Work on Spoken (Multimodal) Language Corpora in South Africa

Jens Allwood†, Harald Hammarström†, Andries Hendrikse*, Mtholeni N. Ngcobo*, Nozibele Nomdebevana*, Laurette Pretorius* and Mac van der Merwe*

† University of Gothenburg, Gothenburg
* University of South Africa, Pretoria
E-mail: jens@ling.gu.se, harald2@chalmers.se, HENDRAP@unisa.ac.za, Ngcobmn@unisa.ac.za, Nomden@unisa.ac.za, pretol@unisa.ac.za, Vdmertm@unisa.ac.za

Abstract

This paper describes past, ongoing and planned work on the collection and transcription of spoken language samples for all the South African official languages and as part of this the training of researchers in corpus linguistic research skills. More specifically the work has involved (and still involves) establishing an international corpus linguistic network linked to a network hub at a UNISA website and the development of research tools, a corpus research guide and workbook for multimodal communication and spoken language corpus research. As an example of the work we are doing and hope to do more of in the future, we present a small pilot study of the influence of English and Afrikaans on the 100 most frequent words in spoken Xhosa as this is evidenced in the corpus of spoken interaction we have gathered so far. Other planned work, besides work on spoken language phenomena, involves comparison of spoken and written language and work on communicative body movements (gestures) and their relation to speech.

1. Background

Linguistic corpora are very important resources for a language, and are increasingly seen as a requirement for maintaining language vitality in the face of global language competition (Allwood 2006, Berment 2004). In particular, multimodal and spoken language corpora are relatively unexplored, and complementary to written language corpora in several senses. For many languages, written language data, as opposed to spoken language data, are relatively easy to collect from the web (Scannell 2007, Streiter et al. 2006). For many languages, spoken language is the dominant modality, and is the only modality for certain genres, e.g. activities involving traditional knowledge.

With this background we decided to set up a collaborative research planning project for the adaptation and development of spoken language corpora software for the African languages of South Africa between the School of Computing and the Department of and Linguistics of The University of South Africa (UNISA). The project has so far achieved the following subgoals:

1. A collaborative research project on developing corpus research software pertinent to the South African indigenous languages between the Linguistics departments at Gothenburg University and UNISA together with the Computer Science department at UNISA has been started.
2. The significance of the spoken language corpus project for the development of the previously disadvantaged indigenous languages was noted by senior management at UNISA as a result of which a UNISA Strategic Project entitled "The UNISA Southern African Spoken and Signed Language Corpus (SASSLC)", was approved and is funded by UNISA for the period 2008-2010.

A third goal not yet achieved is

3. A possibly solar-energy driven e-learning project based on a so called (Wikipedia inspired) "collaborative platform" addressing issues such as participant involvement in creating and using the corpus as well as literacy and numeracy making use of corpus-based material.

2. Some work done so far

The work done so far is based on multimodal (audio-video) recordings of face-to-face communication in different social activities. So far material has been collected for the following languages: Xhosa, Zulu, S. Sotho, Ndebele, Khwedam and concerns the following activities:

Activities with ritualized parts

- Social gatherings, e.g. funerals, weddings, official and unofficial meetings
- Cultural activities, e.g. Thonjane (girls' initiation in S. Sotho); Umkhosi womhlanga (Ceremony of virginity testing in Zulu)
- Reliigious services, e.g. Umgoduso (graduation of traditional healers in Xhosa and Zulu; Bible discussions for Jehova witnesses; Charismatic church services)

Less ritualized activities

• Informal conversations (in bar, shops, by rivers, in family and community gatherings, parties, during school intervals)

The corpus today consists of Xhosa, Zulu and Khwedam spoken language recordings (mostly video, i.e. multimodal) that have been partially transcribed. Below, we present a table that gives the number of recordings and size of transcriptions per language.

Lang-uage	Khwe-dam started in 2007	Zulu Just started	Xhosa started in 2004	Total
Audio rec.	0	0	57	57
Video rec.	6	14	72	98*
All recc.	6	14	29	155*
Audio transcr	0	0	42	42
Video transcr	6	6	55	67
All transcr	6	6	97	109
Chec-ked and edited transcr	0	0	78	78
Total N. of words	6 000	25 000	319 113 Videos 48246 Audios	398 359

Table 1. Recordings and Transcription per language*:
*The corpus also includes 3 video recordings for Southern Sotho (started in 2007) and 3 video recordings for Ndebele (just started).

The transcription standard builds on two decades of experience in working on spoken language corpora for Swedish. On the basis of this, a special manual for the South African environment has been developed (Allwood et al. 2005). The transcription annotates contrastive stress, pauses, lengthening, overlaps, code mixing, code switching, translation to the corresponding written form etc. in a manner that allows computational harvesting.

Some pilot work on the Xhosa spoken corpus has been done, extracting information on word frequencies, feedback, code switching and language mixing. A CD with a compilation of corpus linguistic research training material in Power point has been developed and distributed to various institutions and centers participating in the project. The intention is that this CD should complement and be used with the Spoken Language Corpus Manual that was published in 2005.

An overview of corpus-related research in South Africa in the form of a collection of articles has been published in a thematic volume of Language Matters in 2007. A corpus website (which will serve as a hub for networking institutions participating in the spoken corpus project) has been set up. Its functioning has been tested in an experimental phase in order to iron out problems such as band-with, open-source access, passwords for outside users, etc.

Some San language recordings (Khwedam [xuu], a Khoe-Kwadi language) have been transcribed as part of the Project and currently a corpus for Khwedam is being compiled in Kimberly. The Xhosa corpus has been growing steadily. Reasonable progress has been made with the Southern Sotho corpus and the Zulu corpus for which we have now acquired part-time transcribers, while corpora for Northern Sotho, Venda and Tsonga are being compiled by researchers that we have trained at the relevant language centers at the Universities of Venda and Limpopo. A survey of available corpus tools and their suitability for the 'mining' of corpora of agglutinating languages has been conducted. This survey served as the starting point of a corpus tools development project, we are currently engaged in. A workshop on the problematic nature of words (the typical token units invoked in corpus searches) was conducted by Professor A P Hendrikse, in the Department of Linguistics, at Gothenburg University in September 2007. This workshop explicated a problem that has significant implications for cross-linguistic corpus studies and applications in areas such as speech therapy, child language development, basic vocabularies and word frequency studies. One of the interesting research issues that has emanated from the language corpus project is the status of the notion 'word' in cross-linguistic corpus research and applications (cf. Allwood, Hendrikse and Ahlsén forthcoming). In addition to this paper, work is being done on papers focusing on the problematic issues surrounding words in agglutinating languages and their implications for corpus applications in speech therapy, word-based diagnostics of language disorders and language development. Other work is being done on influences on spoken Xhosa from English and Afrikaans (see pilot study below) as well on gestures in Sotho.

3. Some uses of the corpus

Spoken language corpora can be used in language development, in the development of terminologies and translation data banks, the development of learning materials and the study of indigenous knowledge systems. Spoken language corpora should therefore have a long-term impact on minority languages, their empowerment and status planning. Hopefully, these corpora will also impact on the localisation and adaptation of electronic technologies to the South African indigenous languages.

Basic corpus statistics (words, collocations, MLU (mean length per utterance), vocabulary richness) can be extracted from the corpus without recourse to tools for morphology, syntactic parsing etc. The simplest kind of analysis, namely a word frequency table, already gives a certain insight into the spoken language (as opposed to the written language). Below we show the 20 most frequent words in a Xhosa corpus of 98 056 tokens.

Rank	Word	Freq.	Proportion (relative share)
1	**ke** (so)	1653	0.0168577139594
2	ukuba (if)	1543	0.0157359060129
3	ngoku (now)	990	0.0100962715183
4	m (m, FB)	968	0.00987190992902
5	hayi (no)	915	0.0093314024639
6	nto (thing	875	0.00892347230154
7	le (this)	750	0.00764869054418
8	apha (here)	738	0.00752631149547
9	e (yes)	736	0.00750591498735
10	**so** (so)	689	0.00702659704659
11	xa (when)	678	0.00691441625194
12	**and** (and)	626	0.00638410704087
13	nje (like)	621	0.00633311577058
14	into (thing)	601	0.0061291506894
15	abantu (people)	539	0.00549685893775
16	**ok** (ok)	505	0.00515011829975
17	kodwa (but)	463	0.00472179162927
18	ewe (yes)	403	0.00410989638574
19	**ja** (yes)	365	0.0037223627315
20	mna (me)	364	0.00371216447744

We do not have access to a corresponding written language corpus of comparable size and genre balance, but a few features of spoken language are still apparent. Feedback words such as *ke*, *e*, *ewe*, *m*, *ok, ja, hayi* belong to the most common words (cf. Allwood 1988).

As regards other categories in the frequency list we may make the following observations. It is interesting to note that in the top 10 items the majority is either adverbials or conjunctives. Interestingly enough, even if we look at lower frequencies than the 100 most frequent words there are very few verbs and nouns. Once again, this distortion of the facts is a function of the nature of agglutinating languages. The rich variability of Xhosa agreement and tense/aspect morphology simply obscures the frequency of verbal lexical tokens.

Perhaps the most valuable aspect of collecting authentic, unedited speech is that the data represent 'how people really speak', as opposed to language guided by normative principles. A striking fact is that the English-derived words *so* and *and* turn out to be the 10th and 12th most frequent words in Xhosa! It is one of several witnesses to the presence of English code switching among Xhosa speakers (cf. de Clerk 2006). A typical example runs as follows.

N: *Kanti bekusenzeka ntoni kuqala yintoni le intshintshileyo wena xa u(cinga)* (what was actually happening before.? What has changed according to you?)

T: *o: umbane* (oh electricity)

N: *o: umbane uyabona umbane ubungekho and then namanzi ndivile bathi bayagrumba you know bahambisa amanzi iintwezinjalo* (oh electricity you see electricity was not available and I have heard that they say they are trying to bring water as well, you know, they are bringing water, something like that)

N: *e: so ingathi noko ubomi bungcono* (Yes it seems as if life is better)

It is interesting that few code mixes are attested among the most frequent words. The reason is that in agglutinating languages such as the Bantu languages of South the rich morphological system (e.g. agreement) obscures the English lexical material on which the Xhosa grammar is imposed to such an extent that, although the mixed forms are pervasive they have a misleadingly low token frequency. Thus, we may have an English lexical item such as *wrong* that may occur extensively in the corpus, but because of the variability in the tokens caused by agreement morphology, may be statistically insignificant.
E.g. *urongo, barongo, irongo, zirongo, sirongo*, etc.

This means that the typical statistical analysis of word frequencies in corpus studies needs to be adapted for agglutinating morphological systems, for example.

South Africa is a multilingual country in two senses of the word. On the one hand, it has mother tongue speakers of a multitude of language of which 11 are constitutionally recognized as official languages (viz. Afrikaans, English, Ndebele, Northern Sotho, Southern Sotho, Swati, Tsonga, Tswana, Venda, Xhosa, Zulu). On the other hand, many speakers are conversant in more than two of the official languages. Against this background, it comes as no surprise that cross linguistic influences among these languages are pervasive. Of the languages mentioned above, English is currently the official lingua franca of the country and is the main medium of communication in most official transactions and interaction as well as the main medium of tuition in most secondary and tertiary institutions. Typical of such multilingual situations, the various languages exert some or other influence on each other (e.g. lexical, fixed expressions, discourse particles, feedback expressions and even grammatical structures). On account of the status of English as the lingua franca of the country, all language seem to be influenced more by English than by

any of the other languages.

In our spoken language corpus studies of the various official language of South Africa, we specifically annotated the transcribed recordings for three categories of influences from English on the other languages. The three categories are adoptives, code mixes and code switches. These three categories can be distinguished in the following way:

Adoptives: These are lexical items that over many years have become standardized and accepted as indigenous words that conform to all the relevant linguistic features of the adopting language. Thus although their language of origin may still be identifiable, they are no longer seen as foreign intrusions.
E.g. Xhosa isikolo (Eng: school), utitshala (Eng: teacher)
Sotho: : toropo (Afr: dorp 'town'), koranta (Afr: koerant ' newspaer')

Code mix: These are expressions in which a mixture of the grammar of one language (mostly one of the indigenous languages) and lexical material from another language (mainly English) is manifested. Such instances of code mixing are pervasive, particularly in the urban areas, but they are also attested in the rural areas, i.e. more conservative areas of a specific language. In contrast to adoptives, code mixes are not recognised as standard language by the official bodies (e.g. PanSALB – Pan South African Language Board.)
E.g. Xhosa: *urongo* (you are wrong), *barongo* (they are wrong), *uyafowuna* (he is phoning), *bayafowuna* (they are phoning)

Code switch: These are instance of complete and unaltered forms of another language (in particular, English) that are used in discourses between two speakers of the same language.
E.g. Xhosa discourse with English code switching:
and then *wayiphendula* ('you answered it'),*wayithini* ('how'), *ezi* ('these') *five rands*, *izawuphuma* ('it will be out'), *so then iye elwandle* ('it will go to the sea)

Regarding the other categories in the frequency list we may make the following observations. It is interesting to note that among the top 10 items on the frequency list, the majority are either adverbials or conjunctives. Interestingly enough, even if we look at frequencies lower than the 100 most frequent words, there are very few verbs and nouns. Once again, this distortion of the facts is a function of the nature of agglutinating languages. The rich variability of Xhosa agreement and tense/aspect morphology simply obscures the frequency of verbal lexical tokens.

Given the complex linguistic environment with widespread multilingualism in South Africa (Mesthrie 2002), access to actually occurring spoken data is essential for understanding and studying language contact on the micro level.

Below, we contrast code mixing with code switching by showing the 10 most frequent words of each kind, observing that entirely different classes of words are subject to the respective process. (Code switching involves mostly function words and code mixing involves mostly nouns).

Rank	Code switching	Code mixing
1	and	iidrugs
2	ok	iaids
3	ja	icrime
4	so	iright
5	because	eyione
6	but	ihiv
7	like	etown
8	neh	igovernment
9	man	iifirms
10	and then	ichance

Some of the words in the table like *ok* and *ja* might have been standard in spoken Xhosa for a long time and could for this reason also be seen as adoptives. However, since such feedback words usually have little morphological inflection, a decisive criterion here would be whether they show phonological adaptation to spoken Xhosa.

To make full use of corpus data, existing software tools need to be adapted or redesigned in order to deal with the (specifics of the spoken) language in question and here there are several interesting options. Either a tool for a closely related language (or even the corresponding written standard) can be quickly adapted (Bosch et al. 2008) or one may opt for a more data-driven approach (De Pauw and de Schruver 2009), which promises greater flexibility and less human labor, at the expense of accuracy.

4. Conclusion

This paper has described past, ongoing and planned work on multimodal corpora for all the South African official languages. We have also exemplified some of the possible uses of the corpus by presenting data on spoken Xhosa word frequencies and loans in spoken Xhosa from English and Afrikaans.

5. Acknowledgements

We are grateful to the Swedish Research Links Program and the Cooperative research program between Sweden (VR) and South Africa (NRF) for funding of the project.

6. References

Allwood, J. (1988). *Some Frequency Based Differences between Spoken and Written Swedish*. Papers from the 16th Scandinavian Conference of Linguistics, Turku University, Dept. of Finnish and General Linguistics, pp. 18-29.

Allwood, J. (2006). *Language Survival Kits*. In Saxena, Anju & Borin, Lars (eds.). Lesser-known Languages of South Asia. Mouton de Gruyter, Berlin, New York.

Allwood, J., Hendrikse, A.P., & Ahlsén, E. (forthcoming) *Some problems associated with the word as a basis of interlinguistic comparison*, Submitted.

Allwood, J., Hendrikse, A.P., & Mmemezi, M. (eds.). (2005). Guidelines for Developing Spoken Language Corpora. Spoken African Language Corpora Series, UNISA, Dept. of Linguistics, Pretoria, South Africa.

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues «peu dotées»*. Université Joseph-Fourier, Grenoble I doctoral dissertation.

Bosch, S., Pretorius, L., Podile, K. & Fleisch, A. (2008). *Experimental Fast-Tracking of Morphological Analysers for Nguni Languages.* In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, pp. 2588-2595

De Klerk, V. (2006). *Corpus linguistics and world Englishes : an analysis of Xhosa English* (Corpus and discourse 3). London: Continuum.

Mesthrie, R. (ed.) (2002). *Language in South Africa*. Cambridge: Cambridge University Press.

De Pauw, G,. & de Schryver, G-M. (2009). *African Language Technology: The Data-Driven Perspective*, In: Verena Lyding (ed.) LULCL II 2008 -- Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (EURAC 54), pp. 79-96. Bolzano: European Academy.

Scannell, K. P. (2007). *The Crúbadán Project: Corpus building for underresourced languages*. In C. Fairon, H. Naets, A. Kilgarriff & Gilles-Maurice de Schryver (eds.), "Building and Exploring Web Corpora": Proceedings of the 3rd Web as Corpus Workshop in Louvain-la-Neuve, Belgium, September 2007 (Cahiers du Cental 4), pp. 5-15.

Streiter, O., Scannell, K. P. & Stuflesser, M. (2006). *Implementing NLP projects for noncentral languages: instructions for funding bodies. Strategies for developers*. Machine Translation 20(4), pp. 267–289.