Spoken Language Corpora in South Africa

Jens Allwood

Department of Linguistics, University of Gothenburg, Renströmsgatan 6, S-412 55 Gothenburg, Sweden
Visiting scholar: Department of Linguistics, Unisa, jens@ling.gu.se

A. P. Hendrikse

Department of Linguistics, University of South Africa, P. O. Box 393, Pretoria 0003, South Africa, hendrap@unisa.ac.za

Abstract

In this paper we give an outline of a corpus planning project which aims to develop linguistic resources for the nine official African languages of South Africa in the form of corpora, more specifically spoken language corpora. In the course of the article, we will address issues such as spoken language vs. written language, register vs. activity and normative vs. non-normative approaches to corpus planning. We then give an outline of the design of a spoken language corpus for the nine official African languages of South Africa. We consider issues such as representativity and sampling (urban-rural, dialects, gender, social class and activities), transcription standards and conventions as well as the problems emanating from widespread loans and code switching and other forms of language mix characteristic of spoken language. Finally, we summarise the status of the project at present and plans for the future.

Introduction

In this article we give an outline of a joint corpus linguistics project between the Departments of Linguistics at Unisa and Gothenburg (Sweden). The project aims to develop computer-based linguistic resources for the nine official African languages of South Africa in the form of spoken language corpora. The raw data of the corpora come from audio-visual recordings of natural language used in various social activities.

Although this project is administered by the two linguistics departments mentioned above, we would like to involve as many African linguists and scholars as possible working on these languages as full participants in this project. One of the aims of this article, then, is to publicise this project, its goals, methods and potential outcomes to the relevant community of scholars in South Africa.

The rationale behind the project

Diminished and diminishing linguistic diversity is a characteristic feature of our contemporary world. This feature is, to a large extent, a function of the effects of globalisation on diversity. Factors such as global socio-economic pressures, the need for international communication standards and stable geo-political relations seem to entail inevitable monolingualism at the expense of linguistic diversity. About half of the approximately 6 000 languages spoken in the world today will be extinct by the end of the century for the simple reason that 90% of the world's population speaks the 100 most-used languages (Nettle & Romaine, 2000: 8). Even some of the 100 most-used languages may ultimately succumb to what Granville Price (as quoted by Nettle & Romaine, 2000:5) has aptly called the "killer language", namely English or, more precisely, World Englishes. English in all its varieties is simply the predominant medium of international linguistic interaction.

Why, then, given these overwhelming trends towards global monolingualism, should any speech community channel any efforts and resources towards the maintenance of their language? In a sense, the Asmara Declaration, which was issued by the delegates to a conference entitled *Against All Odds: African Languages and Literatures into the 21st Century* held in Asmara, Eritrea from 11 – 17 January 2000, is an attempt to answer this question.

1. The vitality and equality of African languages must be recognized as a basis for the future empowerment of African peoples.
2. The diversity of African languages reflects the rich cultural heritage of Africa and must be used as an instrument of African unity.
3. Dialogue among African languages is essential: African languages must use the instrument of translation to advance communication among all people, including the disabled.
4. All African children have the inalienable right to attend school and learn in their mother tongues. Every effort should be made to develop African languages at all levels of education.
5. Promoting research on African languages is vital for their development, while the advancement of African research and documentation will be best served by the use of African languages.
6. The effective and rapid development of science and technology in Africa depends on the use of African languages, and modern technology must be used for the development of African languages.
7. Democracy is essential for the equal development of African languages and African languages are vital for the development of democracy based on equality and social justice.
8. African languages, like all languages, contain gender bias. The role of African languages development

must overcome this gender bias and achieve gender equality.

9. African languages are essential for the decolonization of African minds and for the African Renaissance.

The South African spoken language corpus (SASLC) project subsumes, directly or indirectly, all the concerns expressed in this declaration, but more specifically the concerns raised in points 3 – 6, in the sense that it will develop a platform of computer supported basic linguistic resources for applications in translation (point 3), language teaching (point 4), language development (point 5) and language adaptations for science and technology (point 6).

Compared to corpora of English, SASLC is perhaps most similar to the Wellington corpus of spoken New Zealand English (Holmes et al., 1998), to the spoken language part of BNC (British National Corpus) and to the London/Lund corpus (Svartvik, 1990). Compared to spoken corpora of the Nordic languages, SASLC is similar to the Danish BySoc corpus (Gregersen 1991; Henrichsen 1997). The SASLC project is however distinct from these spoken language corpora in that its sampling is activity related, i.e. natural language use in a representative range of socio-economic activities. In this regard, SASLC is very similar to and largely guided by the approach of the Gothenburg spoken language corpus (GSLC).

In the next section we briefly contrast spoken and written language and indicate why we focus on spoken language in this project.

Why spoken language?

Structuralist linguistics for a long time has favoured (explicitly and perhaps mostly implicitly) the view that the difference between spoken and written language is of no relevance to linguistic theory. In addition to the more applied objectives of the SASLC project (such as language development) we also aim at a critical examination of this linguistic orthodoxy. That is, we hope that our study of spoken language will throw some light on the question whether the difference between spoken and written language is of any theoretical significance. We maintain that there is sufficient reason to believe that the difference is indeed theoretically significant and therefore worthy of empirical study. A basic reason is that spoken language has evolutionary primacy over written language, i.e. human beings seem to be genetically predisposed for speech.

Another reason is that the structure of spoken and written language, although similar in some respects, is also very different in many ways. Face-to-face spoken language is interactive (in its most basic form), multimodal (at the very least containing gestures and utterances) and it is also highly context-dependent. Further, spoken discourse very often consists of one word utterances. Written language, on the other hand, in its most typical form is non-interactive, monological and monomodal with a lesser degree of contextualisation. Typically, written language involves sentences which are governed by normative rules that dictate the structure of properly formed sentences. The norms of spoken language are usually of a different sort, rather dictating communicative efficiency enabling high rate processing required by speech.

In spoken language we therefore find linguistic expressions that enable "online" thought processing or expressions that allow for change of mind. From a normative written language perspective these linguistic phenomena might be called "dysfluencies", "false starts", "self-corrections" etc. In spoken language one also finds short and unobtrusive ways of giving discourse feedback, e.g. expressions like *ee, mh*, *yuh* that indicate comprehension, affirmation, surprise and so on.

None of these linguistic phenomena that are so characteristic of spoken language have any place in written language. Through the development of spoken language corpora we therefore hope to broaden the empirical basis for work on what we believe ought to be the central area of linguistic research, namely face-to-face linguistic interaction.

Considerations in the compilation of a spoken corpus

The compilation of a spoken corpus in the multilingual environment in South Africa is seriously affected by at least two features of everyday language use: dialectical variations, on the one hand and, on the other hand, interlingual communicative strategies, such as loans, code-switching, urban koines (cf. Schuring, 1985). If one is aiming at recording natural language use, as we are, all the natural features of language use in a multilingual society, including dialectal variation and language mix, need to be recorded and accounted for. This problem relates to the rather contentious issue of representativity.

Obviously, representativity depends on the kinds of variables that are selected to guide the empirical scope of the study. The deliberate bias of our project is on language use in a representative sample of social activities. This does not mean that we ignore other equally important variables. We deal with these variables in a particular annotational fashion rather than using them in the sampling criteria. In the SASLC representativity does not allude to sociolinguistic variables such as regional dialect, gender, social class or age but rather the range of social activities. We do this because we want to get an ecologically valid picture of the functionality of a language, which would be very difficult to achieve were we to use the traditional interview format which is normally used to capture variation with regard to regional dialect, gender, social class or age.

We now turn to the project itself discussing the various phases and facets in more detail.

The project

Four initial phases are distinguished in the project: a recording phase, a transcription phase, a checking phase and a tagging phase. The overall progress of the project involves concurrent research activities in all four phases. In fact, the developments of various corpus tools, the creation and refinement of an archiving infrastructure, the training of research participants and even trial runs of research outputs require collateral work in all the phases more or less simultaneously.

The recording phase

This phase in the design and development of a corpus presupposes certain fundamental assumptions about various aspects of the data that will form the corpus. Generally speaking, the following parameters seem to guide such assumptions:

- representativity of the corpus
- control of variables in language varieties
- recording medium and storage
- volume/size of the corpus
- length of each sample

Allwood (2001) gives an outline of a different basis for a representativity measure for spoken language-corpora, namely social activities. Social activities have been taken as the basis for decisions on the range and scope of representative samples in the Gothenburg spoken language corpus of Swedish (GSLC). The following activity types are represented in the corpus (cf. Allwood et al., 2001).

Table 1. Activities in the GSLC corpus.

Activity	Recordi	Tokens	Duration
Arranged Discussions	2	9 09	0:47:15
Auction	2	27 89	3:14:11
Bus Driver/Passenger	1	1 34	0:13:37
Church	2	10 23	1:47:02
Consultation	16	34 28	4:08:47
Court	6	33 72	3:58:33
Dinner	5	30 00	2:49:54
Discussion	35	239 40	27:03:39
Factory Conv.	5	28 88	2:54:47
Formal Meeting	15	236 75	28:20:54
Games & Play	1	5 96	0:50:00
Hotel	9	18 13	9:49:55
Informal Conversatio	18	86 81	8:35:19
Interview	57	389 41	45:23:01
Lecture	2	14 66	1:38:00
Market	4	12 17	3:55:07
Party	1	4 35	0:27:01
Phone	32	14 61	2:02:00
Retelling of Article	7	5 29	0:42:00
Role Play	3	8 05	0:57:16
Shop	54	50 49	10:33:33
Task-Oriented Dialogue	26	15 34	2:05:20
Therapy	2	13 52	2:04:07
Trade Fair	16	14 11	1:22:06
Travel Agency	40	39 89	6:00:06
Total	361	1 344 4	171:43:34

In our pilot study on Xhosa we have recorded samples of activities such as meetings, teacher discussions and seminars, student discussion classes, sermons, burial services, kin group meetings, informal discussions and patient interviews in hospitals. It remains, however, a project goal to develop some form of systematisation of the types of activities that will form the basis of the sampling in order to prevent the sampling from remaining merely opportunistic.

Most corpora in their initial stages aimed at what seems to have been the benchmark of a million words. Although our project is open-ended, we have expressed our target in our pilot studies for each one of the languages at roughly 200 hours of recordings. On average, one hour of recording yields a text of approximately 5 000 words.

The transcription phase

There are two facets to the transcription of recorded samples in our project:

- meta-transcription information (the header)
- the transcription of the contributions of all the speakers in an activity with some mark-up (the body).

The meta-transcription information

The transcription that can perhaps best be described by means of an example. [A transcription manual for transcribers will be published shortly and for lack of space we will not go into all the detail concerning the header and the mark-up conventions used in the transcription body.]

Transcription Header
@ Recorded activity ID: V010501
@ Activity type: Informal conversation
@ Recorded activity title: Getting to know each other
@ Recorded activity date: 20020725
@ Recorder: Britta Zawada
@ Participant: A = F2 (Lunga)
@ Participant: B = F1 (Bukiwe)
@ Transcriber: Mvuyisi Siwisa
@ Transcription date: 20020805
@ Checker: Ncedile Saule
@ Checking date: 20020912
@ Anonymised: No
@ Activity Medium: face-to-face
@ Activity duration: 00:44:30
@ Other time coding: Various subsections in the activity
@ Tape: V0105
@ Section: Family affairs
@ Section: Crime
@ Section: Unemployment
@ Section: Closing
@ Comment: Open ended conversation between two adult female speech therapy students Bukiwe and Lunga at Medunsa.

Each information line is marked by the @ sign. The information lines with the exception of a few are self-explanatory and need no further comment. Let us take a closer look at those lines that are perhaps not that self-evident. The information in the recorded activity ID line: V010501 specifies the following: V = Video, 01 = project number, i.e. the current spoken language corpus project, 05 = the number of the tape within this project. Each participant in a recorded activity in the project gets a unique code. That is F1 (where F = female) is uniquely associated with Bukiwe and will again be used if she participates in another recorded activity. The general rule is that participants in the transcription remain anonymous and that all information that could identify them is removed from the transcription and retained in a separate file that is not publicly available. Headers are open-ended information structures and additional information about the participants (for instance their age, level of education, knowledge of other languages) could be freely appended.

The transcription (the body)

It would be quite natural to expect that the transcription of spoken language use should be in the IPA orthography. We have not made this choice for the following reasons:

- It is very difficult to decide how much phonetic detail from IPA should be included.
- It is hard to train transcribers in IPA and to achieve consistency between transcribers in their interpretation of the phonetic data.
- It is very time consuming to do IPA transcriptions (and by implication very costly).
- IPA transcriptions make comparisons between standard written language and spoken language quite difficult.
- There are very few computer-based analytical tools and statistics-based tools available for IPA.
- Finally, the focus of SASLC is not on speech analysis but rather on discourse analysis.
-

The mark-up conventions used in the annotations of the transcriptions of recorded activities in this project follow the transcription standards developed in the Department of Linguistics at Gothenburg University (cf. Nivre no date).

Three types of lines are distinguished in the transcription body – a contribution line preceded by the dollar sign $ (for speaker), a comment line preceded by the @ sign where comments about certain peculiarities in a contribution are provided, and a section line indicated by the § sign where the subsections of a sample text are designated. Consider the example below.

§ At office
Section line
$A: uyakhonza kanene < > Contribution
@ < nod >
Information lin

The section in the sample from which this excerpt comes is ‘at the office’. The contribution represents a complete communicative activity of one participant in the discourse. While the participant is making this contribution she nods and this concurrent non-verbal activity is marked by the angle brackets < > in the contribution and commented on in the comment line <nod>.

The mark-up conventions aim at explicating in the transcription a whole array of typical features characteristic of spoken language. As we noted earlier, these conventions will soon be available in the form of a transcription manual and for lack of space we will not cover all of them here. By way of illustration of the kinds of spoken language features that are represented in our transcription we will highlight some of the more common ones with the help of excerpts from a transcribed sample text.

Elisions, overlaps, comments, pauses, lengthening

§ Religion
$B: uyakhonza kanene
$A: ndiyakhonza owu ndiyamthand{a} [4 < uthixo > ndiyamthanda andisoze ndimlahle undibonisile ukuba mkhulu nantso ke into efunekayo qha]4 kuphela
$B: [4 nantso ke sisi // e: e:]4
@ < personal name: God >

In the contribution of A above there is another instance of an elision indicated by the curly brackets, *ndiyamthand{a}*. Typical of certain spoken language activities is the occurrence of overlaps where some participant(s) say(s) something during the contribution of the participant who has the turn. These overlaps are indicated by means of square brackets (and are numbered because there could be several) in the contribution of the participant whose turn it is. After the completion of this turn the overlaps are transcribed in contribution lines of the overlapping participants. In the excerpt above the bracketed overlap 4 illustrates this convention in the two contributions. Comment information can be of several kinds, for example, gestures, loans, code-switching and also names. The site of a comment is indicated in a contribution by means of angle brackets. In the contribution of A in the excerpt the transcriber wished to comment on the item *uthixo* (which in the written orthography would have appeared with a capital letter, viz. *uThixo* (‘God’) and used the angle brackets to indicate his intention. In the comment line preceded by @ he made the appropriate comment. The convention used to indicate pauses has also been discussed earlier. In this excerpt a pause of medium length is marked in the second contribution of B. Distinct lengthening of utterances, except those that are linguistically standard (as in for instance penultimate syllable lengthening) is indicated by means of a colon as in the overlapping contribution of B.

The checking phase

Each transcription should be checked, ideally independently, by more than one checker. The checking involves viewing a copy of the video recording while following the transcription. In our pilot study so far we have tried to arrange a meeting after each checking phase where the transcriber and the checkers discuss flaws in the transcription and try to resolve differences of opinion. The checking phase is not only important to ensure the reliability and validity of the corpus, but also functions as a feedback to recorders to improve recording techniques.

The tagging phase

We will now briefly comment on the development of a tag set for African languages in our project. The extensive inflectional variety within categories (e.g. up to 23 different classes of nouns with equally extensive concomitant concordial agreement varieties) requires some decision on the scope of the tag set. Should it represent slots/types and leave the paradigmatic varieties/tokens unspecified. For example, should the tag set only represent word classes, say, Noun without further reflection of the category- internal class distinctions, or should it represent the whole range of classes by means of different tags. We have opted for the latter approach in the development of a tag set in our project whereby paradigmatic varieties within a category are differentiated by means of different tags. Needless to say, this resulted in a rather sizeable tag set with rather serious implications for the manual tagging of the samples in the corpus.

The latter problem is addressed in several ways in the project. The tag set has been printed on charts (A1 paper size) in order to facilitate look-up. We are also in the process of developing computer-assisted manual tagging in the form of drag-and-drop tagging from tag set windows. And finally, we are currently developing an automatic computer tagger. Manual tagging is, however, still needed for the development of a training corpus and also for the correction of errors.

Conclusion

In conclusion we would like to briefly outline the scope of the potential research output of the corpus resources that will be developed in this project. Although the project is to some extent still in its beginnings stages where most activities were geared towards the building of an infrastructure as well as the training of researchers in the various facets of the project, sufficient progress has been made in some of our pilot studies to warrant the initiation of some research output activities as well.

Some of the possible long term results we hope to achieve through the project are the following:

(i) A database consisting of corpora based on spoken language from different social activities for the indigenous languages of South Africa. This database will be open to the research community, providing a resource for research and practical applications based on African languages.

(ii) A set of computer based tools for searching, browsing and analyzing the corpus. These tools will be developed in collaboration with the Department of Linguistics, Gothenburg University, Sweden.

(iii) Frequency dictionaries on the word level for the spoken language of the indigenous languages of South Africa. If written language corpora can be secured for these languages, we also expect to be able to provide comparative frequency dictionaries of spoken and written language for the same languages.

(iv) Frequency dictionaries based on morphological analysis of words.

(v) Analyses of a range of spoken language phenomena, such as own communication management and interactive communication (feedback, turn taking and sequencing).

(vi) Frequency based dictionaries for collocations and set phrases.

(vii) Descriptions of the language of different social activities, including, if this is seen as appropriate, frequency listings of words and phrases.

(viii) Syntactic analysis of spoken language and contributions to providing spoken language grammars for different African languages.

(ix) Analyses of spoken language, providing bridges to cultural analysis of narratives, values, politeness, etc.

These are nine possibilities we see at present. Which of them will actually be carried out will depend on the interests of the research team. Probably, as our work develops, also other types of analysis will appear.

Finally, let us reiterare the use that our corpora can have for comparative linguistic studies of African languages and for comparisons of non-African languages with African languages. In such comparisons, we hope to examine some typical spoken language phenomena such as feedback in comparisons between, for example, African languages, Afrikaans, English and Swedish.

The corpus can also be used as a resource for researchers and practitioners outside of linguistics, such as educators and speech therapists, for whom the corpus can serve as a basis for educational or therapeutic material or as an aid to the standardization of evaluative or diagnostic tests.

References

Allwood, J. (1995). An Activity Based Approach to Pragmatics. *Gothenburg Papers in Theoretical Linguistics* 76.

Allwood, J. (2001). Capturing differences between social activities in spoken language. In Kenesei, I. and Harnish, R. M. (eds.) *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam, John Benjamins, pp 301 –319.

Gregersen, F. (1991). *The Copenhagen Study in Urban Sociolinguistics, 1 & 2*. Copenhagen: Reitzel.

Henrichsen, P. J. (1997). Talesprog med Ansigtsløftning, IAAS, Univ. of Copenhagen. *Instrumentalis 10/97.*

Holmes, J. Vine, B. & Johnson, G. (1998). *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: Victoria University of Wellington.

Nettle, D & Romaine, S. (2000). *Vanishing Voices: The Extinction of the World's Languages*. Oxford: Oxford University Press.

Nivre, J. No date. *Transcription Standards: Semantics and Spoken Language*. Göteborg University.

Schuring, G.K. (1985). *Kosmopolitiese omgangstale: Die aard, oorsprong en funksies van Pretoria-Sotho en ander koine-tale*. Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Svartvik, J. (ed.) (1990). The London Corpus of Spoken English: Description and Research. *Lund Studies in English* 82. Lund University Press.

References

Aijmer, K & Altenberg, B. (eds.) (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Allwood, J. (1995). An Activity Based Approach to Pragmatics. *Gothenburg Papers in Theoretical Linguistics* 76.

Allwood, J. (2001). Capturing differences between social activities in spoken language. In Kenesei, I. and Harnish, R. M. (eds.) *Perspectives on Semantics, Pragmatics and Discourse*. Amsterdam, John Benjamins, pp 301 –319.

Allwood, J., Grönqvist, L., Ahlsén, E. & Gunnarson, M. (2001. Annotations and Tools for an Activity Based Spoken Language Corpus. In Kuppevelt J. (ed.), *Current and New Directions in Discourse and Dialogue*. Kluwer: Academic Publishers.

Biber, D. & Finegan, E. (1991). On the exploitation of computerised corpora in variation studies. In Aimer & Altenberg (eds.), *English Corpus Linguistics*. London: Longman, pp 198-2004.

Biber, D, Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Blomberg, M. Carlson, R. Elenius, K. Granström, B. Gustafson, J. Hunnicutt, S. Lindell, R. & Neovius, L. (1993). An experimental dialogue system: WAXHOLM. *Proceedings of EUROSPEECH 93*: 1867-1870.

De Beaugrande, R. Unpublished. *'Corporate Bridges' Twixt Text and Language: Twenty Arguments against Corpus Research And Why They're a Right Load of Old Codswallop*. Universidade Federal de Paraiba.

Gregersen, F. (1991). *The Copenhagen Study in Urban Sociolinguistics, 1 & 2*. Copenhagen: Reitzel.

Halliday, M.A.K. (1994). *An Introduction to Functional Grammar*. 2nd edition. London: Arnold.

Heeman, P.A. & Allen, J. F. (1994). The TRAINS 93 Dialogue. *TRAINS Techical Note 94(2).*

Heine, B & Nurse, D. (eds.) (2000). *African Languages: An Introduction*. Cambridge: Cambridge University Press.

Henrichsen, P. J. (1997). Talesprog med Ansigtsløftning, IAAS, Univ. of Copenhagen. *Instrumentalis 10/97.*

Holmes, J. Vine, B. & Johnson, G. (1998). *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington: Victoria University of Wellington.

Isard, A. & Carletta J. (1995). Transaction and action coding in the Map Task Corpus. Research Paper HCRC/RP-65.

Leech, G & Smith, N. (1999). The Use of Tagging. In van Halteren (ed.), *Syntactic Wordclass Tagging*, pp 23-36.

Leech, G & Wilson, A. (1999). Standards for Tagsets. In van Halteren (ed.), *Syntactic Wordclass Tagging*, pp 55-80.

Nettle, D & Romaine, S. (2000). *Vanishing Voices: The Extinction of the World's Languages*. Oxford: Oxford University Press.

Nivre, J. No date. *Transcription Standards: Semantics and Spoken Language*. Göteborg University.

Schuring, G.K. (1985). *Kosmopolitiese omgangstale: Die aard, oorsprong en funksies van Pretoria-Sotho en ander koine-tale*. Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Sinclair, J.M. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins.

Sinclair, J.M & Renouf, A. (1991). Collocational frameworks in English. In Aijmer & Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pp 128-144.

Svartvik, J. (ed.) (1990). The London Corpus of Spoken English: Description and Research. *Lund Studies in English* 82. Lund University Press.

Tognini-Bonelli, E. (1996). *The role of corpus evidence in linguistic theory and description*. PhD thesis, University of Birmingham. Published as *Corpus Theory and Practice*. Birmingham: twc.

Van Halteren, H. (ed.) (1999). *Syntactic Wordclass-Tagging*. London: Kluwer Academic Publishers.

Wolff, H.E. (2000). Language and Society. In Heine & Nurse (eds.), *African Languages: An Introduction*, pp 298-347.