The Spoken language corpora for the nine official African languages of South Africa

Jens Allwood
Department of Linguistics
University of Gothenburg
Renströmparken
S-312 98 Gothenburg
Sweden
Visiting Scholar
Department of Linguistics
Unisa
jens@ling.gu.se

A P Hendrikse (Contact)
Department of Linguistics
University of South Africa
P O Box 392
Pretoria 0003
hendrap@unisa.ac.za

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 a state of the art article on corpus linguistics, Geoffrey Leech (1991) observes that with the exponential growth of corpus linguistic studies throughout the world, documentation about these studies, particularly transcription standards and annotation schemes, must be made available to the wider scholarly community. With the growing interest in corpus linguistic studies and the initiation of more research projects within this linguistic approach on South African languages, it is important that these efforts be documented and publicised in the wider linguistic community to stimulate scholarly debate and collaboration and to afford the exchange of experiential wisdom.

In this article we would like to 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.

Our spoken language corpus 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).

Point 2 in the declaration addresses the current world-wide concern with indigenous knowledge systems. One of the focus areas of research of the NRF is indeed the indigenous knowledge systems of the various speech communities of South Africa. We would like to believe that our spoken language corpora project could also serve as a resource for research in this domain.

A language in all its varieties is essentially linked to the socio-economic activities of its speakers in the speech community. In fact, a spokesperson for British Telecom (as quoted by Cameron 2001) suggested in 1996 that "... life is in many ways a series of conversations". The survival and maintenance of a language, then, seems to be intimately tied up with its functioning in all the socio-economic activities of a speech community which has been granted the right, scope and opportunities to function as a speech community.

It is against this background that the activity-based spoken corpus project on the nine official African languages of South Africa was initiated. Eventually, a spoken corpus for all the official languages of South Africa should be developed. In anticipation of the envisaged broadening of the scope of this project we will henceforth refer to it as SASLC (South African Spoken Language Corpus).

It is perhaps appropriate to briefly consider the position of SASLC relative to other corpus linguistic projects. The goal of SASLC is to collect samples of spoken language use from as many social activities as possible in order to gain a reasonably comprehensive overview of the role of language and communication in the South African socio-economic life. This type of spoken language corpus is still fairly unique even for English, since most spoken language corpora have been collected for special purposes, among others, speech recognition studies, phonetic studies, dialectal variation studies or studies on the interaction with a computerized dialogue system in a very narrow domain, e.g. Map Task (Isard & Carletta 1995), TRAINS (Heeman & Allen 1994), Waxholm (Blomberg et al. 1993).

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).

To close this section, one final and rather important point about South African speech communities needs to be made, namely the multilingual environment and its impact on the nature of the corpora of language use that we have been collecting. Each and every speech community in South Africa is affected by the multilingual environment in which it functions. There are effects on the choice of language which, in turn, are related to the differences of the functional levels of the various indigenous languages. The languages of South Africa simply occupy different functional spaces, not only because of their historically differentially defined statuses, but more particularly because English is the only language with international status and functions. Wolff (2000: 307, 320) gives a very useful picture (which, for lack of space, we cannot repeat here) of the domains, participants and settings as well as the functions and legal status at various activity levels of indigenous versus international ("colonial") languages in Africa. Although the greatest potential for the survival of a language would be when it can function at all levels in society, this would simply be an unrealistic immediate expectation with regard to all languages spoken in South African. Despite these inequalities, all of these languages could co-exist harmoniously and without threat of extinction within a multilingual environment if there is a stable diglossic situation, i.e. if each language has its own high-valued functional space in the linguistic market place. Be that as it may, we would like to believe that the corpus resources that we will be developing should facilitate the ultimate functioning of previously disadvantaged languages in most, if not all, socio-economic communicative domains in South Africa.

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, and, needless to say, also to research pragmatics as De Klerk (2002: 27) observes:

> In designing any corpus, one also needs to admit that it is virtually impossible to document the full sweep of any language, including dialectical diversity across regions, social classes, ethnic groups and age groups and to include diversity that would allow comparisons across service counters, sermons, doctor/patient interactions, legal proceedings, planned and unplanned class lectures, conversations, and so on. Significant existing corpora have **not** generally aimed at this kind of coverage especially those of spoken language, given the enormous expense involved.

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 outline of the project that follows will discuss and illustrate the various facets of each one of the phases. Although the research activities in the project are necessarily sequenced according to these phases, i.e. first recording, then transcription and so on, 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.

Before we discuss each one of the phases in more detail, it is important to dwell briefly on the relation between theoretical linguistics and corpus linguistics. It is a fairly generally held belief that corpus linguistics is an approach (a set of methods and techniques) rather than a theory. Some corpus linguists (cf. Sinclair 1987, 1991; Sinclair & Renouf 1991) even maintain that corpus linguistic studies, including grammatical studies, should be theory-neutral – the corpus is sufficient in yielding the significant grammatical patterns in terms of statistical methods and criteria. This implicit or explicit distancing of corpus linguistic studies from theoretical linguistics may stem from the deliberate discrediting of corpus linguistics in certain theoretical linguistic circles (cf. de Beaugrande unpublished) or from fear of compromising linguistic data and potential findings by a particular theoretical bias (cf. Sinclair 1991; Leech 1991; Hunston & Francis 2000). The underpinning of corpus linguistics is supposedly then method rather than theory, but as Halliday (1994) suggests there are no "theory-free" descriptions. In fact, corpus linguistics may now, and even more so in the future, become the centre of gravity for across the board linguistic theorising, as has already become demonstrably clear in the effects it has on fairly entrenched theoretical assumptions about, for instance, the relation between grammar and lexis (cf. Sinclair 1991; Moon 1998) abstract linguistic knowledge (formal grammars) and the nature of language use of native speakers (Kjellmer 1991).

The relation between corpus linguistics and theoretical linguistics is important because we believe that the design of the project as well as the choices represented in the various phases of the project are theoretically informed and justified. Thus, although we do not

subscribe to any particular theoretical bias in the project, we do make fundamental theoretical assumptions about such issues as the principled differences between spoken and written language use, normativity, and activity-related spoken varieties as indicated in the previous section.

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

The representativity of a corpus is a contentious issue – the obvious question being representative of what? It is nevertheless an important issue as Biber et al. (1998: 246) note, "A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent." A random and arbitrary collection of texts (written and/or spoken) does not constitute a corpus, but perhaps rather an archive and the difference between the two types hinges on the question of representativity as Leech (1991: 11) suggests when he says "the difference between an archive and a corpus must be that the latter is designed or required for a particular 'representative' function. In the case of written texts, representativity can allude to genres (cf. Biber & Finegan 1991) but in the case of spoken language use very little is known about spoken genres. Text type is another possible basis for representativity and is supposed to complement genre text categories. Some functional labels such as, 'Informal Interaction', 'Learned Exposition', 'Involved Persuasion' have been suggested for the differentiation of texts types (cf. Biber et al. 1998). Register variation seems to be a fairly widely recognised but not necessarily very useful parameter for decisions on representativity simply because different registers may be used in one and the same situation.

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. The following activity types are represented in the corpus (cf. Allwood et al. 2001).

Activity	**Recordings**	**Tokens**	**Duration**
Arranged Discussions	2	9 098	0:47:15
Auction	2	27 890	3:14:11
Bus Driver/Passenger	1	1 348	0:13:37
Church	2	10 235	1:47:02
Consultation	16	34 285	4:08:47
Court	6	33 722	3:58:33
Dinner	5	30 001	2:49:54
Discussion	35	239 409	27:03:39
Factory Conversation	5	28 883	2:54:47
Formal Meeting	15	236 752	28:20:54
Games & Play	1	5 960	0:50:00
Hotel	9	18 137	9:49:55
Informal Conversation	18	86 817	8:35:19
Interview	57	389 416	45:23:01
Lecture	2	14 667	1:38:00
Market	4	12 175	3:55:07
Party	1	4 356	0:27:01
Phone	32	14 614	2:02:00
Retelling of Article	7	5 290	0:42:00
Role Play	3	8 055	0:57:16
Shop	54	50 492	10:33:33
Task-Oriented Dialogue	26	15 347	2:05:20
Therapy	2	13 529	2:04:07
Trade Fair	16	14 116	1:22:06
Travel Agency	40	39 899	6:00:06
Total	**361**	**1 344 493**	**171:43:34**

Socio-economic activities also form the basis for the sampling of language use in our spoken language corpus project. The types of activities that will be represented in the sample will naturally differ. Because of the multilingual situation in South Africa and also because of the unequal status (if not in theory or policy, then at least in practice) between English/Afrikaans on the one hand, and the African languages of South Africa on the other hand, the African languages simply do not function at all in certain socio-economic activities. De Klerk (2002) made a similar observation in her corpus study of Xhosa English – English is not naturally used in certain activities among Xhosa speakers. This is quite natural in a diglossic situation with different functional spaces for different languages. Formal meetings with participants from different speech communities will in all likelihood be conducted in English nowadays. Remarkably, in a recording of part of a pilot project, English was also the medium for about 50 percent of the time during a meeting of a Xhosa language department where all the attendants were native speakers of Xhosa. We believe that this gives a fair reflection of what is probably a characteristic linguistic feature of certain types of activities in the multilingual context of South Africa.

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.

Biber & Finegan (1991: 207ff) give a synoptic overview of inequalities between corpora given the medium of recording. Spoken interaction, contrary to what one might think, involves much more than merely oral communication. A large and very significant part of face-to-face communication is visual, e.g. gestures, facial expressions as well as the deictic context. These non-verbal phenomena accompanying face-to-face communication may be sensitive to activity type. For instance, one would not expect applause in an informal discourse among friends, or spitting during a church sermon while the latter action, though not very civilised, may be very expressive in a quarrel. We have therefore opted for the audio-visual medium of recording to capture both the verbal and the non-verbal facets of spoken language interaction. (Some tools for the correlation of utterances and accompanying gestures and facial expressions on a time

line are currently under development in the Department of Linguistics at Gothenburg.) Audio-visual recordings also facilitate the transcription process as regards speaker identification, observing non-speaking participants' responses and feedback, differentiating overlaps in simultaneous floor-taking, etc.

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.

Another issue that received some attention in the literature is the length of any specific sample text. Biber & Finegan (1991: 213) found in their study that "1 000-word samples reliably represent at least certain linguistic characteristics of a text, even when considerable internal variation is anticipated." Frankly, we have not considered setting any size limits on text samples, but rather to allow pragmatics (the natural progress of an activity up to its end, video tape length, etc.) to dictate the length of a sample. Naturally, some samples then become rather "unwieldy". Typically, texts (spoken or written) are sectionally structured – beginnings and ends, but also topic shifts – and it is important to map out these sections that are characteristic of each activity type (cf. Allwood 2001). In the transcription phase this mapping out of the relevant sections (to the extent that they are recognisable) is done by the transcriber by means of specific comment flags.

Finally, spoken language interaction in a natural setting (as opposed to a controlled studio setting, say, in a role play activity or in a scripted activity such as a news bulletin) more often than not has some or other bearing on the naturalness and quality of the recording. The presence of a camera can have all sorts of effects on the behaviour of the participants in an activity. Our experience, however, is that this over concern with the camera wears off pretty soon and everything returns to some form of naturalness in the discourse. The important point is that the recorder should try to make the camera as invisible as possible, for instance, by not moving around, but rather to select a fixed vantage point that will enable capturing all the participants in a frame. Furthermore, because spoken language involves the communicative interaction of more than one participant and all the communicatively relevant expressions (feedback, gestures and facial expressions) of everybody, it is important that the camera is **not** moved from one

interlocutor to another at each turn, but rather to keep all participants in the frame. For large groups, the use of more than one camera is probably the best solution.

The transcription phase

Certainly the most serious drawback of a spoken language corpus project (as opposed to a corpus of written texts) is the disproportionate demands on resources emanating from the transcription of the recorded samples. Needless to say, it is also the most crucial part in the development of a spoken corpus; without transcriptions there would be an audio-visual archive of recorded activities, but no computer-readable corpus.

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 interlocutors in an activity with some mark-up or annotations (the body).

The meta-transcription information

Every transcription consists of two parts – a header and a body (which is the transcription proper). The header contains the meta-transcription information and is made up of an array of different pieces of information about the recorded activity and 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

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. Admittedly, the use of IPA would have been quite helpful in the transcription of feedback and own communication management expressions, since there are no standardized orthographic correlates for such expressions.

Our choice of transcription standard has therefore been designed to meet the criticism of the use of IPA in spoken language transcriptions listed above. That is, our transcription system should be simple (not include too much phonetic detail), be easy to train transcribers in, be reliable, enabling fairly rapid transcription (lower costs), facilitate the comparison with written language and it should be amenable to computer supported analysis.

The orthography of the transcription is therefore the standard orthography of the indigenous language in question, excluding, however, all punctuation marks including capital letters. In order to make the transcription machine-readable, plain text format is used. Spoken language exhibits certain features that do not always have counterparts in written texts. In the case of African languages tones are a prominent and integral part of spoken language. Unfortunately, it is at the moment not practically possible to include a tone mark-up of our transcriptions. Hopefully, this could be done at a later stage. As mentioned before, the orthographic representation of communication management expressions (e.g. own communication management such as hesitations and interactive communication management such as feedback cf. Allwood 1995) has not been standardised for African languages. These types of utterances are actually very important in spoken language and should therefore be transcribed. Although we depend to some degree on the innovativeness of the transcribers, we are in the process of developing orthographic standards for these utterances as the current pilot transcriptions progress. Pauses and emphasis are also typical of spoken language. Pauses are relative and therefore there is some degree of subjectivity in their perceptual differentiation although some timing techniques can be used. In the transcriptions three pause lengths are distinguished by means of slashes, one / for short pauses, two // for medium length pauses and three /// for distinctly long pauses. Other typical features of spoken

languages are contractions and elisions. These phenomena are transcribed as they are perceived by the transcriber, but the standard written forms are represented by enclosing them in curly brackets in the transcription. The following example from a transcribed recording of a Xhosa discourse illustrates both elision and contraction.

Recorded: *mhlobam* ('my friend')
Transcribed: *mhlob{o}{w}am*

Some cases of contractions and elisions have already found their way into the written standard and are transcribed without any modification. Consider the examples of such a case from Xhosa below.

mtanam < *mntwana wam* ('my child')
kwedini < *nkwenkwendini* ('hey there, youngster')

Finally, we need to comment on the way in which we deal with the pervasive phenomenon of foreign language intrusions (loans and code-switching) in spoken language. Nothing that occurs in a spoken language sample is edited out, i.e. everything, including loanwords and stretches of code-switching, is transcribed but annotated by means of angle brackets together with relevant comments in the comment lines. The retention of these loans and code-switches is important in that they are more prevalent in certain types of activities than in others and as such they are linguistic indices of the nature of the spoken language associated with certain activities. Moreover, they are also indices of the dynamics of language change in South Africa. Very often the foreign intrusions are indigenised in some way or another and these indigenised forms are captured in the transcriptions without any change. Consider the rather interesting example below.

<*empumakoloni*> ('Eastern Cape': English loan: *colony*)

This is an interesting example not only because of the fact that it shows how a loanword has been fully integrated with a grammatical construction, but also because of the term creation strategy that has been followed here. The Xhosa word for the east is *empumalanga* (lit. 'where the sun rises'). The first part of this direction term (carrying the analogical significance of the 'east') has been ingeniously prefixed to the loanword *koloni* ('colony') which refers to the Cape Province yielding the significance 'Eastern Cape'.

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 line

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.

<u>Elisions, overlaps, comments, pauses, lengthening</u>

§ 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.

Contrastive stress

> $B: abanye ke bazihlalele nje: / abanye ABAZANGE bafune sikolo // uyayiqonda ke la meko yokungabikho mzali uqhubayo / uthi aba baza emva kwam bobabini ABAZANGE bafunde kuyaphi // kodwa ke // andigxeki nto kuba ke / ndibakhona ngethuba le ngxaki nobhuti ke [2 abeyinkxaso kakhulu]2
> $A: [2 ya / m: ewe]2 hayi izinto zikuthixo azikho kuthi nam obu bushuman bam ndiseza kutshata ndiseza kutshata

Contrastive stress is indicated by means of capital letters as in the contribution of B in the excerpt above – *ABAZANGE*. Notice also the examples of lengthening, pauses and overlaps in the excerpt.

Unclear speech and code-switching

> $M: loo nto ke njengo{ku}ba sekunyanzeleke ukuba ndiye phaya nje (...) ndikwazi ukuncedisa phaya ndiyiphushile ukwenzela ukuba ndibe <neclaim>

endizakuba nayo <that is why> ndithole <because ndiyaclaimer so that at least> uba <ndiclayimile> ndikwazi ukuhamba
@ <code-switching: English>
$T: ke ngoku ke yenye yezinto endifuna ukuyoyenza
$M: ngolwesithathu (<what she said to me> ngoku bendiphaya) ngecawe besingcwaba umfazi kasicaka jama
@ <code-switching: English>

Unclear speech in a contribution is indicated by means of round brackets. If nothing is audible a dotted line enclosed in round brackets is used as in the contribution of M in the excerpt. If the transcriber is unsure of what is said he/she gives some rendition again enclosed in round brackets, as in the second contribution of M. The code-switching in the first and second contributions of M is left intact but appropriately commented on in a comment line. Notice the degree of indigenization in the code-switching *ndiyaclaimer / ndiyaclayimile* ('I claimed').

One final comment on the transcription phase is in order here. Annotation of samples in a corpus always represents some kind of research and/or theoretical bias. It has therefore been suggested, among others, by Leech (1991: 25) that "an annotated corpus should never totally replace the corpus as it existed prior to annotation. The original 'raw' corpus (including the original sound recordings) should always be available, so that those who find the annotations useless or worse can recover the text in its virgin purity." All the recordings of our project are archived and although we do not maintain 'virgin' transcriptions of the samples, their reinstating should be reasonably easy by fairly straightforward computer-based editing functions.

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

Since we will report in some detail about the development of a tag set and the tagging procedures (manual and automated) in another article in this supplement, a few general comments about this phase in the project will suffice here.

Two general approaches to the retrieval of information from a corpus have been distinguished in corpus linguistics – the corpus-driven approach and the corpus-based approach (cf. Tognini-Bonelli (1996). In the corpus-driven approach information is retrieved from a raw, i.e. un-annotated corpus typically by means of the KWIC (Key Word in Context) method. The work of Sinclair (cf. Sinclair 1987, 1991, Sinclair & Renouf 1991; Hunston & Francis 2000) on grammatical patterns typifies this approach where orthographic space between units in a corpus is taken as the basis for information retrieval. This kind of approach seems to work fairly well in isolating languages particularly in respect of lexical pattern analysis. In the case of agglutinating languages, such as the languages of our corpus, orthographic space is not a very useful basis for information retrieval and some form of annotation is required in order to retrieve the significant patterns. In fact, even in the case of isolating languages, the search for patterns associated with specific linguistic phenomena requires relevant annotations schemes. The annotation of corpora by means of various types of tags is typical of the corpus-based approach. And although one of the strongest advocates of the corpus-based approach, Leech, warns against the danger of bias underlying any form of annotation, the tagging of corpora is now fairly general practice in most corpus linguistic studies (cf. Leech 1991).

There are obviously a whole range of linguistic properties that could be tagged in corpora (cf. Leech 1991; Leech & Smith 1999), but generally speaking, the tagging of morphosyntactic properties, more particularly, word classes is the most common practice. In inflectional languages, however, morphosyntactic units are typically portmanteau morphemes, i.e. several grammatical significances are cumulated in a single morpheme. In some tag sets, these cumulative values are distinguished from each other by means of separate symbols for each significance as in the English tag PPs1N (personal pronoun, 1st person, singular, nominative) for the personal pronoun *I* (cf. Leech & Wilson 1999).

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

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.

Cameron, D. 2001. *Working with Spoken Discourse.* London: Sage.

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.

De Klerk, V. 2002. Towards a Corpus of Black South African English. In *Southern African Linguistics and Applied Language Studies* 20: 25-35

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.

Hunston, S & Francis, G. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

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

Kjellmer, G. 1991. A mint of phrases. In Aijmer & Altenberg (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik,* pp 111-127.

Leech, G. 1991. The state of the art in corpus linguistics. In Aijmer & Altenberg (eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik,* pp 8-29.

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.

Moon, R. 1998. *Fixed expressions and idioms in English.* Oxford: Clarendon Press.

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. 1991. *Corpus, Concordance, Collocation.* Oxford: OUP.

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.