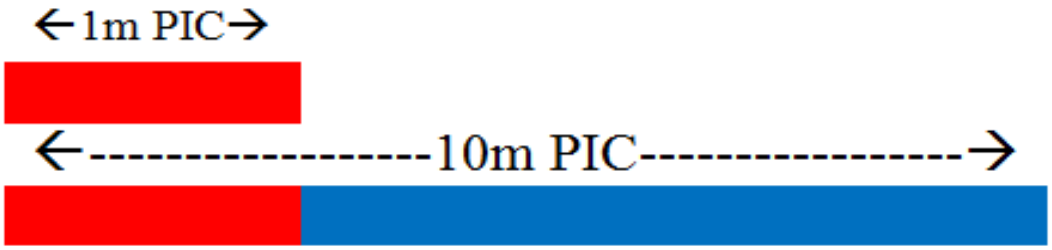


	100	1000	10000	20000	30000
P1 vs. S	7	18.5	26.1	33	38.2

Wordlist comparisons

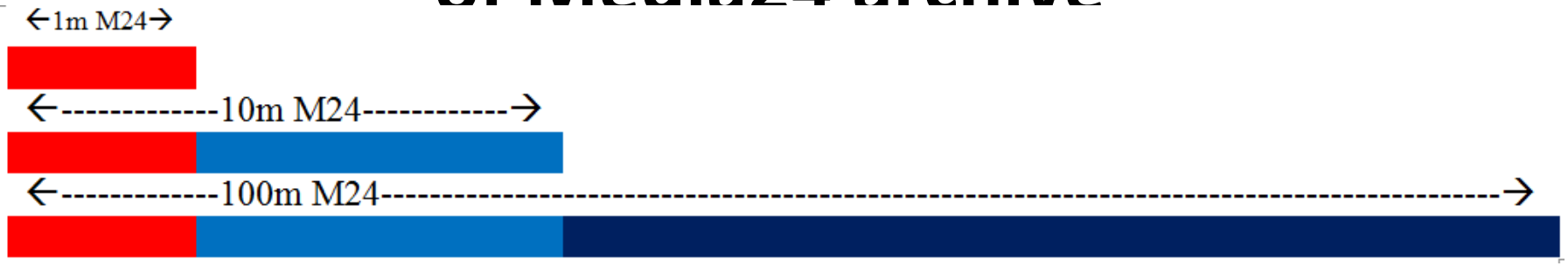
- lemmalists compiled from corpora consisting of tokens:
  - –1 million
  - –10 million
  - –100 million

English: MED versus 1m versus 10m sections of the PIC



MED	1 Million (PIC) (1069,429 tokens)	10 Million PIC (12,398,893 tokens)
2,275(*** starred words) "a word with three stars is one of the most basic words in English" (MED:x)	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 PIC (overlap with MED ***): 2,095 = 91.1% (Lexicographer considers freq. >513) (2,277 words to consider)

Afrikaans: 1m vs. 10m vs. 100m sections of Media24 archive



1 Million (1,011,970 tokens)	10 Million (10,271,880 tokens)	100 Million (119,040,000 tokens)
7,737 Frequency of 11 and more	7,734 Frequency of 100 and more	7,733 Frequency of 1081 and more
Overlap 1 Million versus 10 Million corpora: 6,449 = 83,4%		
Overlap 1 Million versus 1000 Million Corpora 5,991 = 77,5%		

Counting the losses …

Kerfees ‘Christmas’
koningin ‘queen’
eksamen ‘exam’
volk ‘nation’
aardbewing ‘earthquake’
digter ‘poet’
strook ‘strip’
gogga ‘insect’

rubriek ‘report’
toesig ‘caretaking’
koor ‘choir’
aanbreek ‘arrive’
skandaal ‘scandal’
opskrif ‘heading’
tjek ‘cheque’

Sepedi: 1m vs. 10m

1 M PSCl (1,190,583 tokens)	10 PSC(10,242,780 tokens)	100m 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%		

Subjective evaluation of the 2,069 high frequency words in 10M PSC

missed by the 1 M PSC

Progress indicator:

- Macrostructure compilation
- **Microstructural aspects**
- Lexicographic tools

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

Collocations: “GREAT” in Sketch Engine

modifies	36,440	5.2
deal	2,672	10.62
majority	390	7.93
importance	353	7.79
success	383	7.71
difficulty	314	7.55
fun	226	7.53
pleasure	201	7.21
care	291	7.04
advantage	206	7.02
interest	357	6.85
war	269	6.76
hall	199	6.75
help	169	6.7
variety	171	6.66
significance	127	6.56
power	294	6.55
length	143	6.52
extent	138	6.5
emphasis	111	6.33
railway	123	6.3

and/or	5,622	1.4
big	406	8.72
western	185	8.5
spotted	45	7.93
northern	88	7.82
crested	37	7.72
universal	33	7.01
grey	43	6.94
white	103	6.81
russian	38	6.77
victorian	29	6.76
british	137	6.68
central	78	6.56
patriotic	17	6.46
american	59	6.31
eastern	30	6.28
spiritual	18	6.16
whacking	12	6.12
horned	12	6.1
personal	48	6.01
black	58	5.94

modifier	2,154	0.3
bloody	69	9.37
too	385	8.51
truly	45	8.39
as	260	8.1
so	435	8.0
under	26	7.62
very	399	7.28
that	19	7.03
fucking	11	6.92
really	90	6.82
sufficiently	11	6.4
absolutely	14	6.25
potentially	6	5.63
particularly	19	5.61
once	15	5.42
especially	7	5.34
equally	7	5.13
just	44	4.7
relatively	6	4.47
both	9	3.99

adj_subject	1,280	3.7
temptation	9	6.6
creature	24	6.57
shock	7	5.0
distance	8	4.37
risk	12	4.3
fear	9	4.27
difference	15	4.03
influence	9	3.95
damage	6	3.85
God	8	3.63
demand	8	3.29
sound	6	3.25
pressure	7	3.07
need	12	2.98
Britain	7	2.85
loss	6	2.74
power	13	2.66
game	6	2.51
cost	8	2.5
city	7	2.37

adj_comp_of	419	3.3
sound	43	7.21
look	106	5.29
feel	76	5.24
prove	14	4.39
become	31	3.47
grow	7	3.44
play	14	3.18
place	6	2.81
seem	8	1.84
think	9	1.34
put	6	0.97
make	12	0.1

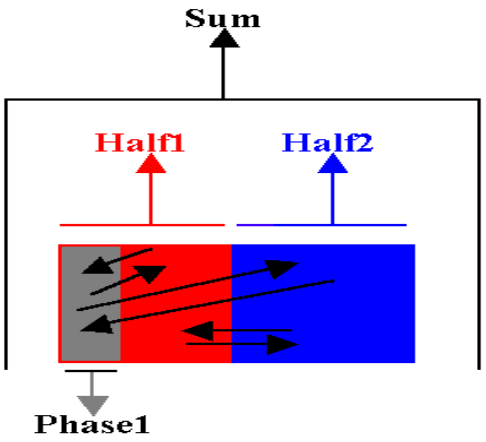
np_adj_comp_of	72	1.6
make	28	1.32

Sketch Engine's GREAT as modifier vs. MED/1mPIC and 12mPIC

GREAT …	MED	1M PIC	12M PIC	Sketch Engine
great deal	yes	yes	yes	yes
great majority	yes	yes	yes	yes
great importance	no	no	yes	yes
great success	no	no	yes	yes
great difficulty	yes	no	yes	yes
great fun	no	no	yes	yes
great pleasure	yes	yes	yes	yes
great care	no	yes	yes	yes
great advantage	yes	no	yes	yes
great interest	no	no	yes	yes
great war	no	yes	yes	yes
great hall	yes	yes	yes	yes
great help	no	no	no	yes
great variety	no	no	yes	yes
great significance	no	yes	yes	yes
great power	no	yes	yes	yes
great length	no	no	yes	yes
great extent	no	no	yes	yes
great emphasis	no	no	yes	yes
great railway	no	no	yes	yes

THE IDIOM *MONNA KE NKU, O LLELA TENG* IN THE GROWING ORGANIC PSC

Context …	Key Word In Context (KWIC)	… Context	Phase1	Half1	Half2	Sum
mo dutšego mogolong. O no re	**monna ke nku, o llela teng**	. A di rumilego bjalo, le tsona tša	✓	✓		✓
! Monna ga se a swanela go lla!	**monna ke nku o llela teng**	. Ge o ka bona monna a ediša dik	✓	✓		✓
ape go sekhumola. "Tiiša thaka,	**monna ke nku o llela teng**	." "... ba boletše ... ba boletše, mot	✓	✓		✓
sa mmone. Ee! Baswana ba re	**monna ke nku, o llela teng**	, fela ge e le Thogorogo yena o il		✓		✓
ba iteile lešepa ka thoka ge ba re	**monna ke nku, o llela teng**	, gomme a bona ba opile kgomo l		✓		✓
a. Fela ka gore bagologolo ba re	**monna ke nku, o llela teng**	, o ile a no ikgata pelo a tšwela pe		✓		✓
na ba be ba no šita kgang, ba re	**monna ke nku o llela teng**	. Ba be ba fetogile difahlegong, b		✓		✓
jo bo rego ke metlae ge go thwe	**monna ke nku o llela teng**	. Ba ile go felela ka mola mphom		✓		✓
ta. Ee, ke therešo. Sesotho se re	**monna ke nku, o llela teng**	. Fela le ge se realo, leo morwa' H		✓		✓
megokgo, motho a lebetše gore	**monna ke nku o llela teng**	. Ga se thaka ya mošemane go go		✓		✓
olo o ile a mo homotša ka go re:	**monna ke nku, o llela teng**	. Ke ge a be a lemogile gore ga se		✓		✓
ela gore ga se nnete ge go thwe	**monna ke nku o llela teng**	. Le go llela teng ga nnete go tleg		✓		✓
ela gore ga se nnete ge go thwe	**monna ke nku o llela teng**	. O ile a bokolela ka pelobohloko		✓		✓
ela bjang, goba ke gona ge ba re	**monna ke nku, o llela teng**	? Gape taba ke ngwana wa rena w		✓		✓
ba rego	**monna ke nku,**	o hwa natšo goba mosadi o fogoh		✓		✓
ema se,	**monna ke nku, o llela teng**	, monna le nku di llela teng			✓	✓
še gore	**monna ke nku o llela teng**	. Ka yeo nako ke ge madira ale a			✓	✓
la gore	**monna ke nku o llela teng**	. O be a sa itiriše ka gore le go m			✓	✓
ka gore	**monna ke nku.**	Aretse, ee, monna ke nku. Mošate			✓	✓
e ba re	**monna ke nku.**	Basadi bale ba bego ba le moo le			✓	✓
e ba re	**monna ke nku**	dihlong tša se mo je. O ratharathi			✓	✓
ma tše,	**monna ke nku, o llela teng**	le Phaga ga e ete, go eta nakedi,			✓	✓
e ba re	**monna ke nku,**	mafelelong a ba a ntša phefo gan			✓	✓
Matlala	**monna ke nku**	o latswa bohloko. E rile e tsena k			✓	✓



Senses: “COUNT” in 1m PIC vs. 10m PIC

1m PIC concordance lines for verbal senses:

- “to add”
- “to consider”

But no nominal senses

10m PIC concordance lines for verbal senses:

- “to add”
- “to consider”

and nominal sense of “a number”

Progress indicator:

- Macrostructure compilation
- Microstructural aspects
- **Lexicographic tools**

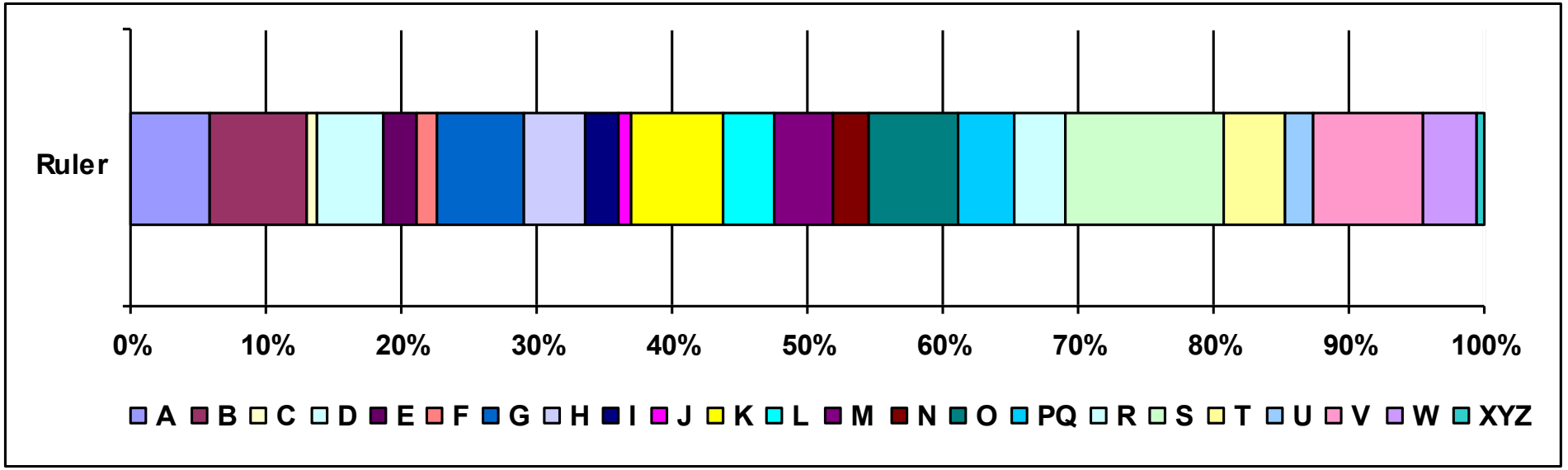
- a relatively small corpus of one million words can be utilized to create useful lexicographic tools:
  - Rulers
  - Block systems
  - Indicators of spreading-across-sources, etc.

Lexicographic rulers

- Balancing alphabetical stretches
  - Number of lemmas
  - Number of pages
  - Time
  - Remuneration of lexicographers

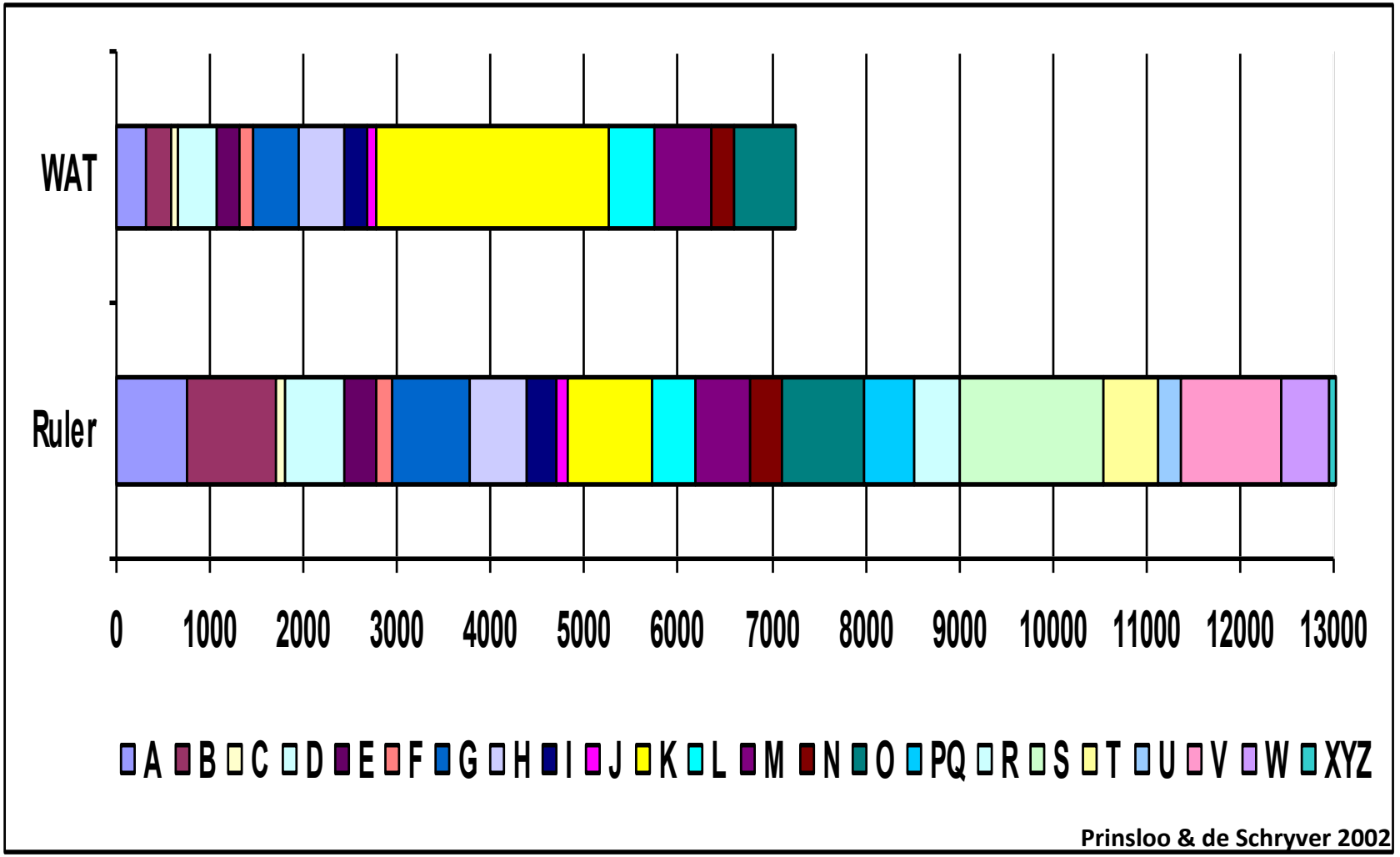
  How to do it:
    - Generate an alphabetical word list
    - Calculate the percentage of words per alphabetical stretch

Ruler for Afrikaans as a « ruler »



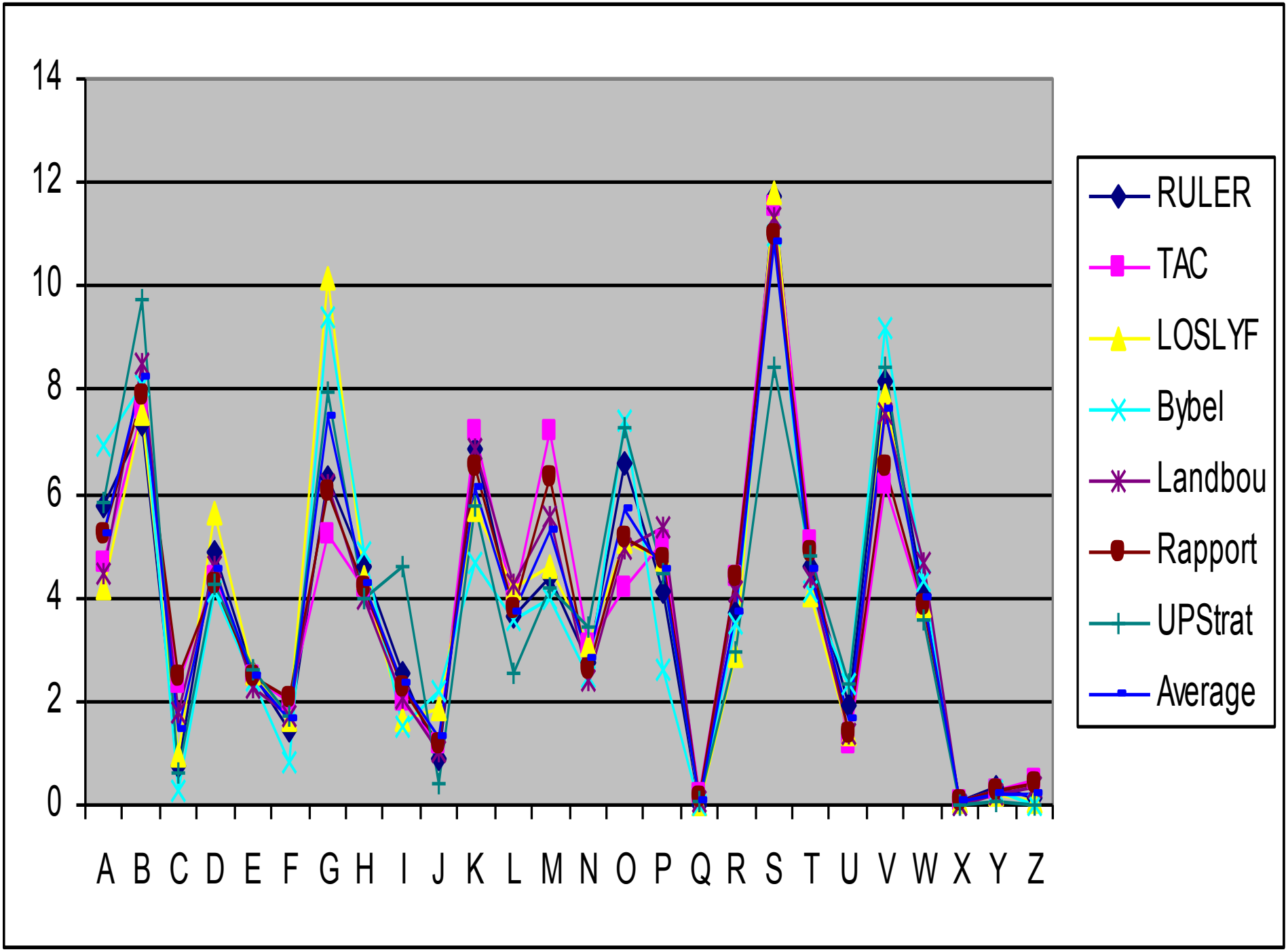
Prinsloo & de Schryver 2002

WAT *versus* the Ruler (# pages)



A Block System for Sesotho sa Leboa

%	Marker	%	Marker	%	Marker	%	Marker	%	Marker
1	ALAF	21	FAHL	41	KUKEL	61	MONO	81	SEET
2	AROG	22	FETL	42	LAMO	62	MOŠA	82	SEJA
3	BAFE	23	FOŠW	43	LEDI	63	MOTO	83	SEMA
4	BANK	24	GALA	44	LEKA	64	MPHO	84	SERE
5	BEAB	25	GAYA	45	LEPO	65	NASW	85	SETS
6	BITL	26	GOLE	46	LETŠ	66	NGWE	86	SITE
7	BOGE	27	HAHA	47	LOGW	67	NKOK	87	STEF
8	BOKO	28	HLAH	48	MABE	68	NTEB	88	SWEL
9	BOMM	29	HLOG	49	MAGA	69	NTSEB	89	TEKE
10	BOPU	30	HOSE	50	MAKA	70	NYAK	90	THATA
11	BOTL	31	IHLO	51	MAMO	71	OLEL	91	THOM
12	BUWA	32	ILWA	52	MARA	72	PANK	92	TIKR
13	DIAP	33	IPIT	53	MATH	73	PHAK	93	TLHA
14	DIIP	34	ITLH	54	MEAG	74	PHET	94	TONA
15	DIKU	35	JESU	55	MELO	75	PIPA	95	TSEN
16	DIPE	36	KATOL	56	MIDI	76	PŠHA	96	TŠHI
17	DITE	37	KGAN	57	MMASE	77	RANG	97	TSOL
18	DITO	38	KGOH	58	MOBO	78	RETA	98	TUME
19	DUDI	39	KGWA	59	MOHL	79	RRAG	99	WABO
20	EMAE	40	KLAS	60	MOKO	80	SATH	100	ZOUN


14
12
10
8
6
4
2
0
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
RULER
TAC
LOSLYF
Bybel
Landbou
Rapport
UPStrat
Average

Conclusion

- Raw and even un-annotated 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.
- **Macrostructure**: Stability in core vocabulary 1m vs. 10m vs. 100m
- **Microstructure**: most frequent senses, collocations, idioms/ proverbs detected
- **Lexicographic tools**: 1m corpus sufficient for basic tools
- **Priority on actual compilation** rather than expensive corpus enlargement and corpus cleaning

Thank you! / Dankie!/ Ke a leboga!

References

- *Brown Corpus of Standard American English*. http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/ corpora/list/private/brown/brown.html.
- Dante: http://www.webdante.com/
- Johansson, Stig, Geoffrey Leech and Helen Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/bergen Corpus of British English, for Use with Digital Computers.* Oslo: Department of English, University of Oslo.
- Sketch Engine: http://www.sketchengine.co.uk/
- WordSmith Tools: http://www.lexically.net/wordsmith/index.html