CHAPTER 8

Managing eleven parallel corpora and the extraction of data in all official South African languages

D.J. PRINSLOO & GILLES-MAURICE DE SCHRYVER
University of Pretoria & Ghent University

Introduction: a decade of African-language corpus building

Corpora have been built in the Department of African Languages of the University of Pretoria since the early 1990s. Work began on a corpus for Sesotho sa Leboa, the *Pretoria Sesotho sa Leboa Corpus* (PSC), which gradually grew from 156 000 running words or "tokens" in 1990 (see Prinsloo 1991) to 5.8 million words a decade later (see De Schryver & Prinsloo 2001a). Today, this corpus stands at 8.7 million words and is still growing. In addition, corpora had been built for all other official South African languages by the end of the 1990s, although the sizes of these corpora remained rather small compared to PSC (see De Schryver & Prinsloo 2000a). Around the turn of the millennium, however, a dedicated major effort brought about an important change. A joint project between the African-language departments of the University of Pretoria on the one hand, and Ghent University (Flanders, Belgium) on the other hand, resulted in relatively large corpora for *all* official South African languages, with sizes averaging several million tokens per language. In the process, quite a number of other African-language corpora were built as well, such as corpora for Cilubà, Kiswahili, Hausa, Somali and Lingala (see Van der Veken & De Schryver 2003).

The foundations for a South African discipline of corpus linguistics were laid in a series of articles published from 2000 onwards. In De Schryver and Prinsloo (2000a) the creation of such corpora was discussed in great detail and a wide range of already-existing applications for languages such as isiZulu, Setswana, etc., were reviewed in Prinsloo and De Schryver (2001a). Crucial corpus stability issues, applied to both Sesotho sa Leboa and Xitsonga, were expounded in Prinsloo and De Schryver (2001b). One year later it was shown how African-language Internet data

© Van Schaik Publishers

could already be used *for* corpus creation and how the Internet itself could be directly queried *as* a corpus (see De Schryver 2002). A set of African-language applications based on such Internet data was presented next in Van der Veken and De Schryver (2003) and includes, for instance, a spellchecker for isiXhosa.

It is thus clear that this is the golden age of corpus "building" for the (South) African languages, while corpus "applications" for these languages have become a *sine qua non* in the large field of Human Language Technologies (HLTs). In this article an overview is given of the major characteristics of the authors' South African language corpora, as well as of some major current and planned research. Tools and applications with regard to these corpora are also highlighted, and two groundbreaking corpus-based case studies are presented, the first on ways to semi-automatically translate words between all official South African languages, and the second, on multidimensional lexicographic Rulers for those languages.

Available corpora for the official South African languages

Basically, three sets of corpora have been built by the authors of the current article. They include the language for general-purpose (LGP) corpora, the language for special-purpose (LSP) corpora and the true "parallel" corpora consisting of translations of the same document in all eleven official South African languages. The various LSP corpora will not be discussed in any detail here. However, it may be noted at this point that, for instance, an original use of a Sesotho sa Leboa linguistics LSP corpus is described in Taljard and De Schryver (2002).

The main characteristics of the eleven South African LGP corpora are shown in Table 8.1.

Table 8.1 LGP Pretoria Corpora

LGP Corpus Name	Acronym	Tokens	Types
Pretoria isiNdebele Corpus	PNC	1 959 482	250 990
Pretoria siSwati Corpus	PSwC	4 442 666	293 156
Pretoria isiXhosa Corpus	PXhC	8 065 349	846 162
Pretoria isiZulu Corpus	PZC	5 783 634	674 380
Pretoria English Corpus	PEC	12 799 623	119 235
Pretoria Afrikaans Corpus	PAfC	11 602 276	373 497
Pretoria Xitsonga Corpus	PXiC	4 556 959	115 848
Pretoria Tshivenda Corpus	PTC	4 117 176	118 771
Pretoria Setswana Corpus	PSTC	6 130 557	157 274
Pretoria Sesotho sa Leboa Corpus	PSC	8 749 597	165 209
Pretoria Sesotho Corpus	PSSC	4 513 287	107 102

© Van Schaik Publishers

The main characteristics of the eleven South African parallel corpora, which do not necessarily have to be used simultaneously, are shown in Table 8.2.

Table 8.2 Parallel Pretoria Corpora

Parallel Corpus Name	Acronym	Tokens	Types
Parallel Pretoria isiNdebele Corpus	//PNC	22 362	7 317
Parallel Pretoria siSwati Corpus	//PSwC	22 054	6 529
Parallel Pretoria isiXhosa Corpus	//PXhC	22 675	7 106
Parallel Pretoria isiZulu Corpus	//PZC	23 948	7 573
Parallel Pretoria English Corpus	//PEC	32 320	3 065
Parallel Pretoria Afrikaans Corpus	//PAfC	31 869	3 425
Parallel Pretoria Xitsonga Corpus	//PXiC	33 884	2 762
Parallel Pretoria Tshivenda Corpus	//PTC	38 603	2 707
Parallel Pretoria Setswana Corpus	//PSTC	37 535	2 840
Parallel Pretoria Sesotho sa Leboa Corpus	//PSC	38 716	2 840
Parallel Pretoria Sesotho Corpus	//PSSC	44 501	2 615

Conjunctivism versus disjunctivism for the official South African languages

In order to truly appreciate the above sizes and values, one needs to look briefly into what is known as "conjunctivism" versus "disjunctivism". This is best done by studying Figures 8.1 versus 8.2.

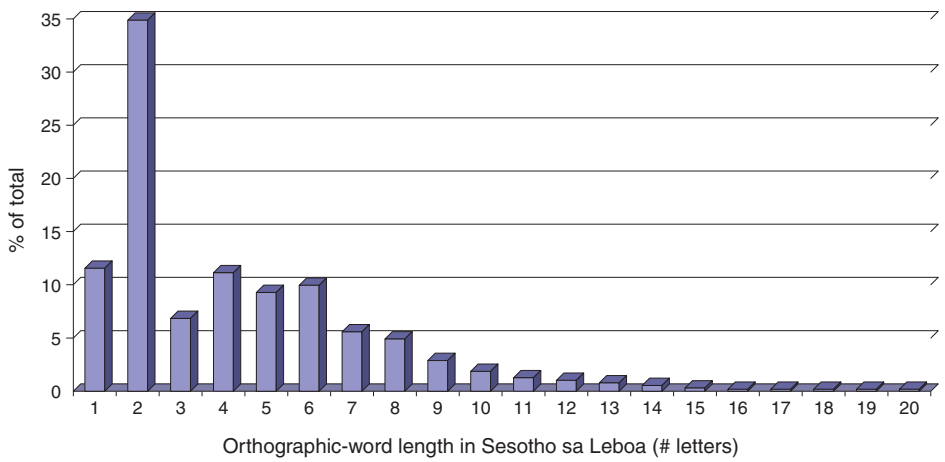


Figure 8.1 Distribution in % of the average length of orthographic words in Sesotho sa Leboa (overall average = 3.88)

© Van Schaik Publishers

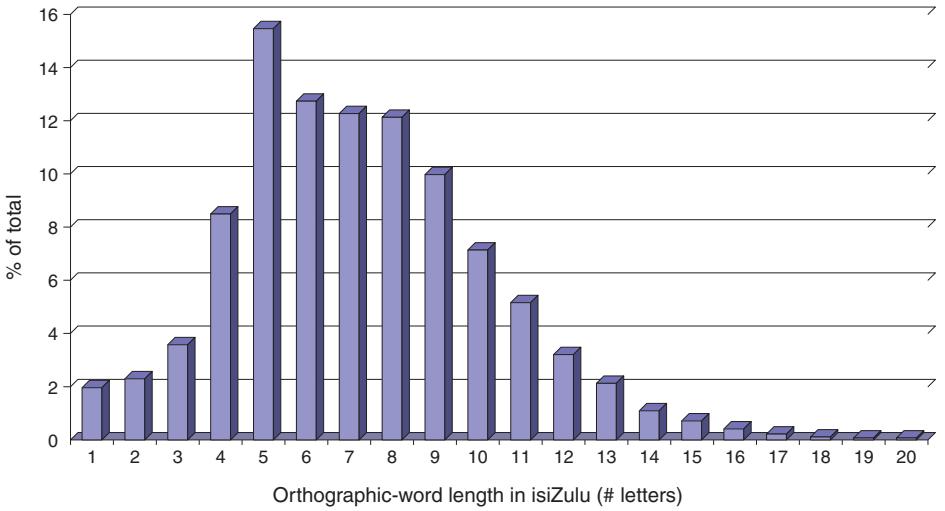


Figure 8.2 Distribution in % of the average length of orthographic words in isiZulu (overall average = 7.18)

Figure 8.1 shows that 35% of all running words in a Sesotho sa Leboa corpus are just two letters long. It is also clear that more than 80% of these tokens consist of less than seven letters. In the case of isiZulu in Figure 8.2, words that are only two letters long represent a mere 2% of the tokens, but five-letter-words are in the majority, representing roughly 15% of the tokens. In contrast to Sesotho sa Leboa, approximately 80% of the tokens in isiZulu are words that have four to ten letters.

In Prinsloo and De Schryver (2002b: 261–262) it was established that the eleven by eleven matrix shown in Table 8.3 can be drawn up, providing a revealing insight into the *relative* degree of conjunctivism / disjunctivism of all eleven official South African languages.

By means of Table 8.3, the relative degree of conjunctivism / disjunctivism of each language can be compared to that of any other language. For instance, the matrix indicates that one orthographic word in isiZulu corresponds to 1.66 orthographic words in Sesotho but only to 0.96 in isiNdebele. Clearly, the conjunctively written South African languages, i.e. the Nguni group consisting of isiNdebele, siSwati, isiXhosa and isiZulu, are located on the left (and at the top) of the matrix; the disjunctively written South African languages, i.e. Xitsonga and Tshivenda, and the Sotho group including Setswana, Sesotho sa Leboa and Sesotho, are located on the right (and at the bottom) of the matrix. English and Afrikaans happen to separate the conjunctive from the disjunctive languages.

Managing eleven parallel corpora: corpus-query software

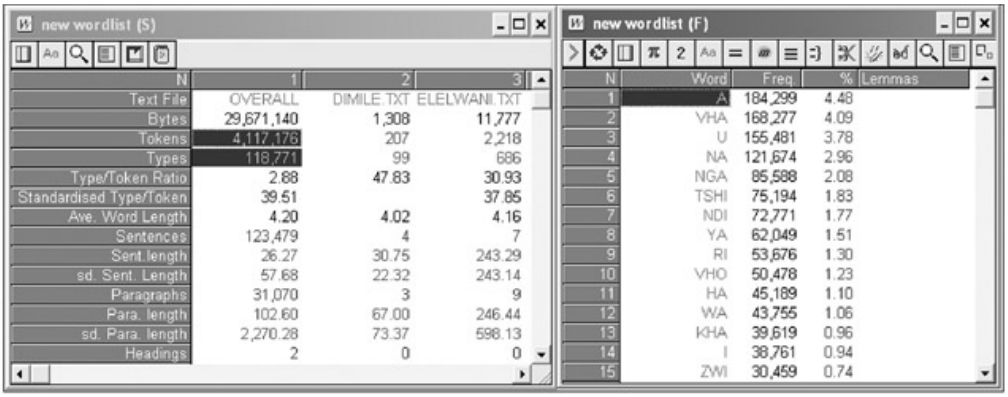
There is no point in developing corpora without adequate corpus-query software. The LGP corpora can be studied with the language-independent *WordSmith Tools* (Scott 1999), while the parallel corpora can be studied by means of the language-

© Van Schaik Publishers

Table 8.3 Eleven by eleven conjunctivism / disjunctivism matrix based on orthographic word counts derived from 55 two-by-two parallel corpora

	isiNdebele	siSwati	isiXhosa	isiZulu	English	Afrikaans	Xitsonga	Setswana	Tshivenda	SsaL	Sesotho
isiNdebele	**1.00**	1.01	1.01	1.04	1.41	1.41	1.61	1.63	1.67	1.73	1.77
siSwati	0.99	**1.00**	1.03	1.04	1.41	1.41	1.61	1.62	1.69	1.72	1.77
isiXhosa	0.99	0.97	**1.00**	1.01	1.36	1.37	1.58	1.58	1.75	1.67	1.71
isiZulu	0.96	0.97	0.99	**1.00**	1.32	1.34	1.54	1.55	1.58	1.60	1.66
English	0.71	0.71	0.74	0.76	**1.00**	1.00	1.15	1.16	1.19	1.24	1.25
Afrikaans	0.71	0.71	0.73	0.75	1.00	**1.00**	1.15	1.16	1.19	1.23	1.24
Xitsonga	0.62	0.62	0.63	0.65	0.87	0.87	**1.00**	1.01	1.05	1.06	1.08
Setswana	0.62	0.62	0.63	0.64	0.86	0.86	0.99	**1.00**	1.03	1.07	1.08
Tshivenda	0.60	0.59	0.57	0.63	0.84	0.84	0.96	0.97	**1.00**	1.03	1.08
SsaL	0.58	0.58	0.60	0.62	0.81	0.81	0.94	0.94	0.97	**1.00**	1.02
Sesotho	0.57	0.57	0.58	0.60	0.80	0.80	0.92	0.92	0.93	0.98	**1.00**

independent *ParaConc* (see Barlow 2003). Some main features of these programs are shown by means of screenshots and brief legends in Figures 8.3 to 8.6, first for WordSmith.



© Van Schaik Publishers

Figure 8.3 Tshivenda LGP corpus (PTC) – Statistical and Word Frequency windows of WordSmith's *WordList* function

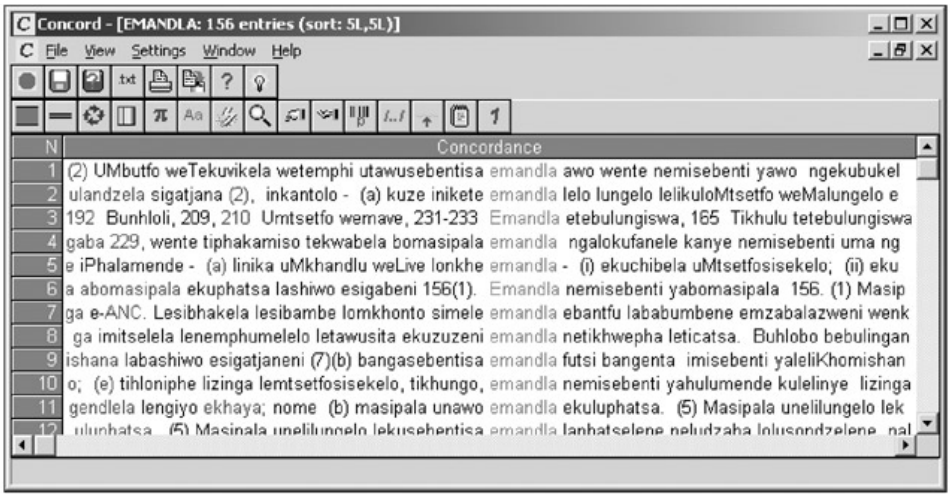


Figure 8.4 siSwati parallel corpus (//PSwC) – Concordance lines window of WordSmith's *Concord* function

ParaConc basically behaves like WordSmith, but up to four languages can be worked with and viewed in an aligned mode simultaneously. As an illustration, in Figure 8.5 a Sesotho sa Leboa text on South Africa's new Coat of Arms is compared with its isiZulu counterpart.[1]

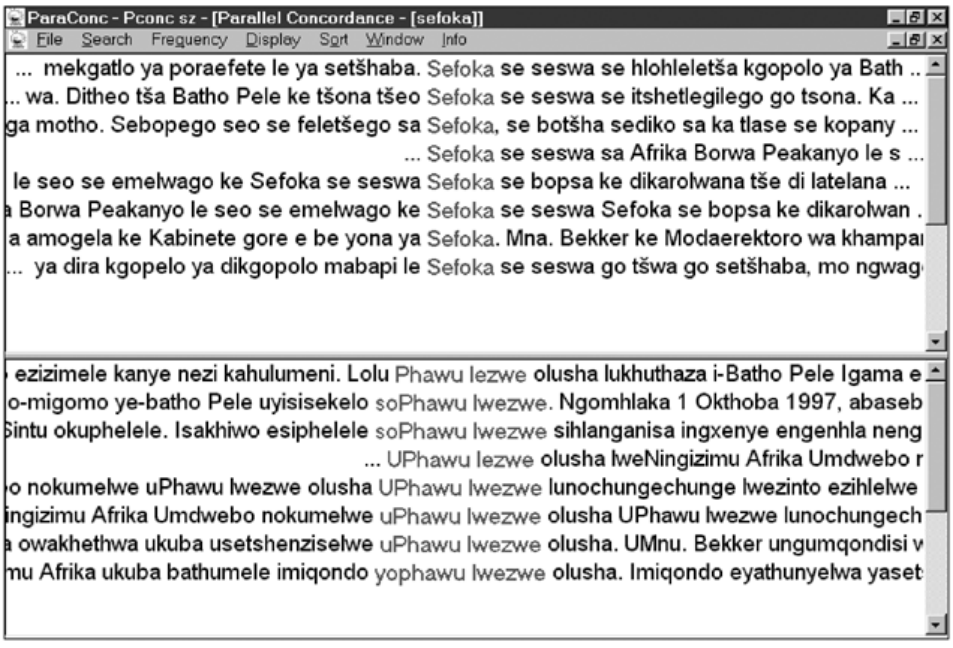


Figure 8.5 Sesotho sa Leboa / isiZulu concordance lines for **coat of arms** in ParaConc

© Van Schaik Publishers

In the concordance lines the so-called "node" for both languages is **coat of arms**, which is **sefoka** in Sesotho sa Leboa and **-phawu l(w)ezwe** in isiZulu. It is clear that, at least on this level, it is not particularly cumbersome to work with a disjunctively written language (in this case Sesotho sa Leboa) and a conjunctively written language (in this case isiZulu) simultaneously.

When studying multi-word units (MWUs) across languages, computing collocates in ParaConc is especially useful. For the same translations of the Coat of Arms document, Figure 8.6 shows the concordance lines for the node **elephant tusks**, a MWU consisting of two words in English, two in isiZulu (**-amazinyo endlovu**), yet three in Sesotho sa Leboa (**manaka a tlou**).

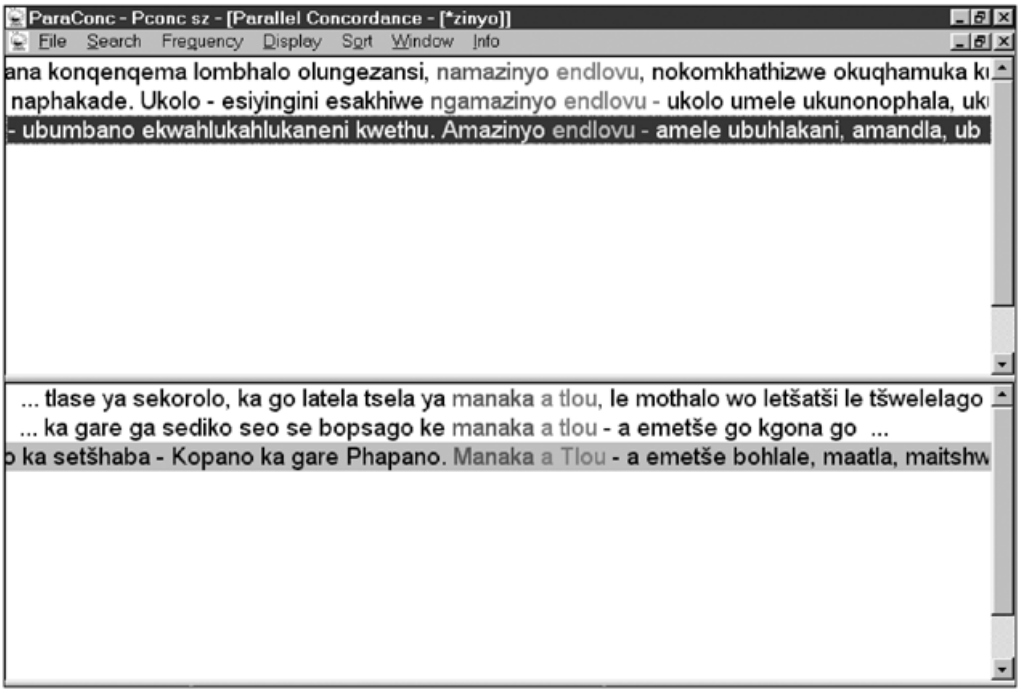


Figure 8.6 isiZulu / Sesotho sa Leboa concordance lines and associated collocates (in red on screen) for **elephant tusks** in ParaConc

Linguistics and applied linguistics

After the brief overview of the various corpora provided above, some of the major corpus-based research results and real-world tools and applications that have materialised over the past few years are presented.

Linguistics:

In linguistics, fundamental new insights were obtained with regard to the class 17 locative prefixes **ku-**, **kwi-** and **ko-** in isiZulu, following a tandem synchronic /

© Van Schaik Publishers

diachronic study in the five-million-word *Pretoria isiZulu Corpus* (PZC). The outcome of this research was described in De Schryver and Gauton (2002) and represents a true methodological breakthrough, since it comprises the first diachronic study for an African language that is based on trends across time in actual language data (as opposed to the method that had hitherto been used, namely that of "reconstructing" towards a hypothetical proto-language).

Phonetics:

A new approach to phonetic research, based on corpus frequencies, was proposed – and applied to Cilubà – by De Schryver (1999). With this method a maximum number of claims, based on a minimum number of words, can be made about the most frequent section of a language's lexicon.

Language teaching and learning:

In the field of language teaching and learning, advances were made in the corpus-aided compilation of textbooks, as well as in facilitating the teaching of morpho-syntactic and contrastive structures, as discussed in Prinsloo and De Schryver (2001a) for Sesotho sa Leboa.

Translation:

In the field of corpus-based translation studies (CTS) it was shown by Gauton et al. (2003) that it is possible at this stage to find translation equivalents across the various South African languages by means of straightforward corpus queries.

Terminology:

In Taljard and De Schryver (2002) a procedure was developed for the semi-automatic term extraction of both single- and multi-word terminology from African-language corpora. The approach was illustrated for Sesotho sa Leboa linguistics terms by comparing the results of a manual excerption with those proffered by software.

Spellcheckers:

Corpus-based spellcheckers for the South African languages, commissioned by the *Department of Arts and Culture* (DAC), were released in June 2003. The methodology, including a large section devoted to Tshivenda, was described in Prinsloo and De Schryver (2003a; 2003c) and Van der Veken and De Schryver (2003).

Lexicography:

Both research and practical tools are wide-ranging in corpus lexicography. On the macrostructural level various innovative approaches for the construction of lemma-sign lists were suggested in De Schryver and Prinsloo (2000b; 2001a; 2003) and in De Schryver (2003c). These approaches are illustrated for languages such as Sesotho sa Leboa and isiNdebele. De Schryver and Prinsloo (2000c) focus on the corpus-based microstructural compilation, while the general corpus-based construc-

© Van Schaik Publishers

tion of dictionary articles was treated, *inter alia*, in De Schryver and Lepota (2001), De Schryver and Prinsloo (2001b) and Nong et al. (2002). In these contributions, Sesotho sa Leboa is often brought into focus in combination with languages such as Cilubà and Kiswahili. This work has also resulted in corpus-based dictionaries, among them a dictionary for Cilubà by De Schryver and Kabuta (1998) and one for Sesotho sa Leboa by Prinsloo and De Schryver (2000).

Extracting data in all official South African languages

Case study 1: In search of ways to find translation equivalents between all official South African languages

At present, the research focus and development of new tools revolves around the *management* of parallel corpora of all eleven official South African languages and the parallel *extraction* of useful data from any combination of these corpora. Until recently WordSmith was used for this purpose, whereby data were extracted from each LGP corpus separately and then cross-compared. ParaConc is currently used to query any selection of parallel corpora simultaneously. In this regard, one may consider the manipulation of the eleven-language-versions of the document "What is the ANC?"[2], as the first set of illustrations of how particular words and their translation equivalents across languages may be semi-automatically extracted by means of ParaConc queries. Figure 8.7 is an extract of the alignment for two of the eleven languages, viz. siSwati and Sesotho.

File Alignment Search Frequency Window Info

NGABE IYINI I-AFRICAN NATIONAL CONGRESS?	AFRICAN NATIONAL CONGRESS (ANC) KE ENG?
* I-ANC yenhlangano yavelonkhe yenkhululeko.	* ANC ke mokgatlo o kwanetsang tokoloho ya setjhaba.
Yasungulwa nga-1912 kutawuhlanganisa bantfu labaMnyama lkanye nekuba sembili kumzabalazo wenqucuko lesisekelo kutepolitiki, tenhlalakahle netemnotfo.	O thehilwe ka selemo sa 1912 ho kopanya ma-Afrika le ho etella pele boitseko ba diphetoho tsa sethato dipolotiking, kahisanong le moruong.
* Eminyakenishumi leyimfica i-ANC iye yahola umzabalazo lomelene nebuhlanga nencindzetelo, yahlelembisa kudvuba kwelinyenti, ivusa lugcozi kumphakatsi wamhlabawonkhe yaphindze yatsatsa umzabalazo wetikhali kumelana nelubandlululo.	* Dilemomng tse mashome a robong kaofela ANC ha esale e ntse e etelletse pele boitseko kgahlanong le kgetholloya bo-morabe, mme e hlophisa matsholo a bongata a kganyetso nmoho le ho susumelletsa dinaha tsa matjhaba ho ntshetsa pele boitseko, esita le ho nka dihlomo twantshong ya apartheid.
* I-ANC iye yazuza imphumelelo lebonakalako yedemokhrasi elukhetfweni lwanga-1994, lapho yaniketwa ligunya lelicacile lekubamba tinkhulumiswano ngeMtsetfosisekelo lomusha wedemokhrasi weNingizimu Afrika.	* ANC e ile ya fihlela katleho e kgolo dikgethong tsa dimokerasi tsa 1994, moo e ileng ya fuwa thomo e tileng ya ho rerisana ka Molaotheo o motjha wa demokerasi wa Afrika Borwa.
LoMtsetfosisekelo lomusha wemukelwa nga-1996.	Molaotheo oo o motjha o ile wa amanelwa ka 1996.
* I-ANC iphindze yakhetfwa nga-1999 kuhulumende wavelonkhe nakumaphrovinsi ngeliguruva lelengetiwe.	* ANC e ile ya boela ya kgethwa ka 1999 boemong ba mmuso wa naha le mebusong ya diprofensi ka thomo e ekeditsweng.
* Tinchubomgomo te-ANC tesekeleke kumalunga ayo kantsi nebuholi bayo bunekutiphendvulela kumalunga wonkhe.	* Maano aANC a ralwa ke ditho tsa yona mme boetapele ba yona bo ikarabela ho ditho tsa yona.
* Bulunga be-ANC buvulekele bonkhe baseNingizimu Afrika labangetulu kweminyaka lengu-18, kungakhatsalekile kutsi wakuphi luhlanga, umbala nobe inkholelo, lowemukela tesekelo, tinchubomgomo kanye netinhlelo tayo.	* Botho ba ANC bo buletswe Ma-Afrika Borwa ohle a dilemong tse ka hodimo ho tse 18, ho sa natswe morabe, mmala le bodumedi, mme e be batho ba tla amohela metheo, maano le mananeo a mokgatlo.
NGABE TYINI TINJONGO NETINHLOSO TE-ANC?	SEPHEO LE MAIKEMISETSO A ANC KE AFE?
* Inhloso lesembili ye-ANC kwakhiwa kwemphakatsi wedemokhrasi lobumbene, longakhetsi ngebuhlanga futsi longakhetsi ngebulili.	* Sepheo sa bohlokwa ke ho theha setjhaba sa demokerasi se kopaneng, se hlokang kgethello ya bo-morabe le kgetholloya bong.
* Loku kusho kukhululiwa kwebantfu be-Afrika ikakhulukati kanye nebantfu labamnyama jikelele ekuboshweni kutepolitiki netemnotfo.	* Hona ho bolela tokoloho dipoloyiking le moruong haholo-holo tokoloho ya Ma-Afrika le batho ba batsho ka kakaretso.
Kusho kukhuphula lzinga lemphilo labo bonkhe baseNingizimu Afrika, ikakhulukati labaphuyile.	Ho bolela ho phahamisa boleng ba bophelo ba Ma-Afrika ohle, haholo-holo ba hutsanehileng.
* Umzabalazo wekuzuza lenhloso ubitwa ngeMzabalazo weNtsandvo yeliNyenti kuVelonkhe.	* Boitseko ba ho fihlela sepheo sena bo bitswa Ntwa ya Naha ya ho tlisa Diphetoho tsa demokerasi.
NGABE YINI LEYINKHOMBANDLELA KUNCHUBOMGOMO YE-ANC?	KE SE FE SE TATAISANG LEANO LA ANC?
I-Freedom Charter, leyemukelwa kuKhongolose weBantfu nga-1955, seloku ingumculu losisekelo senchubomgomo ye-ANC.	Lengolo le Phatlalatsang Ditokelo tsa Tokoloho (Freedom Charter), le ileng la amanelwa ke seboka sa Batho ka 1955, le dutse e ntse ele lengolo la motheo wa leano la ANC.
I-Freedom Charter imemetela kutsi	Lengolo lena, laDitokelo tsa Tokoloho le phatlalatsa ho re:
* Bantfu batawubusa	* Batho ba tla busa
* Onkhe emacembu avelonkhe atawuba nemalungelo lalinganako	* Batho ba mefuta yohle ba tla ba le ditokelo tse lekanang
* Bantfu batawuba nesabelo kunmotfo welive	* Batho bohle ba tla arolelana moruo wa naha ka ho lekana
* Umhlaba utawabuya kulabo labawusebentako	* Naha e tla aroleiwa bohle ba sebetsang ho yona
* Bonkhe batawulingana embi kwemtsetfo	* Bohle re tla lekana ka pela molao

Figure 8.7 siSwati / Sesotho alignment of the document "What is the ANC?" in ParaConc

Figure 8.8 shows parallel concordance lines for **people** in the Afrikaans, Setswana, isiXhosa and Xitsonga versions of the same document.

Figure 8.8 Afrikaans / Setswana / isiXhosa / Xitsonga parallel concordance lines for **people** in ParaConc

So-called "hot words", i.e. likely translation equivalents, in relation to the concept **people**, may be isolated from the different-language-versions and calculated in ParaConc. Compare the output in Table 8.4 for the Afrikaans **mense** ("people") in Setswana, isiXhosa and Xitsonga respectively.

Table 8.4 Setswana / isiXhosa / Xitsonga hot words for the Afrikaans **mense** ("people") in ParaConc [bolded hot words are correct]

Setswana ✓		**isiXhosa ✓**		**Xitsonga ✓**	
Rank	Hot word	Rank	Hot word	Rank	Hot word
41.16	ba	27.44	**abantu**	25.87	**vanhu**
39.20	**batho**	12.62	umbala	16.50	a
21.76	lefatshe	9.64	ye-anc	16.50	muvala
19.53	la	9.64	basemzantsi	12.35	yimele
15.49	kaya	4.08	engundoqo	8.93	misava
15.49	bo	4.08	se-anc	6.85	matimba
12.28	tla	4.08	ngokubanzi	6.76	leyi
10.32	baagi	4.08	amandla	6.13	lava
6.71	a	3.68	afrika	4.78	xiavo
4.05	kgololesego	2.59	yabo	4.06	afrika

© Van Schaik Publishers

For Setswana the correct (✓) translation equivalent (**batho**) is listed in second position, while the correct equivalent is the top hot word for isiXhosa (**abantu**) and Xitsonga (**vanhu**).

As pointed out above, semi-automatic term extraction in a single language in isolation was described in Taljard and De Schryver (2002) and the feasibility of subsequently finding translations for those terms across languages was described in Gauton et al. (2003). A fully automated extraction of keywords from the above-mentioned Sesotho sa Leboa text on the new South African Coat of Arms (CoA), for example, by means of the *KeyWord* function of WordSmith, results in the data shown in Table 8.5 (see Taljard & De Schryver 2002: 52–54).

Table 8.5 Automatic keyword extraction from the Sesotho sa Leboa version of a document on the new South African Coat of Arms (CoA)

N	Keyword	Translation	CoA Count	CoA %	PSC Count	PSC %	Keyness
1	sefoka	coat of arms	10	0.96	231		87.3
2	ditirelo	services	8	0.77	114		77.3
3	bontšha	show	11	1.06	640	0.01	76.1
4	seswa	(something) new cl. 7	6	0.58	18		75.2
5	Afrika	Africa(n)	9	0.87	941	0.02	51.9
6	tlhame	secretary bird	4	0.39	19		46.9
7	barulaganyi	designers	3	0.29	3		42.8
8	Borwa	South	8	0.77	1 137	0.02	41.4
9	badiriši	users	3	0.29	9		37.6
10	setšhaba	nation	10	0.96	3 204	0.06	36.2
11	lebišitšwe	is / are aimed at	3	0.29	14		35.3
12	batho	people	16	1.54	12 232	0.24	33.1
13	se	subj. conc. cl. 7; dem. cl. 7; ...	39	3.76	73 986	1.43	27.6
14	mmušo	government	6	0.58	1 214	0.02	26.9
15	manaka	tusks	3	0.29	72		25.9
16	leswa	(something) new cl. 5	3	0.29	73		25.9
17	emela	represent(s)	4	0.39	308		25.5
18	tshedimošo	information	3	0.29	85		25.0
19	mabapi	with regard to, regarding	5	0.48	830	0.02	24.4

© Van Schaik Publishers

By simply scrolling over the keywords in Table 8.5 – *keywords that are proffered fully automatically by the KeyWord function* – one can deduce that the text in ques-

tion must provide information regarding the new South African coat of arms. It also seems as if the designers aimed at showing symbols, such as a secretary bird and tusks, and that the text represents the government's attempt to provide new services to the nation's people / users.[3]

As illustrated in Figure 8.8 and Table 8.4, using ParaConc it is possible to call up concordance lines in a certain language and to instruct the software to calculate the hot words in the languages for which there are aligned parallel corpora. Instead of calculating the keywords, for example in Setswana, by means of WordSmith's KeyWord function, and then cross-comparing the above Sesotho sa Leboa keywords with the Setswana keywords in order to try to "match" them, it is now possible to investigate how well ParaConc is able to *automatically* translate Sesotho sa Leboa into Setswana.

Parallel concordance lines for the Sesotho sa Leboa and Setswana CoA document were run for each of the 19 keywords in Table 8.5 and the software was instructed to calculate the hot words in each case. In 14 (74%) of the 19 cases correct Setswana translation equivalents for the Sesotho sa Leboa keywords were found among the hot words. Moreover, in eight instances, the correct translation equivalent was the top hot word and in 12 instances the translation was among the top three. One is therefore bound to conclude that, in searching for translation equivalents between two disjunctively written South African languages (as is the case for Sesotho sa Leboa and Setswana), ParaConc's hot-word function is perfectly capable of finding the requested translation in three quarters of the cases.

However, finding translation equivalents between a disjunctively and a conjunctively written language or even between conjunctively written languages by means of the same technique, is less successful. Upon studying the isiZulu hot words suggested as translation equivalents for the CoA Sesotho sa Leboa keywords, for example, it becomes clear that correct equivalents are only supplied in seven out of the 18 cases (or 39%).[4] For terms such as **sefoka**, "coat of arms", **seswa**, "(something) new cl. 7" and **tlhame**, "secretary bird", the correct equivalents are indeed generated among the top hot words, as can be seen from Table 8.6.

Table 8.6 isiZulu hot words generated for the Sesotho sa Leboa words **sefoka**, "coat of arms", **seswa**, "(something) new cl. 7" and **tlhame**, "secretary bird" [bolded hot words are correct]

sefoka ✓		**seswa ✓**		**tlhame ✓**	
Rank	Hot word	Rank	Hot word	Rank	Hot word
78.56	olusha	100.25	**olusha**	67.16	**intinginono**
69.06	**lwezwe**	57.74	lwezwe	63.49	–
44.03	**uphawu**	53.74	afrika	42.77	kakhulu
34.53	**sophawu**	40.50	sophawu	24.73	kanye
26.51	**lezwe**	36.49	uphawu	18.38	ukuhlangana

© Van Schaik Publishers

Table 8.6 *Continued*

24.51	afrika	34.49	pele	18.38	endlovu
15.00	pele	32.48	lezwe	10.36	futhi
7.50	imiqondo	17.24	imiqondo		
7.50	eningizimu	17.24	eningizimu		
7.50	baseningizimu	17.24	baseningizimu		

No suitable translation equivalents (**✗**) were generated for **ditirelo**, "services", **bontšha**, "show" and **barulaganyi**, "designers", as can be seen from Table 8.7.

Table 8.7 isiZulu hot words generated for the Sesotho sa Leboa words **ditirelo**, "services", **bontšha**, "show" and **barulaganyi**, "designers" [not a single hot word is correct]

ditirelo ✗		**bontšha ✗**		**barulaganyi ✗**	
Rank	Hot word	Rank	Hot word	Rank	Hot word
28.87	amazinga	17.62	kanye	46.47	ukuthi
17.24	amakhasimende	15.93	–	28.43	–
3.61	ukuze	10.65	kakhulu	25.24	imiqondo
		2.64	futhi		
		2.32	ukuhlangana		
		2.32	eningizimu		

Closer manual comparison of the texts reveals that a number of factors contribute to this relatively low strike rate, among them disjunctivism versus conjunctivism. The first and probably the most obvious case occurs where, in contrast to the Sesotho sa Leboa translation, no attempt was made to translate *Government Communication and Information System* (GCIS) in the isiZulu translation: "Ba ditirelo tša Mmušo tša Dikgokagano le Tshedimošo" (Sesotho sa Leboa) versus "I-Government Communication and Information System" (isiZulu). ParaConc's potential to find hot words for keywords such as **mmušo**, "government" and **tshedimošo**, "information" is impeded by the adoption of the English wording for GCIS. Consequently, options such as **uhulumeni** or **umbuso** for "government" and **ulwazi**, **umbiko**, **ukwaziswa**, **ukwazi**, **imfundiso**, etc., for "information" have not been considered, as was indeed done for the underlined options elsewhere in the translation. The fact that **tshedimošo** occurs only three times in the text means that ParaConc's chances of correctly selecting an appropriate isiZulu equivalent are heavily reduced. In the

© Van Schaik Publishers

case of **mmušo**, which occurs six times in the text, five potential chances remain and one could expect that the software would find **uhulumeni** as a translation equivalent with relative ease. Clearly, in this instance, the negative effect of finding hot words in a conjunctively written language comes to the fore. Although **mmušo** is easily detectable as a word in all six instances in Table 8.8, irrespective of the co-text, **uhulumeni** is found only once as a stand-alone "orthographic word".

Table 8.8 Sesotho sa Leboa versus isiZulu with regard to the translation of **government**

Sesotho sa Leboa	isiZulu	Translation
... la **Mmušo**	... likaHulumeni	... of the **government**
... tša **mmušo**	... kahulumeni	... of the **government**
... **Mmušo**	... I-Government	... **Government**
... bodirela **mmušo**	– (*not translated*)	... those working for the **government**
... bodirela **mmušo**	... abasebenzela **uhulumeni**	... those working for the **government**
... bjalo ka **mmušo**	... njengohulumeni	... like the **government**

This implies that a potential of six chances to spot **uhulumeni** as the hot word for **mmušo** is reduced to one. Dedicated software will thus need to be developed that can perform a morphological analysis in order to determine the relation between **uhulumeni**, **kahulumeni**, **likaHulumeni**, **njengohulumeni**, etc., in the detection of hot words. Hence it is not surprising that the appropriate hot word was only ranked in sixth place and that it occurred as **kahulumeni** rather than **uhulumeni**, as shown in Table 8.9.

Table 8.9 isiZulu hot words generated for the Sesotho sa Leboa word **mmušo**, "government"

mmušo ✓	
Rank	Hot word
26.52	ukuthi
19.89	pele
15.26	baseningizimu
5.42	afrika
4.63	kakhulu
4.63	**kahulumeni**
2.62	ukuze

© Van Schaik Publishers

Note also that the phrase "those working for the government" in Table 8.8 was translated twice in the same way as **bodirela mmušo** in Sesotho sa Leboa but only once in a directly comparable way as **abasebenzela uhulumeni** in isiZulu.

This presentation demonstrates that one may be cautiously optimistic when it comes to the use of ParaConc to further query parallel corpora. Nevertheless, for the conjunctively written languages in particular, it seems wise to embark on developing customised tools that include (language-specific) morphological analysis components.

Case study 2: Multidimensional lexicographic Rulers for all official South African languages

Among the practical lexicography tools that are currently being developed, it is necessary to include the design of a set of pioneering measurement and prediction instruments aimed at assisting the South African lexicographers with the compilation of their national dictionaries. These so-called multidimensional lexicographic Rulers draw heavily on statistics derived from electronic corpora on the one hand, as well as on existing dictionary data, on the other hand.

In order to treat a representative and balanced section of the lexicon in any given dictionary, it must be borne in mind that the various alphabetical categories or stretches – i.e. those sections containing all lemma signs that start with the letter **A**, then **B**, etc. – obviously do not contain the same number of articles in each category. A quick glance at any English LGP dictionary, for example, immediately reveals that the largest categories by far are **C** and **S** in English and that categories **X**, **Y** and **Z** contain only a few lemma signs. Compared to the category **S**, categories **X**, **Y** and **Z** are actually virtually empty in English, which means that only a few dictionary pages are needed for the latter. The exact allocations to each category are clearly language-dependent and the question is whether a specific distribution, preferably one that could be accurately measured and predicted, exists for the different categories in any given language. In-depth and exhaustive research for a number of languages (and for various types of dictionaries) carried out by the authors of the current article has proved that this is indeed possible.

A remarkable consistency in respect of the balance between alphabetical stretches has been detected in dictionaries and in corpora. This consistency is observed with regard to the number of lemma signs treated in each alphabetical category or the number of pages dedicated to each alphabetical stretch of a dictionary on the one hand, and lemmatised as well as unlemmatised alphabetically sorted wordlists culled from corpora on the other hand. Compare, for example, the data in Table 8.10 for Afrikaans where the average breakdown based on the number of pages dedicated to the treatment of each alphabetical category in five different Afrikaans LGP dictionaries is compared to an alphabetical word list culled from the *Pretoria Afrikaans Corpus* (PAfC), which is an Afrikaans LGP corpus (see Prinsloo & De Schryver 2003b: 110).

One can conclude from Table 8.10 that, except for the smaller alphabetical categories **C**, **X**, **Y** and **Z**, there is a rather striking correlation between the average of

© Van Schaik Publishers

Table 8.10 Average page allocation in five Afrikaans dictionaries (D1–D5) versus PAfC

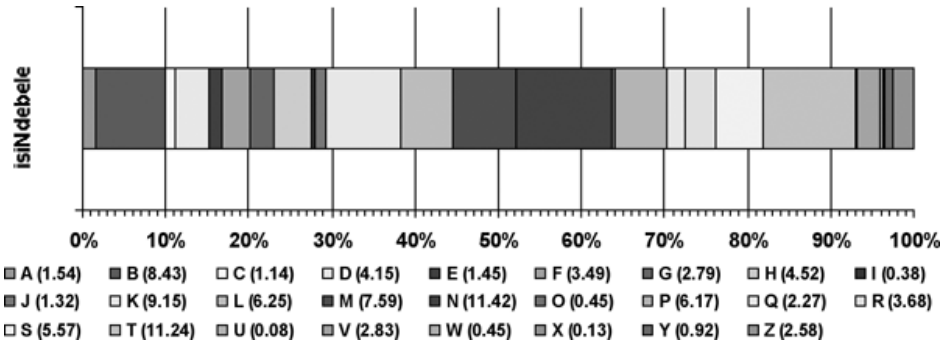
	D1-D5	DIFFERENCE		PAfC	
	%	abs. %	rel. %	%	
A	5.72	–0.06	–1.08	5.79	**A**
B	7.44	+0.26	+3.63	7.18	**B**
C	0.29	–0.91	–76.03	1.20	**C**
D	4.70	–0.34	–6.80	5.05	**D**
E	2.16	–0.70	–24.56	2.87	**E**
F	1.25	–0.43	–25.52	1.68	**F**
G	5.97	–0.64	–9.73	6.61	**G**
H	4.76	+0.32	+7.31	4.44	**H**
I	2.34	–0.38	–13.94	2.72	**I**
J	0.76	–0.33	–30.25	1.09	**J**
K	7.55	+1.43	+23.45	6.12	**K**
L	3.34	–0.62	–15.57	3.96	**L**
M	4.10	–0.55	–11.86	4.65	**M**
N	2.36	–0.71	–23.02	3.07	**N**
O	6.95	+0.67	+10.74	6.28	**O**
P	4.21	+0.23	+5.78	3.98	**P**
Q	0.02	+0.00	+22.41	0.02	**Q**
R	3.58	–0.29	–7.49	3.87	**R**
S	12.72	+1.98	+18.47	10.74	**S**
T	4.40	–0.39	–8.07	4.79	**T**
U	1.95	+0.08	+4.34	1.87	**U**
V	9.03	+1.77	+24.41	7.26	**V**
W	4.09	+0.14	+3.45	3.95	**W**
X	0.06	–0.06	–53.03	0.12	**X**
Y	0.17	–0.29	–62.61	0.46	**Y**
Z	0.05	–0.20	–78.36	0.25	**Z**
	100.00	***r* = 0.983**		100.00	

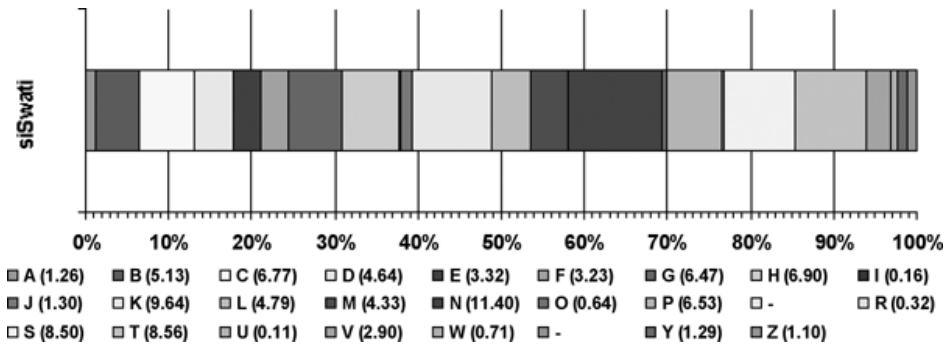
© Van Schaik Publishers

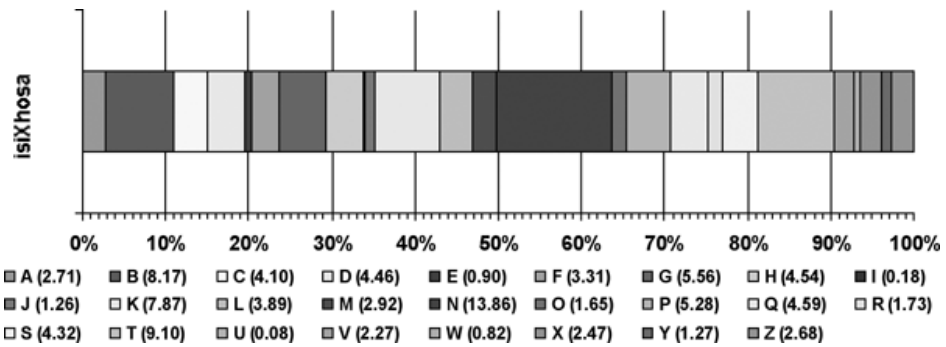
the five dictionaries and the corpus suggestion. The correlation coefficient *r* is as high as 0.983.[5]

Rulers are built from such statistics and were introduced for Afrikaans and English in Prinsloo and De Schryver (2002a; 2003b), for one of the conjunctively written languages (i.e. isiNdebele) in De Schryver (2003c), and for one of the disjunctively written languages (i.e. Sesotho sa Leboa) in Prinsloo and De Schryver (2004). For the purpose of this article the full set has been completed (based on De Schryver 2003b) and Rulers for *all eleven* official South African languages are presented for the very first time in print below.

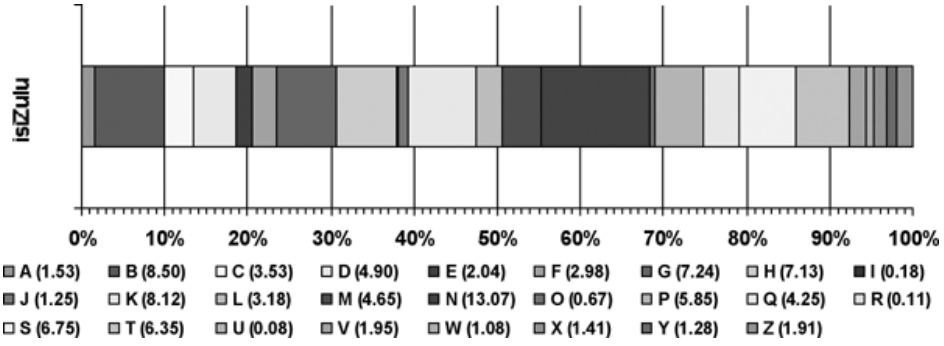
Rulers for the South African Nguni languages (isiNdebele, siSwati, isiXhosa and isiZulu):



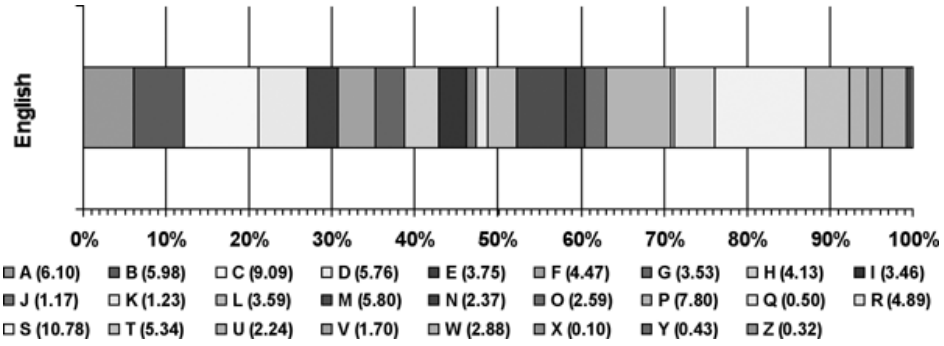


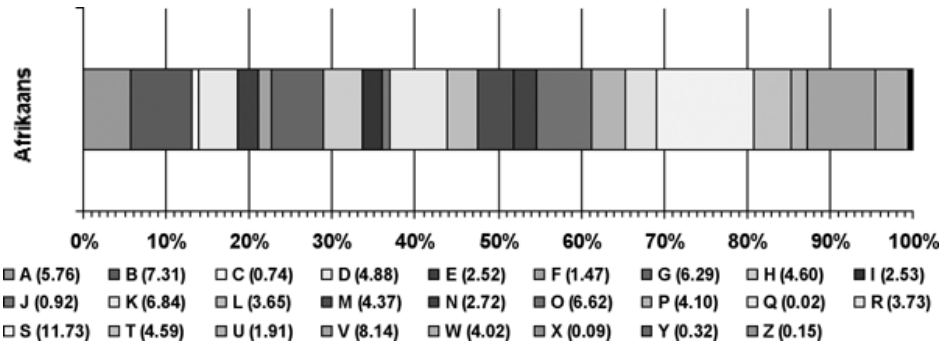


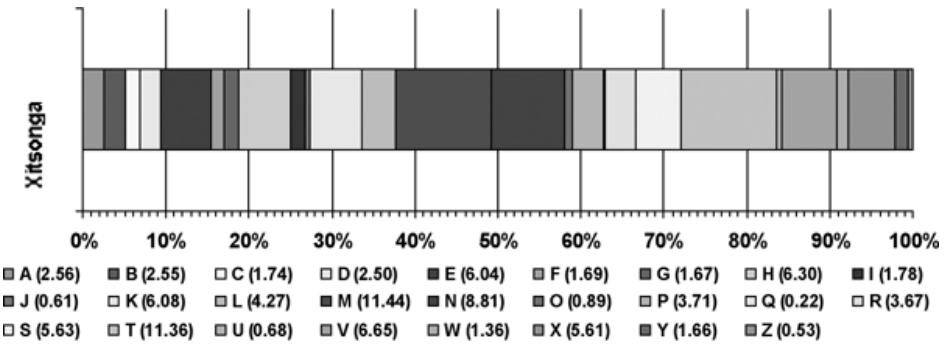
© Van Schaik Publishers



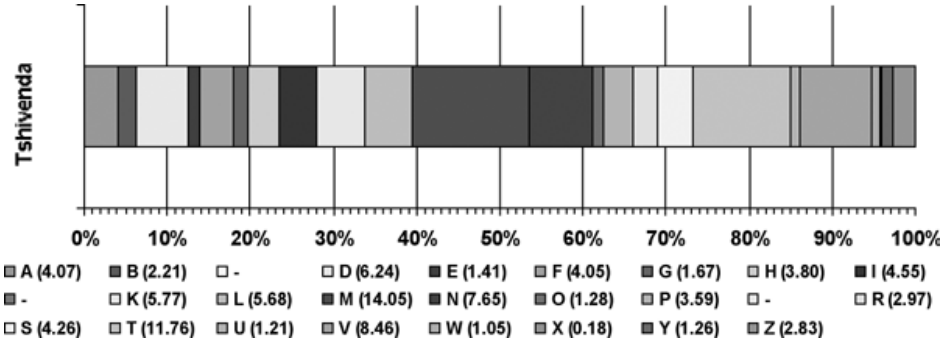
Rulers for English, Afrikaans, Xitsonga and Tshivenda:



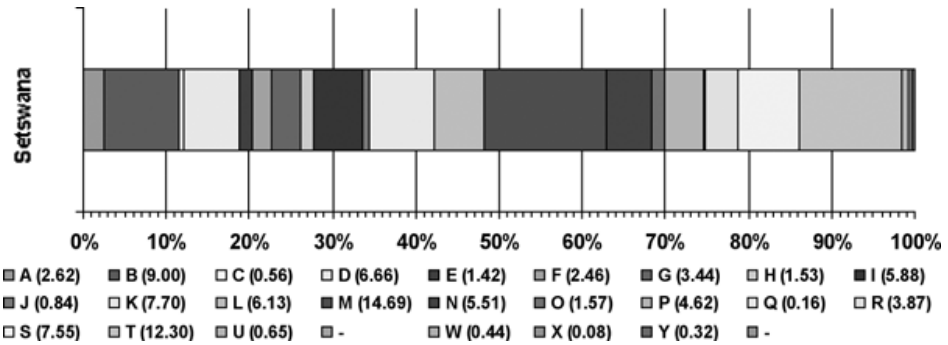


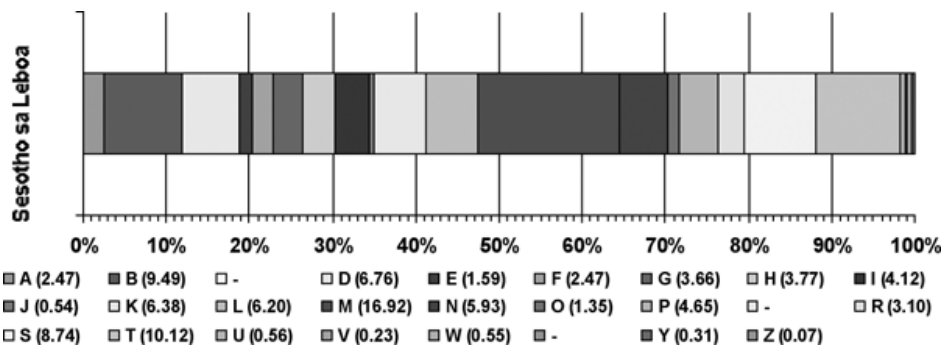


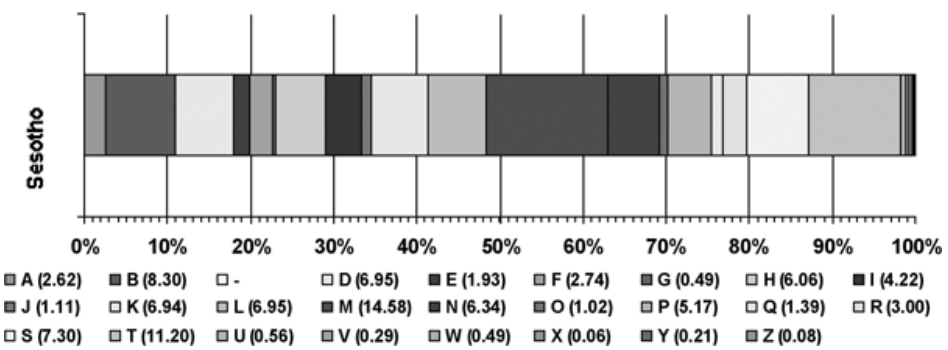
© Van Schaik Publishers



Rulers for the South African Sotho languages (Setswana, Sesotho sa Leboa and Sesotho):







© Van Schaik Publishers

In their most basic forms, Rulers are abstract entities, as percentages straightforwardly correlate with the alphabetical distribution in semasiological dictionaries. These percentages (which can of course be made as fine-grained as one wishes) can be converted in terms of the number of pages, the number of lemma signs and the required compilation time for any (sub)section of the dictionary. They can furthermore be used to measure aspects of published dictionaries that are about to be revised; they may guide the further compilation of already existing lexicography projects; and they can even predict features of envisaged reference works.

If the dictionaries are being compiled from **A** to **Z**, the set of Rulers indicates that lexicographers at the English Unit will still be compiling the stretch **C** at the 20% mark, while lexicographers at the Xitsonga Unit should already have reached the stretch **H** at the same 20% mark. Even among related languages there is a considerable difference. At the 20% mark in the Nguni group, isiNdebele lexicographers are about to start the stretch **G**, isiZulu and isiXhosa lexicographers are completing **E**, yet siSwati lexicographers are only halfway through **E**. Whereas **N** is the largest alphabetical stretch for the Nguni languages, **M** is the largest for all Sotho languages, as well as for Xitsonga and Tshivenda. For English and Afrikaans, **S** comprises the largest alphabetical stretch.

For example, as a practical illustration, say three years (at five days a week) are to be devoted to the compilation of a Tshivenda dictionary consisting of a projected 400 pages and 7 000 lemma signs; the Tshivenda Ruler *predicts* that 9.0 weeks will have to be spent on the compilation of the stretch **K**, for which 404 lemma signs will cover 23.1 pages, and that work on **W** should last 1.6 weeks for the compilation of 73 articles on approximately 4.2 pages, etc. It is evident that corpora, with Rulers acting as go-betweens, can have a *direct* influence on management strategies at all eleven official South African *National Lexicography Units* (NLUs).

The future: TshwaneLex, TshwaneConc and TshwaneSpell – a suite of HLT products for the South African languages

The next major challenge consists of designing a fully integrated dictionary compilation package cum corpus tool cum spellchecker. The basic dictionary compilation software has already been created by *TshwaneDJe HLT* (http://tshwanedje.com/) and is known as *TshwaneLex* (Joffe et al. 2003a; 2003b). All the dictionaries that are currently being compiled with TshwaneLex are corpus-based and WordSmith is used to query these corpora running *in concurrence* with TshwaneLex. The intention is to extend the functionality of TshwaneLex by linking it with *TshwaneConc* so that corpora will be directly accessible from within TshwaneLex. Furthermore, Rulers are built-in components of TshwaneLex. In a subsequent stage, corpus-based spellchecker functionality, in the form of *TshwaneSpell,* will be added to enable the South African lexicographers to simultaneously check the orthography of their newly created dictionary text.

TshwaneLex, TshwaneConc and TshwaneSpell are but the first components of a modern and fully integrated suite of HLT products specifically designed for the true *management* of data in all official South African languages – and beyond.

© Van Schaik Publishers

REFERENCES

Barlow, M. 2003. *ParaConc: a concordancer for parallel texts*. Houston, TX: Athelstan. For this software, also see <http://www.athel.com/>.

De Schryver, G-M. 1999. *Cilubà phonetics, proposals for a "corpus-based phonetics from below"-approach*. Ghent: Recall.

De Schryver, G-M. 2002. Web for/as corpus: a perspective for the African languages. *Nordic Journal of African Studies*, 11(2): 266–282.

De Schryver, G-M. (Ed.). 2003a. *TAMA 2003 South Africa: CONFERENCE PROCEEDINGS*. Pretoria: (SF)^2 Press.

De Schryver, G-M. 2003b. *Rulers & block systems for South African lexicography*. Paper presented at the Lexicography Seminar of the Pan South African Language Board (PanSALB). Pretoria, 12 November.

De Schryver, G-M. 2003c. 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. & Gauton, R. 2002. The isiZulu locative prefix ku- revisited: a corpus-based approach. *Southern African Linguistics and Applied Language Studies*, 20(4): 201–220.

De Schryver, G-M. & Kabuta, N.S. 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)*. Ghent: Recall.

De Schryver, G-M. & Lepota, B. 2001. The lexicographic treatment of days in Sepedi, or When mother-tongue intuition fails. *Lexikos*, 11 (AFRILEX-reeks/series 11: 2001): 1–37.

De Schryver, G-M. & Prinsloo, D.J. 2000a. The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies*, 18(1–4): 89–106.

De Schryver, G-M. & Prinsloo, D.J. 2000b. 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. & Prinsloo, D.J. 2000c. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure. *South African Journal of African Languages*, 20(4): 310–330.

De Schryver, G-M. & Prinsloo, D.J. 2001a. 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. & Prinsloo, D.J. 2001b. Towards a sound lemmatisation strategy for the Bantu verb through the use of *frequency-based tail slots* – with special reference to Cilubà, Sepedi and Kiswahili. In J.S. Mdee & H.J.M. Mwansoko (Eds), *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam, 216–242, 372.

De Schryver, G-M. & Prinsloo, D.J. 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.

Gauton, R., Taljard, E. & De Schryver, G-M. 2003. Towards strategies for translating terminology into all South African languages: a corpus-based approach. In G-M. de Schryver (Ed.), 2003a: 81–88.

Joffe, D., De Schryver, G-M. & Prinsloo, D.J. 2003a. Introducing TshwaneLex – a new computer program for the compilation of dictionaries. In G-M. de Schryver (Ed.), 2003a: 97–104.

© Van Schaik Publishers

Joffe, D., De Schryver, G-M. & Prinsloo, D.J. 2003b. Computational features of the dictionary application "TshwaneLex". *Southern African Linguistics and Applied Language Studies*, (Special issue on "Language Technology in Southern Africa: resources and applications"), 21(4): 239–250.

Nong, S., De Schryver, G-M. & Prinsloo, D.J. 2002. Loan words versus indigenous words in Northern Sesotho – a lexicographic perspective. *Lexikos*, 12 (AFRILEX-reeks/series 12: 2002): 1–20.

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. & De Schryver, G-M. (Eds). 2000. *SeDiPro 1.0, First Parallel Dictionary Sepêdi – English*. Pretoria: University of Pretoria.

Prinsloo, D.J. & De Schryver, G-M. 2001a. Corpus applications for the African languages, with special reference to research, teaching, learning and software. *Southern African Linguistics and Applied Language Studies*, 19(1–2): 111–131.

Prinsloo, D.J. & De Schryver, G-M. 2001b. Monitoring the stability of a growing organic corpus, with special reference to Sepedi and Xitsonga. *Dictionaries: Journal of The Dictionary Society of North America*, 22: 85–129.

Prinsloo, D.J. & De Schryver, G-M. 2002a. Designing a measurement instrument for the relative length of alphabetical stretches in dictionaries, with special reference to Afrikaans and English. In A. Braasch & C. Povlsen (Eds). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*. Copenhagen: Center for Sprogteknologi, Københavns Universitet, 483–494.

Prinsloo, D.J. & De Schryver, G-M. 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. & De Schryver, G-M. 2003a. Towards second-generation spellcheckers for the South African Languages. In G-M. de Schryver (Ed.), 2003a: 135–141.

Prinsloo, D.J. & De Schryver, G-M. 2003b. 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]. In W. Botha (Ed.), *'n Man wat beur: huldigingsbundel vir Dirk van Schalkwyk*. Stellenbosch: Buro van die WAT, 106–126.

Prinsloo, D.J. & De Schryver, G-M. 2003c. Non-word error detection in current South African spellcheckers. *Southern African Linguistics and Applied Language Studies* 21(4) (Special issue on "Language technology in Southern Africa: resources and applications"): 307–326.

Prinsloo, D.J. & De Schryver, G-M. 2004. Crafting a multidimensional Ruler for the compilation of Sesotho sa Leboa dictionaries. In J. Mojalefa (Ed.). *Rabadia ratšhatšha: in-depth literature, linguistics, translation and lexicography studies in African languages. Festschrift in honour of P.S. Groenewald*. Pretoria: J.L. van Schaik.

Scott, M. 1999. *WordSmith Tools version 3*. Oxford: Oxford University Press.

Taljard, E. & De Schryver, G-M. 2002. Semi-automatic term extraction for the African languages, with special reference to Northern Sotho. *Lexikos*, 12 (AFRILEX-reeks/series 12: 2002): 44–74.

Van der Veken, A. & De Schryver, G-M. 2003. Les langues africaines sur la Toile: étude des cas haoussa, somali, lingala et isiXhosa [The African languages on the Internet: case studies for Hausa, Somali, Lingala and isiXhosa]. *Cahiers du Rifal*, 23 (Thème: Le traitement informatique des langues africaines [Theme: The computational processing of the African languages]): 33–45.

© Van Schaik Publishers

1 This document is available in all eleven official South African languages from <http://www.gov.za/symbols/coatofarms.htm>.

2 This document is available in all eleven official South African languages from <http://www.anc.org.za/about/anc.html>.

3 Note that all keywords from Table 8.5, except for the subject concord and/or demonstrative **se** (which do not have translation equivalents in English), are used and underlined in this description. This clearly suggests that keywords pinpoint the "aboutness" of a document.

4 Attempting to find an isiZulu equivalent for the Sesotho sa Leboa subject concord and/or demonstrative **se** is not relevant, given the isiZulu conjunctive way of writing.

5 Calculated with *Pearson's Product Moment Correlation* formula; a value of 1 would correspond to a perfect positive linear relationship.

© Van Schaik Publishers