

Corpus applications for the African languages, with special reference to research, teaching, learning and software¹

DJ Prinsloo* and Gilles-Maurice de Schryver

Department of African Languages, University of Pretoria, Pretoria 0002, South Africa.

*Corresponding author, e-mail: prinsloo@postino.up.ac.za

Abstract: The point of departure of the present article is the realisation that more and more serious contemporary linguistic applications are based on electronic corpora. If African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should therefore become an absolute priority. In order to illustrate the feasibility of *corpus applications* for the African languages at present, the article first considers 'fundamental linguistic research' in the fields of phonetics and question particles. It is shown how that research was boosted as a result of the utilisation of corpora. In a second section 'language teaching and learning' is given due attention by means of the corpus-aided compilation of pronunciation guides and textbooks, and the teaching of morpho-syntactic and contrastive structures. Finally, in the field of 'language software', a series of first-generation spellcheckers based on corpora is reviewed. All applications are exemplified with reference to one or more of the following African languages: Cilubà, Sepedi, isiXhosa, isiZulu, and Setswana.

Introduction

The present article focuses on various *corpus applications* in the broad field of linguistics, with special reference to the African languages. In a previous article we argued that '[c]ompiling and querying electronic corpora has become a *sine qua non* as an empirical basis for contemporary linguistic research. As a result, around the world, corpus applications now abound in all fields of linguistics' (De Schryver & Prinsloo, 2000:89). The *compilation* of African-language corpora and *corpus query tools* are the subjects of that previous article.

Present-day scholars are unanimous when it comes to the crucial role corpora play in the broad field of modern linguistic applications:

'As a consequence of the growing global interest in large electronic text corpora in the past few years, th[e present-day Dutch] corpus will be a component of a multifunctional collection of electronic texts, rather than used for lexicographical purposes only' (Kruyt, 1995:19)

'Carefully constructed, large written and spoken corpora are essential sources of linguistic knowledge if we hope to provide extensive and adequate descriptions of the concrete use

of the language in real text. These types of descriptions certainly remain impossible if we only rely on introspection and native speaker intuition' (Calzolari, 1996:4)

'It is now almost inconceivable that worthwhile and comprehensive lexical descriptions can be undertaken without a corpus' (Kennedy, 1998:91)

According to Kruyt, firstly: 'At the level of word form, [...] analysis of corpus data may be supported by statistical tools', secondly: 'The analysis at other language levels than word form requires a corpus encoded for linguistic features', and thirdly: 'Statistic devices can be applied on encoded linguistic features as well' (1995:126–127). As indicated in De Schryver and Prinsloo (2000:95–96), African-language corpus linguistics goes a long way with corpora clear of any codes (cf. also Hurskainen, 1998, § 2). The present article will therefore deal with corpus applications for the African languages on Kruyt's 'first level', which means that 'raw' corpora are used which have not been supplemented by a series of so-called 'standard corpus pre-processing' annotations. In contrast to the different levels suggested by Kruyt, Calzolari stresses the complexity of the

mutual interactions between lexicon and corpus:

'We can summarise, without claiming to be exhaustive, the lexicon (L) — corpus (C) interactions in the following list, where an arrow from L to C means, in general, the projection/mapping of some lexical data on the corpus, while an arrow from C to L means acquisition of lexical information from corpora.

- L → C tagging
- C → L frequencies (of different linguistic "objects")
- C → L proper nouns
- L → C parsing
- C → L updating
- C → L "collocational" data (MW, idioms, gram. patterns...)
- C → L "nuances" of meanings & semantic clustering
- C → L lexical (syntactic/semantic) knowledge acquisition
- L → C semantic tagging
- ↓
- C → L more semantic information on the lexical entry
- L → C semantic disambiguation
- C → L corpus based computational lexicography
- C → L validation of lexical models' (Calzolari, 1996:7–8)

As we are operating on Kruyt's first level, we see that, within Calzolari's framework, we are essentially dealing with 'acquisition of lexical information from corpora' (C → L). In future, when corpora for African languages will also be pre-processed linguistically, 'projection/mapping of some lexical data on the corpus' (L → C) will also become possible. At that point, we will be able to implement Calzolari's 'bootstrapping methodology' (1996:14), which implies that a continuous projection/mapping of C on L, and vice versa, results in successive analyses of the corpus which increase in richness.

Corpora uses in the broad field of linguistics are virtually unlimited, and are even found outside linguistics (such as in anthropology, history, sociology, etc., cf. Hurskainen, 1998, §2). Within linguistics, Calzolari (1996:4–5) mentions NLP (Natural Language Processing) and Speech systems, the evaluation of syntactic theories, a variety of phenomena occurring in

'real' texts (such as underestimated/underdiscussed structures, linguistically uninteresting phenomena, and deviations from linguistic models), the construction of stochastic models, the identification and characterisation of sub-languages, language teaching and learning, literary surveys, sociolinguistic considerations, lexicographic compilations, stylistic studies, etc.

Due to space restrictions, we can obviously only discuss a fraction of all potential linguistic applications of corpora for African languages. As such, the present article will touch upon three different linguistic facets: (a) fundamental linguistic research, (b) language teaching and learning, and (c) language software. The first facet (fundamental linguistic research) is exemplified by means of a thorough discussion of Cilubà phonetics on the one hand, and an overview of question particles in Sepedi on the other hand. The second facet (language teaching and learning) is illustrated with compilations of a pronunciation guide for Cilubà and a textbook for Sepedi on the one hand, and the teaching of morpho-syntactic and contrastive structures in Sepedi on the other hand. Finally, for the third facet (language software) the first-generation spellcheckers developed for isiXhosa, isiZulu, Sepedi and Setswana are reviewed.

Corpus applications in the field of fundamental linguistic research, Part 1: phonetics²

Formulation of the basic aim: 'corpus-based phonetics from below'

Traditionally, phonetic research has been undertaken only in order to proceed to phonology. More recently, phonetic research has been carried out in order to frame the results in a global perspective by cross-comparing occurrence frequencies of different phone categories in various languages. The utilisation of corpora in the field of phonetics, however, opens new and even more exciting doors. We will illustrate this with reference to some phonetic aspects of Cilubà.

If we look at phonetic studies that have been undertaken throughout the world, we see that the great majority of them are based on a 'translation' of a story, very often an English one at that, or worse, a 'translation' of randomly sampled (English) words. In order to avoid an

ethnocentric approach, different so-called ‘every-day, non-cultural lists of words’ have been assembled over the years (cf. Swadesh, 1952, 1953:349, 1955). Even though the label for such (English) lists was later changed to ‘basic vocabulary’ (cf. Bastin, Coupez & De Halleux, 1983:174), one will always recognise a ‘foreign bias’ for as long as one does not take the language being studied as one’s point of departure. The minority of scholars who did take the language itself as their point of departure would randomly sample a small selection of recordings and/or texts from that language from which to work. Here, of course, one strongly doubts the representativeness of such random samples.

In order to pursue truly modern phonetics one should therefore *do away with the ethnocentric approach* on the one hand, and *eliminate the random factor* on the other hand. Formulated differently, one needs to arrive at a phonetic description which emanates solely from the language itself — hence a ‘phonetics from below’-approach; and this must be a description with well-founded claims. To comply with both these points of departure at the same time, we argue that one can simply turn to top-

frequency counts derived from a corpus of the language under study. The amazing thing is that such a corpus does not even need to be large, while the actual words one works on can moreover be ridiculously small. The methodology we suggest is therefore a ‘*corpus-based phonetics from below*’-approach.

Previous ‘traditional’ scholarship in the field of Cilubà phonetics

As far as previous ‘traditional’ scholarship is concerned, only three publications explicitly discuss phonetic aspects of Cilubà. The first attempt is the one by Gabriël and dates from the 1920s. His phone inventory (originally a running text) has been summarised in Table 1.

As can be seen from Table 1, Gabriël does not use any phonetic symbols. He rather describes the “sounds” and/or uses the roman alphabet patterned on French and Flemish. In addition, one huge omission in his description concerns the tonal dimension of Cilubà.

The first scholar to stress the crucial role of tone in Cilubà is Burssens in his *Tonologische schets van het Tshiluba* (1939). His phone inventory (originally also a running text) has been summarised in Table 2.

Table 1: Cilubà phone inventory according to Gabriël (s.d.⁴ [(1921³):7–11])

ingressive sounds						
<ul style="list-style-type: none"> • monosyllabic affirmation en! • dental click 						
egressive sounds						
vowels		consonants		combinations		
a e i o u • they can be short, medium, or long • they can be nasalised		voiced	voiceless	vowels	consonants	
		soft b d j v z	hard p f t k s sh tsh	V+V	nasal+C	C+semivowel
		semivowel w y	• p = voiceless bilabial affricate	ai ei au eu io iu ...	• mb mp mf mv mm	• bw fw kw lw nw pw sw tw vw ...
		trill l	• mp = explosive		• nd ng nk ns nz nj nsh ntsh nt nn nw ny	• by dy ky my ny py ...
		nasal n m velar-n			• preceding vowels are slightly nasalised	

Table 3: Cilubà phone inventory according to Muyunga (1979:48–49, 52)

<i>Simple Consonants</i>	bilabial	labio-dental	dental-alveolar	palato-alveolar	palatal	velar	glottal
plosive	p b		t d			k	
nasal	m		n		ɲ	ŋ	
lateral			l				
fricative	ɸ	f v	s z	ʃ ʒ			h
affricate				tʃ			
semivowel	w				j		
<i>Prenasalised Consonants</i>	bilabial	labio-dental	dental-alveolar	palato-alveolar	palatal	velar	glottal
plosive	mp mb		nt nd				
fricative		mf mv	ns nz	ɲʃ ɲʒ		ŋk	
affricate				ntʃ			
<i>Simple Vowels</i>	front	back		<i>Diphthongs (We are puzzled by Muyunga's notation of diphthongd.)</i>		front	back
close	i	u		close	ɨi	ɨu	
half-close	e	o		close to half-close	ie ue io uo		
open	a			close to open	ia ua		
Remarks:	<ul style="list-style-type: none"> vowels can be short, environmentally lengthened or inherently long p is always prenasalised 						

is not explicit in this regard, the 2 333 words are *not* 2 333 *different* words. Anyhow, by randomly choosing 11 short texts he hopes to arrive at a representative sample of the Lubà language.

Corpus-based fieldwork and the Cilubà Phonetic Database (CPD)

While Muyunga's method can be regarded as a 'phonetics from below'-approach, one still needs to eliminate the random factor. It is precisely here that our suggestion to utilise a corpus comes in. To that end *Recall's Cilubà*

Corpus (RCC), a small-size, structured corpus of just 300 000 running words (tokens) was queried (cf. De Schryver & Prinsloo, 2000:98–102), and a corpus of that size turned out to contain approximately 35 000 different words (types). Now, the top ONE PERCENT of the types with an even distribution across the different sub-corpora (cf. De Schryver & Prinsloo, *forthcoming*), or thus just 350 words, not only turned out to provide enough data for a thorough phonetic analysis, but also to complement all existing phonetic descriptions. In other words, although this study only deals with

the top one percent of the types in RCC, the results are far-reaching. Indeed, all claims about the frequencies of occurrence of certain phones imply that these claims are valid for those words that are most frequent in Cilubà.

Fieldwork was carried out with a male native speaker of standard Cilubà. For each of the 350 words, he was asked to pronounce a short sentence chosen from the concordance lines extracted from RCC. After repeating this sentence a second time, the word was pronounced two more times in isolation. With this procedure we hoped to obtain a pronunciation as close to natural spoken language as possible. During the recordings, an initial transcription was made. In order to complete the purely auditory and visual cues, the informant was often asked to describe — in his own words —

the articulation of this or that phone. In addition, we read out our own transcriptions time and again.

Following the fieldwork, our initial transcriptions were verified with the recordings. Samples of the resulting (detailed) transcriptions are shown in Table 4.

The phonetic transcriptions of the 350 most frequent Lubà words constitute the backbone of the statistical database which was subsequently set up — the *Cilubà Phonetic Database (CPD)*. In total, CPD contains 1 709 phones. Each phone's phonetic description was coded in various ways to enable a thorough distributional analysis. An overview of the different phones attested in CPD is shown in Table 5.

Compared to the inventories presented in Tables 1, 2 and 3, the CPD inventory reveals a

Table 4: Samples of the phonetic transcriptions of the 350 most frequent words in Cilubà

#	phonetic transcription	#	phonetic transcription	#	phonetic transcription
153	b ^h úkɔ̀lè	180	wè:bè	207	kúfétá
154	ɲɔ̀:ndɔ̀	181	kú:ɲimà	208	múfɪ:ɲá
155	djè:ndè	182	kúvwá	209	mát ^{wh} ù:ɲá
156	wá:mbá	183	lú:ɲèɲi	210	b ^h úkwà
157	wá:mbá	184	mwá:kù	211	djè:t ^{wh} ù
158	kúmɔ̀ɲá	185	tá:	212	ɲbùzɪ
159	bɔ̀:bɔ̀	186	á:bɔ̀	213	kálé
160	dít ^{wh} úkú	187	kwá:ɲátá	214	ɲvùlá
161	bjá:mzá	188	tʃɪ:t ^{wh} ú	215	máɲɪ
162	kwá:mbá	189	kwí:kálá	216	fáɲɪ:fé
163	á:bá	190	míkà:ndà	217	kê:á
164	kwé:lá	191	tʃɪt ^{wh} ùfà	218	b ^h úkálé:ɲá
165	bùdʃi	192	mê:sú	219	t ^w wá
166	múswè	193	kú:lú	220	disà:ɲká
167	tʃè:má	194	mú:ɲkátʃɪ	221	ná:bjò

Table 5: Cilubà phone inventory derived from the *Cilubà Phonetic Database (CPD)*

CONSONANTS

	bilabial	labiodental	alveolar	palato-alveolar	palatal	velar
oral stop	p b		t d			k
nasal stop	m		n		ɲ	ŋ
trilled stop			(r)			
fricative	ɸ	f v	s z	ʃ ʒ		
resonant					j	
lateral resonant			l			

VOCALIC RESONANTS

	front	central	back
	(j) i		u
-mid	e		o
-mid	ɛ	(ə)	ɔ
	a		

OTHER SYMBOLS

- w** voiced labial-velar consonantal resonant
ʃ voiceless palato-alveolar affricate

number of striking differences, such as the presence of the voiced alveolar trilled stop [r] and the high number of vocalic resonants. The voiced alveolar trilled stop [r], for instance, is a phone not to be found in genuine ‘standard Cilubà’, so phoneticians have tended to overlook its importance. From a frequency point of view however, it is clear that this phone rightfully deserves its place on phonetic charts of Cilubà. Yet, in order for such and similar claims to be valid, one must be sure that there is a good correlation between the overall distribution of the phones mentioned in the literature and those in CPD.

Comparison between phone frequencies in the literature and those in CPD

On a first level, a comparison can be made

between the phone frequencies found in Muyunga and those derived from CPD. The results of this comparison are summarised in Table 6.

From Table 6 it is clear that there is excellent agreement between the two frequency studies. There are only a few discrepancies, and even these can be explained. As far as the consonants are concerned, there is just one phone for which there is a big difference between the studies, namely the voiced labial-velar consonantal resonant [w], for which Muyunga shows 1.04% while CPD has 5.21%. To a much smaller extent, an analogous difference can be observed for the voiced palatal consonantal resonant [j], for which Muyunga shows 1.36% while CPD has 2.52%. The reason for this is obvious once one realises that Muyunga includes diphthongs into his frequen-

Table 6: Cilubà phone frequencies in Muyunga (1979:58, 62–63) compared to those in CPD

CONSONANTS			VOCALIC RESONANTS			
symbol	Muyunga-%	CPD-%	symbol	Muyunga-%	CPD-%	
p	0.31	0.35	i	8.47	9.28	7.20
b	6.46	6.61	i(:)	0.81		
t	3.77	2.22	i:		0.39	1.17
d	5.29	5.15	e	4.03	4.83	3.86
k	5.25	5.68	e(:)	0.80		
m	7.28	7.25	e:		0.88	3.80
n	7.51	6.55	a	11.54	12.90	12.05
ɕ	0.63	0.59	a(:)	1.36		
ŋ	2.34	1.29	a:		0.48	4.80
(r)	—	0.12	u	10.62	11.48	9.13
ɛ	1.24	0.88	u(:)	0.86		
f	0.43	0.23	u:		0.22	1.11
v	0.81	1.23	o	1.46	1.92	2.05
s	1.38	1.40	o(:)	0.46		
z	0.48	0.70	o:		0.09	0.82
ʃ	0.98	0.82	(j)		—	0.06
ʒ	0.73	0.53	(ə)		—	0.06
j	1.36	2.52	DIPHTHONGS			
l	4.48	3.63	length	Muyunga-%	CPD-%	
w	1.04	5.21	short	2.41	—	
tʃ	0.76	0.94	long	2.59	—	
Total	52.53	53.90	Total	47.47	46.11	

cy counts. As a result, CPD's [**wa**], [**we**] and [**ja**] for instance, are considered [**ua**], [**ue**] and [**ia**] respectively by Muyunga. The 5.00% (= 2.41 + 2.59) diphthongs Muyunga counts roughly correspond to the 5.33% (= (5.21 - 1.04) + (2.52 - 1.36)) more [**w**] and [**j**] in CPD. To be able to compare the vocalic resonants, CPD's [**ε**] was added to [**e**], and CPD's [**ə**] was added to [**o**]. Also, as Muyunga distinguishes between 'short, environmentally lengthened and inherently long vowels' (cf. Table 3) while CPD is based on 'words in isolation' (thus excluding environmentally lengthened vocalic resonants), Muyunga's short and environmentally lengthened vocalic resonants had to be counted together in order to compare the two studies. As far as the short vocalic resonants are concerned, they agree rather well. For the long vocalic resonants, however, [**a:**] (0.48% *versus* 4.80%) and [**ε:**] (0.88% *versus* 3.80%) seem too incongruous. Upon consulting our transcriptions, we noted that the majority of [**a:**] and [**ε:**] come from demonstratives. Yet, for this part of speech, Muyunga (1979: 150–152) consistently (and wrongly) writes *short* vocalic resonants.

As far as the phones in brackets in Table 6 are concerned, we can note that, besides being attested solely in CPD, they are extremely infrequent. They therefore do not distort the inventory.

In order to calculate the correlation coefficient *r* between the two frequency studies, it is clear from the foregoing that counts for vocalic resonants, and for [**w**], [**j**] and diphthongs cannot be included. For the remaining phones one obtains a near-perfect correlation, as *r* = 0.98. On the whole, we must conclude that the proportional distribution of the phones in the small-scale CPD (1 709 phones) corresponds to the distribution found in Muyunga, which is as much as six times larger (10 726 phones). Doubtless, this clearly supports a *corpus-based* phonetics from below approach.

On a second level, the proportional occurrence of the different tones in vocalic resonants can also be considered. As far as number of words is concerned, the largest study was undertaken by Kabuta, as he transcribed one and a half hour of unscripted conversation and concluded that '[c]ounts carried out on a 90-minute ordinary conversation recorded on cas-

sette revealed [...] that there are 62% of H [high tones] vs. 38% L [low tones]' (1998b:57). The detailed analysis stored in CPD attests 61.04% high and 35.28% low tones (together with 3.30% falling, 0.13% rising, 0.13% middle and 0.13% voiceless). The fact that the tonal dimension in just 350 top-frequency words corresponds extremely well with the tonal dimension in a one-and-a-half-hour-long natural conversation once more clearly supports a *corpus-based* phonetics from below approach.

Complementing existing phone inventories for Cilubà

Accepting the validity of a corpus-based approach instantly implies that one must also seriously consider the peripheral phenomena attested by means of such an approach. Thus, phones like the voiced alveolar trilled stop [**r**] and the vocalic resonant schwa [**ə**], hence phones that do not belong to genuine 'standard Cilubà', should nonetheless be mentioned on future phonetic charts of Cilubà — precisely because they too presently belong to the frequent phones of the language.

Surprisingly enough, one word [**s̥i**] (a particle used to confirm a statement and for which 'isn't it?' might be a close equivalent) contained a phone never mentioned in the literature so far. The fact that the vocalic resonant [**i**] showed up as voiceless in the particle [**s̥i**] was really surprising to both the researchers as well as to the informant. This very particle was recorded very often, and in many different contexts. At times the informant even forced himself to make it voiced — as for one reason or another it was *thought* that this was the way it had to be pronounced — but in the end the informant was bound to conclude about the voiced attempt: "No! People do not speak like that." The voiceless vocalic resonant [**i̥**] should therefore also be mentioned on future phonetic charts of Cilubà.

Framing Cilubà phonetics in a global perspective

Once one realises that a minimum number of words representing the most frequent section of a language's lexicon are sufficient as a basis for a phonetic description, one can easily take existing research one step further and frame the results in a global perspective. The largest database for which systematic data is readily

available is *UPSID* (an acronym for *UCLA Phonological Segment Inventory Database*). This database was compiled under Maddieson's supervision at the University of California, Los Angeles, and contains the phonemic inventories of 317 languages (Maddieson, 1984). By way of example we can consider the distribution of the different places of articulation in stops for Cilubà, as shown in Figure 1.

From Figure 1 it can be seen that the most frequent places of articulation in stops for Cilubà are located forward in the oral cavity, viz. bilabial and alveolar, which together account for roughly four fifths of the places. The velar place of articulation roughly accounts for the remaining fifth. The UPSID database has: 23.50% labial, 0.13% labiodental, 33.48% dental-alveolar, 7.00% postalveolar, 5.09% retroflex, 5.70% palatal, 19.63% velar, 2.01% uvular and 3.46% glottal. Hence, one must conclude that here the Cilubà distribution broadly follows the general pattern seen in the world's languages.

On the other hand, exactly three quarters of the Cilubà stops are voiced, the remaining quarter being voiceless. The UPSID database has: 52.49% voiced *versus* 47.51% voiceless. Hence Cilubà here does not follow the general pattern seen in the world's languages.

Towards a sound treatment of the Cilubà vocalic resonants

The study of CPD also reveals the lackadaisical approach of any phonetic description of Cilubà thus far when it comes to the vocalic resonants. Firstly, through a purely auditive comparison with the taped pronunciation of the Cardinal Vowels (CVs) by Daniel Jones himself, we have come to the conclusion that a total of nine vocalic-resonant values are attested in CPD, cf. Table 7.

Nine vocalic-resonant values is a high number for a language traditionally considered as having *only five* vocalic resonants. In the entire literature in our possession, only three authors mention the existence of more than five vocalic resonants. Stappers (1949:xi) devotes just one sentence to the observation that there is no phonological opposition between *o* and *ɔ*, and *e* and *ɛ* in Cilubà. Kabuta (1998a:14) devotes only one short, obscure rule in which he argues that /e/ is pronounced [ɛ] whenever the preceding syllable contains /e/ or /o/. He gives only two examples: [kupɔnɛʃa] and [kukɛmɛʃa], which are not really helpful.⁵ In addition, one is at a total loss when it comes to the phones [o] *versus* [ɔ], for nothing is mentioned about them. The only serious attempt to clarify the matter is found in Muyunga (1979:49–51).

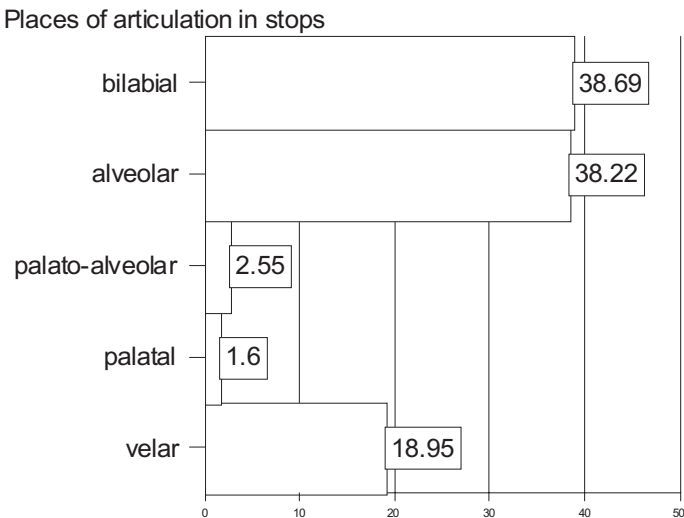


Figure 1: Proportional occurrence of each place of articulation in stops for Cilubà

Table 7: Vocalic resonants attested in CPD

[i]	CV1	[ɛ]	CV3	[u]	CV8
[e]	CV2	[ɚ]	CV4 somewhat retracted	[ɔ]	CV7 somewhat lowered
[ɘ]	CV2 somewhat lowered	(ə)	(IPA symbol for schwa)	[ɤ]	CV6 somewhat raised

His study brings him to the conclusion that ‘[t]he degree of openness of these vowels [e/ and /o/] is conditioned by the final vowel of the word’ (1979:49), a phenomenon he calls ‘a kind of retrogressive vowel harmony’. Unfortunately, upon scrutinising CPD, this suggested harmony cannot be supported. This has an important consequence. Even though Stappers’ observation still holds, the occurrence of a particular vocalic-resonant value not being predictable in a specific environment, one should seriously reconsider the many different orthographies used for Cilubà, for they are all restricted to just five ‘vowel symbols’.

Secondly, even more lackadaisical throughout the literature is the treatment of the tonal dimension of vocalic resonants, and this despite the fact that tones are used to make both semantic and grammatical distinctions. We are convinced that, if one is to expound on the real nature of the vocalic resonants in Cilubà, one needs a three-dimensional approach with a quantity level, a tonality level and a frequency level — and this for each vocalic resonant.⁶ As an illustration, two such three-dimensional approaches are shown in Figures 2 and 3.

Rare phenomena and the corpus-based phonetics from below approach

We must note that a method based on top-frequencies of occurrence will not — by definition! — show the rather rare phenomena of a language. In this respect, Gabriël is the *only* author to mention the presence of two ingressive phones, namely the ‘monosyllabic affirmation en!’ and the ‘dental click’ (cf. Table 1). These facets are certainly crucial if one pursues an *exhaustive* phonetic description of Cilubà.

Rather accidentally, what Gabriël calls the

‘monosyllabic affirmation en!’ was recorded during the sessions with the informant. Indeed, in one of the utterances to illustrate [s̺] (the particle used to confirm a statement) the informant starts off with a phone we could, tentatively, pinpoint as [ɸ̺], a breathy voiced glottal fricative pronounced on an indrawn breath. As it stands there, the ‘confirmation particle’ [s̺] is preceded by the ‘affirmation particle’ [ɸ̺]. It is however not simply a pleonasm to strengthen the ensuing statement even more. Rather, [ɸ̺] seems to be a paralinguistic use of the pulmonic ingressive airstream mechanism in order to express sympathy.

The Balubà rarely swear but whenever they do, they use [ʔ], the voiceless dental click. Just as [ɸ̺] (made on a pulmonic ingressive airstream), [ʔ] (made on a velaric ingressive airstream) is only used in a paralinguistic function.

The corpus-based phonetics from below approach as a powerful tool

To summarise this section on corpus applications in the field of fundamental phonetic research, one can safely claim that a ‘*corpus-based phonetics from below*’-approach is a powerful tool. Specifically for Cilubà, it has revealed previously underestimated phones, led to the discovery of one new phone, enabled framing the phonetic inventory in a global perspective, and pointed out some serious lacunae in the literature. For any language one can claim that this approach entails a new methodology in terms of which the phonetic description of a language is obtained in which one starts from the language itself and eliminates the random factor. In addition, this methodology makes it possible to make a maximum number of distributional claims, based on a minimum number of words, about the most frequent section of a language’s lexicon.

Tones for the vocalic resonant [ε]

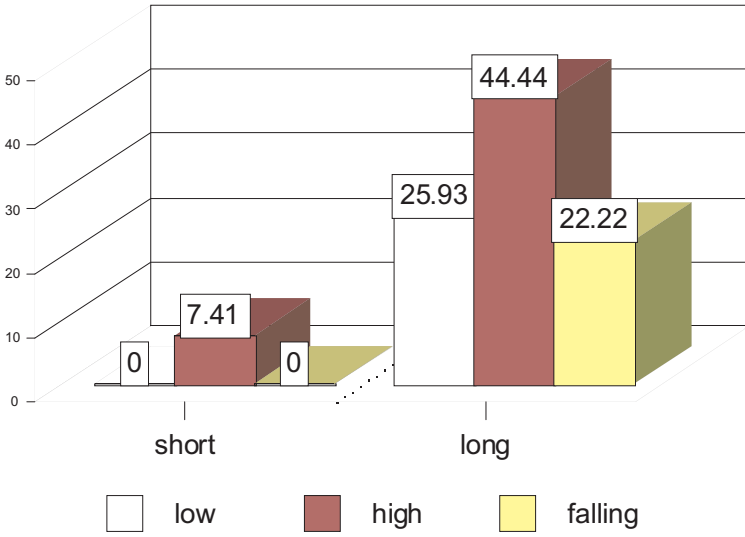


Figure 2: Three-dimensional approach to the vocalic resonant [ε] for Cilubà

Tones for the vocalic resonant [a]

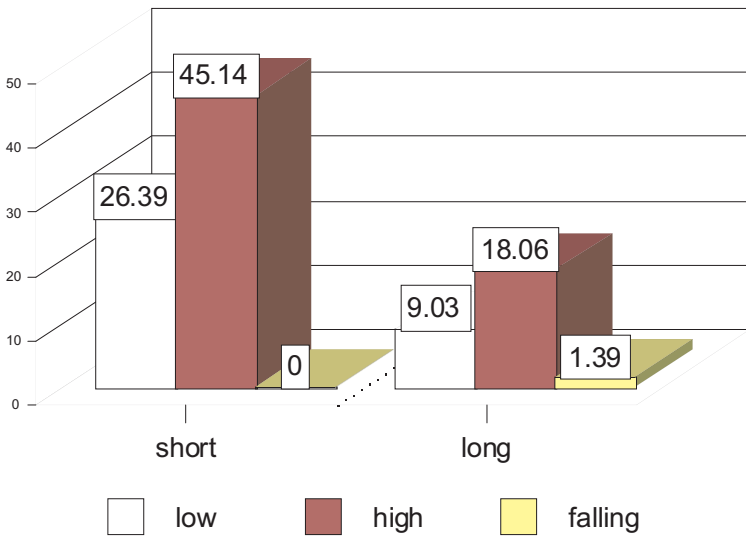


Figure 3: Three-dimensional approach to the vocalic resonant [a] for Cilubà

Corpus applications in the field of fundamental linguistic research, Part 2: question particles
Question particles in Sepedi: introspection-based and informant-based approaches

As a second example of how the corpus can revolutionise fundamental linguistic research into African languages, more specifically for the interpretation and description of problematic linguistic issues, we can look at how the corpus adds a new dimension to the traditional *introspection-based* and *informant-based* approaches. In these approaches a researcher had to rely on his/her own native speaker intuition or, as a non-mother tongue speaker, on the opinions of one (or more) mother tongue speaker(s) of the language. If conclusions which were made by means of introspection or in utilising informants are reviewed against corpus-query results, quite a number of these conclusions can be *confirmed* whilst others, however, are proven *incorrect*.

Prinsloo (1985), for example, made an in-depth study of the interrogative particles *na* and *afa* in Sepedi, in which he analysed the different types of questions marked by these particles. He concluded that *na* is used to ask questions of which the speaker does not know the answer, while *afa* is used if the speaker is of the opinion that the addressee knows the answer. Compare (1) and (2) respectively (adapted from Prinsloo, 1985:93).

(1) **Na** o tseba go beša nama? 'Do you know how to roast meat?'

(2) **Afa** o tseba go beša nama? 'Do you know how to roast meat?'

In terms of Prinsloo (1985), the first question will be asked if the speaker does not know whether or not the addressee is capable of roasting meat, and the second if the speaker is under the impression that the addressee is capable of roasting meat but observes that he/she is not performing well. Louwrens (1991:140), in turn, states that the use of *na* demands an answer, but that the use of *afa* indicates a rhetorical question.

Both Prinsloo (1985:93) and Louwrens (1991:143) emphasise that *afa* cannot be used with question words, and give the examples shown in (3) – (4), and (5) – (6) respectively.

(3) ***Afa** go hwile mang?

(4) ***Afa** ke mang?

(5) ***Afa** o ya kae?

(6) ***Afa** ke ngwana ofe yô a llago?

From (3) – (6) it is clear that according to Prinsloo and Louwrens the occurrence of *afa* with question words such as *mang*, *kae*, *-fe*, etc. is not possible in Sepedi.

Furthermore, they agree that *afa* cannot be used in sentence-final position:

'Sekere vraagpartikels tree [... s]legs in die inisiële sinsposisie [op]:

(3ii) * O tšwa ka gae ge o etla fa ka gore o šetše o fela pelo **afa**?' (Prinsloo, 1985:91)

'the particle *na* may appear in either the initial or the final sentence position, or in both these positions simultaneously, whereas *afa* may appear in the initial sentence position only' (Louwrens, 1991:140)

Thus Prinsloo's and Louwrens' presentation of the data suggests that: (a) *na* and *afa* mark different types of questions, (b) *afa* will not occur with question words such as *mang*, *kae*, *-fe*, etc., and (c) *afa* cannot be used in the sentence-final position.

Question particles in Sepedi: corpus-based approach

Querying the large, structured *Pretoria Sepedi Corpus (PSC)* when it stood at 4 million running words, *confirms* the semantic analysis of Prinsloo and Louwrens in respect of (b). The fact that not a single example is found where *afa* occurs with question words such as *mang*, *kae*, *-fe*, etc. validates their finding regarding the interrogative character of *afa*.

As for (c), however, compare the example found in PSC and shown in (7) where *afa* is, contrary to Prinsloo's and Louwrens' claim, used in sentence-final position.

(7) **Mokgalabje wa mereba ge! Naa e ka ba kgomolekokoto ye e mo hlotšeng afa? E ka ba ...** 'The cheeky old man! Can it be something big, immense and strong that created him? It can be ...'

Here, we must conclude that the corpus indicates that the analysis of both Prinsloo and Louwrens was too rigid.

Finally, contrary to claim (a), numerous examples are found in the corpus of *na* in combination with *afa*, but only in the order *na afa* and not vice versa. At least Louwrens, in principle, suggests that '[*n*]a and *afa* may in certain

instances co-occur in the same question' (1991:144), yet the only example he gives shows the co-occurrence of *a* and *naa*. Louwrens gives no actual examples of *na* occurring with *afa*, especially not when these particles *follow one another directly*. Table 8 lists the concordance lines culled from PSC for the sequence of question particles *na afa*.

The lines listed in Table 8 provide the empirical basis for a challenging semantic/pragmatic analysis in terms of the theoretical assumptions (and rigid distinction between *na* and *afa* especially) made in Prinsloo (1985) and Louwrens (1991). As far as the relation between this empirical basis and the theoretical assumptions is concerned, one would be well-advised to take heed of Calzolari's suggestion:

'In fact corpus data cannot be used in a simplistic way. In order to become usable they must be analysed according to some theoretical hypothesis, that would model and structure what would be otherwise an unstructured

set of data. The best mixture of the empirical and theoretical approaches is the one in which the theoretical hypothesis is itself emerging from and is guided by successive analyses of the data, and is cyclically refined and adjusted to textual evidence' (Calzolari, 1996:9)

The corpus is indispensable in highlighting the co-occurrences of *na* and *afa*. No researcher would have persevered in reading the equivalent of 90 Sepedi literary works and magazines to find such empirical examples. In fact, he/she would probably have missed them anyway, being 'hidden' in 4 million words of running text.

To summarise this section, we see that the corpus comes in handy when pursuing fundamental linguistic research into African languages. When a corpus-based approach is contrasted with the so-called 'traditional approaches' of introspection and informant elic-

Table 8: Concordance lines for the sequence of question particles *na afa* in Sepedi

1	tša Dikgoneng. Ruri re paletšwe. (Letl. 47)	Na afa Kgoteledi o tla be a gomela gae a hweditše
2	16) dedio. Bjale gona bothata bo agetše.	Na afa Kgoteledi mohla a di kwa o tla di thabela?
3	ba go forolle! MOLOGADI: Sešane sa basadi!	Na afa o tloga o sa re tswa, Feba? Ke eng tše o di
4	molamo. Sa monkgwana se gona ke a go botša.	Na afa o kwele gore mmotong wa Lekokoto ga go mpša
5	ya tšewa ke badimo, ya ba gona ge e felela.	Na afa o sa gopola ka lepokisana la gagwe ka nako
6	mpušeletša matšatšing ale a bjana bja gago.	Na afa o a tseba ka mo o kilego wa re hlomola pelo
7	a tla ba a gopola Bohlapanonwana gaMashilo.	Na afa bosola bjo bja poso ...? Yeo ya ba potšišo
8	hwetša ba re ga a gona. MODUPI: Aowa, ge!	Na afa rena re tla o gotša wa tuka? MOLOGADI: Se
9	ka mabaka a mabedi. La pele e be e le gore	na afa yola monna wa gagwe o be a sa fo ithomela
10	lebeletše gomme ka moka ba gagabiša mahlo):	Na afa le a di kwa? Ruri re tlo inama sa re
11	ba a hlamula.) MELADI (o a hwenahwena):	Na afa o di kwele? NAPŠADI (o a mmatamela): Ke
12	boa mokatong. (Go kwala khwaere ya Kekwele)	Na afa le kwa bošaedi bjo bo dirwago ke khwaere ya
13	Aowa, fela ga re kgole le kgole. KEKWELE:	Na afa matšatši a le ke le bone Thellenyane?
14	le baki yela ya go aparwa mohla wa kgoro?	Na afa dikobo tšeo e be e se sutu? Sutu? Ee, sutu,
15	iša pelo kgole, e fo ba metlae. KOTENTSHO:	Na afa le ke le hlole mogwera wa rena bookelong?
16	a go loba ga morwa Letanka e be e le, "	Na afa ruri ke therešo? Tša ditsotsi tšona ga di
17	le boNadinadi le boMatonya. MODUPI:	Na afa o a bona gore o a itahlela? O re sentše
18	yoo wa gago. NTLABILE: Kehwile mogatšaka,	na afa o na le tlhologelo le lerato bjalo ka
19	se nnete! Gape go lebala ga go elwe mošate.	Na afa baisana bale ba ile ba bonagala lehono. MDI
20	tlogele tšeo tša go hlaletšana. (Setunyana)	Na afa o ile wa šogašoga taba yela le Mmakoma? MDI
21	"mo ke lego gona ge o ka mpona o ka sola".	Na afa e ka be e le Dio? Goba ke Lata? Aowa, monna
22	ba oretše wo o se nago muši. "Hei thaka,	na afa yola morwedi wa Lenkwang o ile wa mo
23	le mmagwe ba ka mo feleletša. THOMO:	Na afa o a lemoga gore motho yo ga se wa rena? O
24	be re hudua dijanaga tša rena kua tseleng?	Na afa o lemoga gore mathaka a thala a feta

itation, corpora reveal both correct and incorrect traditional findings.

Corpus applications in the field of language teaching and learning ***Compiling pronunciation guides***

The corpus-based approach to the phonetic description of a language's lexicon that was described above, has, in addition, a first important application in the field of language acquisition. For Cilubà, the described study instantly lead to the compilation of two concise corpus-based pronunciation guides, a *Phonetic frequency-lexicon Cilubà-English* consisting of 350 entries, and a converted *Phonetic frequency-lexicon English-Cilubà* (De Schryver, 1999:55–68, 69–87). Provided that the target users know the conventions of the International Phonetic Alphabet (IPA), these two pronunciation guides give them the possibility to 'retrieve', 'pronounce' and 'learn' — and hence to 'acquaint themselves with' — the 350 most frequent words from the Lubà language.

Compiling modern textbooks, syllabi, workbooks, manuals, etc.

Pronunciation guides are but one instance of the manifold contributions corpora can make to the field of language teaching and learning. In general, one can say that learners are able to master a target language faster if they are presented with the most frequently used words, collocations, grammatical structures and idioms in the target language — especially if the quoted material represents authentic (also called 'naturally-occurring' or 'real') language use. In this respect Renouf, reporting on the compilation of a 'lexical syllabus', writes:

'With the resources and expertise which were available to us at Cobuild, [... a]n approach which immediately suggested itself was to identify the words and uses of words which were most central to the language by virtue of their high rate of occurrence in our Corpus' (Renouf, 1987:169)

The consultation of corpora is therefore crucial in compiling modern textbooks, syllabi, workbooks, manuals, etc.

The compiler of a specific language course for scholars or students may decide, for example, that a basic or core vocabulary of say 1 000 words should be mastered. In the past the com-

piler had to select these 1 000 words on the basis of his/her intuition or through informant elicitation which was not really satisfactory since, on the one hand, many highly used words were accidentally left out, and, on the other hand, such a selection often included words of which the frequency of use was questionable. According to Renouf:

'There has also long been a need in language-teaching for a reliable set of criteria for the selection of lexis for teaching purposes. Generations of linguists have attempted to provide lists of 'useful' or 'important' words to this end, but these have fallen short in one way or another, largely because empirical evidence has not been sufficiently taken into account' (Renouf, 1987:68)

With frequency counts derived from a corpus at his/her disposal, the basic or core vocabulary can easily and accurately be isolated by the course compiler and presented in various useful ways to the scholar or student — for example by means of full sentences in language laboratory exercises. Compare Table 9, which is an extract from the first lesson in the *First Year's Sepedi Laboratory Textbook* used at the University of Pretoria and the Technikon Pretoria, reflecting the five most frequently used words in Sepedi in context.

Here, the corpus allows learners of Sepedi, from the first lesson onwards, to be provided with naturally-occurring text revolving around the language's basic or core vocabulary.

Teaching morpho-syntactic structures

It is well-known that African-language teachers have a hard time teaching morpho-syntactic structures and getting learners to master the required analysis and description. This task is much easier when authentic examples, taken from a variety of written and oral sources, are used rather than artificial ones made up by the teacher. This is especially applicable to cases where the teacher has to explain more advanced or complicated structures and will have difficulty in thinking up suitable examples. According to Kruyt, such structures were largely ignored in the past:

'Very large electronic text corpora [...] contain sentence and word usage information that was difficult to collect

Table 9: Extract from the *First Year's Sepedi Laboratory Textbook*

M: <i>gore</i> 'that, so that'	
M: <i>Ke nyaka gore o nthuše</i> 'I want you to help me'	S:
M: <i>bona</i> 'see'	
M: <i>Re bona tau</i> 'We see a lion'	S:
M: <i>bona</i> 'they/them'	
M: <i>Re thuša bona</i> 'We help them'	S:
M: <i>bego</i> 'which was busy'	
M: <i>Batho ba ba bego ba reka</i> 'The people who were busy buying'	S:
M: <i>ila</i> 'come, shall/will'	
M: <i>Tla mo!</i> 'Come here!'	S:

until recently, and consequently was largely ignored by linguists' (Kruyt, 1995:126)

As an illustration we can look at the rather complex and intricate situation in Sepedi where up to five *le*'s or up to four *ba*'s are used in a row. In Tables 10 and 11 a selection of concordance lines extracted from PSC is listed for both these instances.

The relation between grammatical function and meaning of the different *le*'s in Table 10 can, for example, be pointed out. In corpus line #1 the first *le* is a conjunctive particle, followed by the class 5 relative pronoun and the class 5 subject concord. The sequence in corpus line #8 is copulative verb stem, class 5 relative pronoun and class 5 prefix, while in corpus line #29 it is class 5 relative pronoun, 2nd person plural subject concord and class 5 object concord, etc.

As the concordance lines listed in Tables 10 and 11 are taken from the living language, they represent excellent material for morpho-syntactic analysis in the classroom situation, as well as workbook exercises, homework, etc. In retrieving such examples in abundance from the corpus, the teacher can focus on the daunting task of guiding the learner in distinguishing between the different *le*'s and *ba*'s, instead of trying to come up with such examples on the basis of intuition and/or through informant elicitation. In addition, in an educational system where it is expected from the learner to perform

a variety of exercises/tasks on his/her own, basing such exercises/tasks on 'real' language can only be welcomed.

Teaching contrasting structures

Singling out top-frequency words and top-frequency grammatical structures from a corpus should obviously receive most attention for language teaching and learning purposes. Conversely, rather *infrequent* and *rare* structures are often needed in order to be contrasted with the more common ones. For both these extremes, where one needs to be *selective* when it comes to frequent instances and *exhaustive* when it comes to infrequent ones, the corpus can successfully be queried. Renouf argues:

'we could seek help from the computer, which would accelerate the search for relevant data on each word, allow us to be selective or exhaustive in our investigation, and supplement our human observations with a variety of automatically retrieved information' (Renouf, 1987:169)

Formulated differently, in using a corpus certain related grammatical structures can easily be contrasted and studied, especially in those cases where the structures in question are rare and hard if not impossible to find by reading and marking. Following exhaustive corpus queries, these structures can be instantly

Table 10: Morpho-syntactic analysis of up to five /e's in a row in Sepedi

Legend:							
①	relative pronoun class 5	⑤	subject concord 2nd person plural				
②	copulative verb stem	⑥	object concord 2nd person plural				
③	conjunctive particle	⑦	subject concord class 5				
④	prefix class 5	⑧	object concord class 5				
1	Go ile gwa direga mola malokeišene a mantši a thewa, gwa agiwa...	le	le	le	...bitšwago Donsa. Lona le ile la thongwa ke ba "municipality" ka go aga		
		③	①	⑦			
8	Taba ye e tšwa go Morena; rena ga re kgone go go botša le lebe le ge e...	le	le	le	...botse. Rebeka šo, o a mmona! Mo tšee o sepele, e be mosadi wa morwa wa		
		②	①	④			
13	go tšwa ka sefero a ngaya sethokgwa se se bego se le mokgahlo ga lapa...	le	le	le	le	le	...latelago. O be a tseba gabotse gore se bego se le mokgahlo ga lapa...
		①	⑦	②	①	⑦	
16	go tlošana bodutu ga rena go tlamegile go fela, ga ešita le lona leeto...	le	le	le			...swereng, ge nako ya lona ya go fela e fihla, le swanetše go fela. Bjale,
		①	⑤	⑧			
18	yeo e lego gona ke ya gore a ka ba a bolailwe ke motho. Ga se fela lehu...	le	le	le			...golomago dimpa tša ba motse, e šetše e le a mmalwa. Mabakeng ohle ge
		①	⑦	⑥			
29	ke yena monna yola wa mohumi le bego le le ka gagwe maabane. Letsogo...	le	le	le			...bonago le golofetše le, e sa le le gobala mohlang woo." Banna ba
		①	⑤	⑧			
32	seo re ka se dirago. Ga se ka ka ka le bona letšatši la madi go swana...	le	le	le			...hlabago le! Bona mahlasedi a mahubedu a lona a tsotsometša dithaba
		③	①	⑦			

Table 11: Morpho-syntactic analysis of up to four /ba's in a row in Sepedi

Legend:						
①	relative pronoun class 2	③	copulative verb stem			
②	auxiliary verb stem	④	subject concord class 2			
		⑤	object concord class 2			
22	meloko ya bona, ba tlišitše dineo; e be e le bagolo bala ba meloko, balaodi...	ba	ba	ba	...badilwego. Ba tlišitše dineo tša bona pele ga Morena; e be e le dikoloi	
		①	①	④		
127	mediro ya bobona yeo ba bego ba sešo ba e phetha malapeng a bobona. Ba ile...	ba	ba	ba	...tlogela le tšona dijo tšeo ba bego ba dutše ba dija. A ešita le bao ba beg	
		④	②	④		
185	ya Modimo 6 Le le ba go dirišana le Modimo re Le eletša gore le se ke la...	ba	ba	ba	...amogetšego kgaugelo ya Modimo mme e se ke ya le hola selo. Gobane o re: Ke	
		③	①	④		
259	ba be ba topa tša fase, baeng bao bona ba ile ba ba amogela ka tše pedi,...	ba	ba	ba	ba	...bea fase ka a mabedi, ka gore lešago la moeng le bewa ke mongwotse gae. Baen
		④	②	④	⑤	
272	tle go ya go hwetša tšela di bego di di kokotela. BoPoromane le bona ga se...	ba	ba	ba	...hlwa ba laela motho. Sa bona e ile ya fo ba go tšwa ba tlemolla makaba a bo	
		④	②	④		
312	itlela go itiša le koma le legogwa, fela ka tsebo ya gagwe ya go tsoma...	ba	ba	ba	...mmea yo mongwe wa baditi ba go laola lesolo. O be a fela a re ka a mangwe ge	
		④	②	④		
317	etšega go mpolediša. Mola go bago bjalo le ditaba di emago ka mokgwa woo, ...	ba	ba	ba	...bea marumo fase. Ga se ba no a lahlela fase sesolo. Ke be ke thathankg	
		④	②	④		

indexed and studied in context, and contrasted with their more frequently used counterparts.

As an example we can consider two different locative strategies used in Sepedi: 'prefixing of *go*' versus 'suffixing of *-ng*'. Teachers often err in regarding these two strategies as mutually exclusive, especially in the case of human beings. Hence, they regard *go monna* 'at the man' and *go mosadi* 'at the woman' as the accepted forms, while not giving any attention to, or even rejecting, forms such as *monneng* and *mosading*. This is despite the fact that Louwrens attempts to point out the difference between them:

'There exists a clear semantic difference between the members of such pairs of examples: *kgôšing* has the general meaning 'the neighbourhood where the chief lives', whereas *go*

kgôši clearly implies 'to the particular chief in person' (Louwrens, 1991:121)

Although it is clear that prefixing *go* is by far the most frequently used strategy, some examples are found in PSC substantiating the use of the suffixal strategy. Even more important is the fact that these authentic examples clearly indicate that there is indeed a *semantic difference* between the two strategies. Compare the general meaning of *go* as 'at' with the specific meanings which can be retrieved from the corpus lines shown in Tables 12 and 13.

Louwrens' semantic distinction between these strategies goes a long way in pointing out the difference. However, once again, careful analysis of corpus data reveals semantic connotations other than those described by researchers who solely rely on introspection and informants. So, for example, the meaning

Table 12: Corpus lines for *monneng* (suffixing the locative *-ng* to 'a human being' in Sepedi)

1	e eja mabele tšhemong ya mosadi wa bobedi	monneng	wa gagwe. Yoo a rego ge a lla senku a
2	a napa a ineela, a tseba gore o fihlile	monneng	wa banna, yo a tlogo mo khutšiša maima a
3	gore ka nnete le nyaka thušo, nka go iša	monneng	e mongwe wa gešo yo ke tsebago gore yena
4	bose bja nama. Gantši kgomo ya mogoga	monneng	e ba kgomo ye a bego a e rata kudu gare
5	gagwe a ikgafa go sepela le nna go nkiša	monneng	yoo wa gabo. Ga se ba bantši ba ba ka
6	ke mosadi, gobane ke yena e a ntšhitsuwego	monneng	. Ka baka leo monna o tlo tlogela tatagwe
7	botšobana bja lekgarebe. Thupa ya tefo	monneng	e be e le bohloko go bona kgomo e etšwa
8	ka mosadi. Gobane boka mosadi ge a tšwile	monneng	,le monna o tšwelela ka mosadi. Mme
9	a tšwa mosading; ke mosadi e a tšwilego	monneng	. Le gona monna ga a bopelwa mosadi; ke

Table 13: Corpus lines for *mosading* (suffixing the locative *-ng* to 'a human being' in Sepedi)

1	seleka. (Setu.) Bjale ge a ka re o boela	mosading	wa gagwe ke reng? PEBETSE: Se tshwenyeye,
2	thuše selo ka gobane di swanetše go fihla	mosading	. A tirišano ye botse le go jabetšana
3	a yo apewa a jewe. Le rile go fihla	mosading	la re mosadi a thuše ka go gotša mollo le
4	o tlogetše mphufutšo wa letheka la gagwe	mosading	yoo e sego wa gagwe, etšwe a boditšwe
5	a nnoši lenyalong la rena. IKGETHELE:	Mosading	wa bobedi? INAMA: Ke mo hweditše a na le
6	a lahla setala, a sekamela kudu ka	mosading	yo monyenyanane, mererong gona o tla
7	lona leo. Ke maikutlo a bona a kgatelelo	mosading	." Taamane a sega, "Ke be ke sa tsebe seo
8	ka namane", re hwetša se na le kgononelo	mosading	gore ge a ka nyalwa e sa le lekgarebe go
9	tšhelete le botse bja gagwe mme a kgosela	mosading	o tee. O dula gona Meadowlands, Soweto.
10	dirwa ke gore ke bogale. Ke bogale kudu	mosading	wa go swana le Maria. Nna ke na le
11	ke wa ntira mošemanyana. O boletše maaka	mosading	wa gago gore ke mo hweditše a itia bola
12	. Gobane monna ge a bopša, ga a tšwa	mosading	; ke mosadi e a tšwilego monneng. Le gona
13	lethabo le tlhompho ya maleba go tšwa	mosading	wa gagwe. O tla be a intšhitše seriti ka
14	le ba bogweng bja gagwe gore o sa ya	mosading	kua Ditsobotla a bone polane yeo a ka e
15	ka dieta ...?" "O ile a ntlogela, a ya	mosading	yo mongwe. Ke yena mosadi yoo yo a

of phrases such as *Gobane monna ge a bopša, ga a tšwa mosading; ke mosadi e a tšwilego monneng* 'Because when man was created he did not come out of a woman, it is the woman who came out of the man' (cf. corpus line #9 in Table 12 and corpus line #12 in Table 13) in the Biblical sense, is not catered for.

Corpus applications in the field of language software: spellcheckers

According to the *Longman Dictionary of Contemporary English*, a spellchecker is 'a computer PROGRAM that checks what you have written and makes your spelling correct' (Summers, 1995³). Today, such language software is abundantly available for Indo-European languages. Yet corpus-based frequency studies may enable African languages to be provided with such tools as well.

Basically there are two main approaches to spellcheckers. On the one hand one can program software with a proper description of a language, including detailed morpho-phonological and syntactic rules, together with a stored list of word-roots; and on the other hand one can simply compare the spelling of typed words with a stored list of word-forms. The latter, indeed, forms the core of the *Concise Oxford Dictionary's* definition of a spellchecker: 'a computer program which checks the spelling of words in files of text, usually by comparison with a stored list of words' (1996).

While such a 'stored list of words' is often assembled in a random manner, we argue that much better results are obtained when the compilation of such a list is based on high frequencies of occurrence. Formulated differently, a first-generation spellchecker for African languages can simply compare typed words with a stored list of the top few thousand word-forms. Actually, this approach is already a reality for isiXhosa, isiZulu, Sepedi and Setswana, as first-generation spellcheckers compiled by DJ Prinsloo are commercially available in *WordPerfect 9* within the *WordPerfect Office 2000* suite. Due to the conjunctive orthography of isiXhosa and isiZulu the software is obviously less effective for these languages than for the disjunctively written Sepedi and Setswana.

To illustrate this latter point, tests were conducted on two randomly selected paragraphs.

In (8) the isiZulu paragraph is shown, where the word-forms in bold are not recognised by the *WordPerfect 9* spellchecker software.

(8) Spellchecking a randomly selected paragraph from *Bona Zulu* (June 2000:114)

Izingane ezizichamelayo zivame ukuhlala ngokuhlukumezeka kanti akufanele ziphathwe ngaleyondlela. Uma ushaya ingane ngoba izichamelile usuke uyihlukumeza ngoba lokho ayikwenzi ngamabomu njengoba iningi labazali licabanga kanjalo. Uma nawe mzali usubuyisa ingqondo, usho ukuthi ikhona ingane engajatsuliswa wukuvuka embhedeni obandayo omanzi njalo ekuseni?

The stored isiZulu list consists of the 33 526 most frequently used word-forms. As 12 out of 41 word-forms were not recognised in (8), this implies a success rate of 'only' 71%.

When we test the *WordPerfect 9* spellchecker software on a randomly selected Sepedi paragraph, however, the results are as shown in (9).

(9) Spellchecking a randomly selected paragraph from the telephone directory *Pretoria White Pages* (November 1999–2000:24)

Dikarata tša mogala di a hwetšagala ka go fapafapana goba R15, R20, (R2 ke mahala) R50, R100 goba R200. Gomme di ka šomišwa go megala ya Telkom ka moka (ye metala) Ge tšhelete ka moka e fedile karateng o ka tsentšha karata ye nngwe ntle le go šitiša poledišano ya gago mogaleng.

Even though the stored Sepedi list is smaller than the isiZulu one, as it only consists of the 27 020 most frequently used word-forms, with 2 unrecognised words out of 46, the success rate is as high as 96%.

The four available first-generation spellcheckers were tested by *Corel's Beta Partners* and the current success rates were approved. Yet it is our intention to substantially enlarge the sizes of *all* our corpora for South African languages, so as to feed the spellcheckers with, say, the top 100 000 word-forms. The actual success rates for the conjunctively written languages (isiNdebele, isiXhosa, isiZulu and siSwati) remains to be seen, while it is expected that the performance for the disjunctively written languages (Sepedi, Sesotho, Setswana, Tshivenda and Xitsonga) will be more than acceptable with such a corpus-based approach.

Conclusion

We have shown clearly that applications of electronic corpora in various linguistic fields have, at present, become a reality for the African languages. As such, African-language scholars can take their rightful place in the new millennium, and mirror the great contemporary endeavours in corpus linguistics achieved by scholars of, say, Indo-European languages.

In this article, together with a previous one (De Schryver & Prinsloo, 2000), the *compilation*, *querying* and possible *applications* of African-language corpora have been reviewed. In a way, these two articles should be considered as foundational to a discipline of corpus linguistics for the African languages — a discipline which will be explored more extensively in future publications.

From the different corpus-project applications that have been used as illustrations of the theoretical premises in the present article, we can draw the following conclusions:

- In the field of fundamental linguistic research we have seen that, in order to pursue truly modern *phonetics*, one can simply turn to top-frequency counts derived from a corpus of the language under study — hence a ‘corpus-based phonetics from below’-approach. Such an approach makes it possible to make a maximum number of distributional claims, based on a minimum number of words, about the most frequent section of a language’s lexicon.
- Also in the field of fundamental linguistic research, the discussion of *question particles* brought to light that when a corpus-based approach is contrasted with the so-called ‘traditional approaches’ of introspection and informant elicitation, corpora reveal both correct and incorrect traditional findings.
- When it comes to corpus applications in the field of *language teaching and learning*, we have stressed the power of corpus-based pronunciation guides and corpus-based textbooks, syllabi, workbooks, manuals, etc. In addition, we have illustrated how the teacher can retrieve a wealth of morpho-syntactic and contrasting structures from the corpus — structures he/she can then put to good use in the classroom situation.
- Finally, we have pointed out that at least one set of corpus-based language tools is already

commercially available. With the knowledge we have acquired in compiling the software for first-generation *spellcheckers* for four African languages, we are now ready to undertake the compilation of spellcheckers for *all* African languages spoken in South Africa.

Notes

¹This article is based on a paper read by the authors at the *First International Conference on Linguistics in Southern Africa*, held at the University of Cape Town, 12–14 January 2000. G-M de Schryver is currently Research Assistant of the Fund for Scientific Research — Flanders (Belgium).

²A different approach to the research presented in this section can be found in De Schryver (1999).

³Laver’s phonetic taxonomy (1994) is used as a theoretical framework throughout this section.

⁴Strangely enough, Muyunga seems to feel the need to combine the different phone inventories into one new inventory. In this respect, he distinguishes the voiceless bilabial fricative *and* the voiceless glottal fricative, claiming that ‘Each simple consonant represents a phoneme, except ϕ and h which belong to a same phoneme’ (1979: 47). Here however, Muyunga is mixing different dialects. While [ϕ] is used by, for instance, the Bakwà Diishò — their dialect giving rise to what is presently known as ‘standard Cilubà’ (De Clercq & Willems, 1960³:7) — the glottal variant [h] is used by, for instance, some Bakwà Kalonji (Stappers, 1949:xi). The glottal variant, not being the standard, is seldom found in the literature. A rare example is the dictionary by Morrison, Anderson, McElroy & McKee (1939).

⁵High tones being more frequent than low ones, Kabuta restricts the tonal diacritics to low tone, falling tone and rising tone. The first example should have been [$ku\phi\grave{o}ne\grave{s}a$].

⁶Considering tone (and quantity) as an integral part of vocalic-resonant identity does not seem far-fetched as long as ‘words in isolation’ are concerned. The implications of such an approach for ‘words in context’, however, definitely need further research.

References

(URLs last accessed on 16 April 2001)

- Bastin Y, Coupez A & De Halleux B.** 1983. Classification lexicostatistique des langues bantoues (214 relevés). *Bulletin des Séances de l'Académie Royale des Sciences d'Outre-Mer* 27(2): 173–199.
- Bona Zulu, Imagazini Yesizwe,** Durban, June 2000.
- Burssens A.** 1939. *Tonologische schets van het Tshiluba (Kasayi, Belgisch Kongo)*. Antwerp: De Sikkel.
- Calzolari N.** 1996. Lexicon and Corpus: a Multifaceted Interaction. In Gellerstam M *et al.* (eds) *Euralex '96 Proceedings I*. Gothenburg: Gothenburg University. pp 3–16.
- Concise Oxford Dictionary, Ninth Edition, On CD-ROM.** 1996. Oxford: Oxford University Press.
- De Clercq A & Willems E.** 1960³. *Dictionnaire Tshilubà-Français*. Léopoldville: Imprimerie de la Société Missionnaire de St. Paul.
- De Schryver G-M.** 1999. *Cilubà Phonetics, Proposals for a 'corpus-based phonetics from below'-approach*. Ghent: Recall.
- De Schryver G-M & Prinsloo DJ.** 2000. The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies* 18: 89–106.
- De Schryver G-M & Prinsloo DJ.** *forthcoming*. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. *South African Journal of African Languages* 21.
- Gabriël [Vermeersch].** s.d.⁴ [(1921³)] *Etude des langues congolaises bantoues avec applications au tshiluba*. Turnhout: Imprimerie de l'École Professionnelle St. Victor.
- Hurskainen A.** 1998. *Maximizing the (re)usability of language data*. Available at <<http://www.hd.uib.no/AcoHum/abs/hursk.htm>>.
- Kabuta NS.** 1998a. *Inleiding tot de structuur van het Cilubà*. Ghent: Recall.
- Kabuta NS.** 1998b. Loanwords in Cilubà. *Lexikos* 8: 37–64.
- Kennedy G.** 1998. *An Introduction to Corpus Linguistics*. London: Longman.
- Kruyt JG.** 1995. Technologies in Computerized Lexicography. *Lexikos* 5: 117–137.
- Laver J.** 1994. *Principles of Phonetics*. Cambridge: Cambridge University Press.
- Louwrens LJ.** 1991. *Aspects of Northern Sotho Grammar*. Pretoria: Via Afrika Limited.
- Maddieson I.** 1984. *Patterns of Sounds*. Cambridge: Cambridge University Press. See also <<http://www.linguistics.rdg.ac.uk/staff/Ron.Brasington/UPSID.interface/Interface.html>>.
- Morrison WM, Anderson VA, McElroy WF & McKee GT.** 1939. *Dictionary of the Tshiluba Language (Sometimes known as the Buluba-Lulua, or Luba-Lulua)*. Luebo: J. Leighton Wilson Press.
- Muyunga YK.** 1979. *Lingala and Cilubà Speech Audiometry*. Kinshasa: Presses Universitaires du Zaïre.
- Pretoria White Pages, North Sotho, English, Afrikaans Information Pages,** Johannesburg, November 1999–2000.
- Prinsloo DJ.** 1985. Semantiese analise van die vraagpartikels *na* en *afa* in Noord-Sotho. *South African Journal of African Languages* 5(3): 91–95.
- Renouf A.** 1987. Moving On. In Sinclair JM (ed.) *Looking Up, An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London: Collins ELT. pp 167–178.
- Stappers L.** 1949. *Tonologische bijdrage tot de studie van het werkwoord in het Tshiluba*. Brussels: Koninklijk Belgisch Koloniaal Instituut.
- Summers D** (director). 1995³. *Longman Dictionary of Contemporary English, Third Edition*. Harlow: Longman Dictionaries.
- Swadesh M.** 1952. Lexicostatistic Dating of Prehistoric Ethnic Contacts. *Proceedings of the American Philosophical Society* 96: 452–463.
- Swadesh M.** 1953. Archeological and Linguistic Chronology of Indo-European Groups. *American Anthropologist* 55: 349–352.
- Swadesh M.** 1955. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics* 21: 121–137.