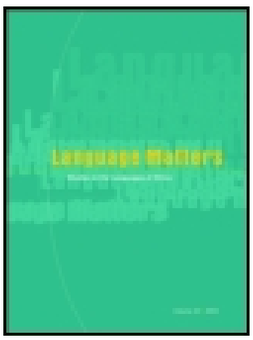

ISSN: 1022-8195 (Print) 1753-5395 (Online) Journal homepage: https://www.tandfonline.com/loi/rlms20

Educated mother-tongue South African English: A corpus approach

Vivian de Klerk , Ralph Adendorff , Mark de Vos , Sally Hunt , Ron Simango , Louise Todd & Thomas Niesler

To cite this article: Vivian de Klerk , Ralph Adendorff , Mark de Vos , Sally Hunt , Ron Simango , Louise Todd & Thomas Niesler (2006) Educated mother-tongue South African English: A corpus approach, , 37:2, 206-226, DOI: 10.1080/10228190608566261

To link to this article: https://doi.org/10.1080/10228190608566261

Educated mother-tongue South African English: A corpus approach

Vivian de Klerk, Ralph Adendorff, Mark de Vos, Sally Hunt, Ron Simango, Louise Todd
Department of English Language and Linguistics
Rhodes University
Grahamstown
Eastern Cape Province
South Africa
v.deklerk@ru.ac.za

Thomas Niesler
Department of Electrical and Electronic Engineering
Stellenbosch University
Western Cape Province
South Africa

Abstract

South Africa is anecdotally known for its complex system of speech varieties correlating with variables such as ethnicity, first language, class and education. These intuitions (e.g. Lass 1990) require further investigation, especially in the context of a changing South Africa where language variety plays a key role in identifying social, economic and ethnic group membership. Thus, in this research, the extent to which these variables play a role in variety is explored using a corpus approach (the nature of class and race in the corpus is discussed more fully later in the article). The corpus project, focusing primarily on accent, has been undertaken by members of the Department of English Language and Linguistics at Rhodes University in South Africa, collaborating with staff from the Department of Electrical and Electronic Engineering from Stellenbosch University, South Africa. A corpus (the first of its kind) is being compiled, comprising the speech of educated, white, mother-tongue speakers of South African English (as distinct from Afrikaans English, Indian English, and the second language (L2) varieties of English used by speakers of indigenous African languages), and data collection is well under way. This short article aims to describe the aims of the project, and the methodological approach which underpins it, as well as to highlight some of the more problematic aspects of the research.

Keywords: corpus; Lancaster–Burgen–Oslo–Bergen; LOB; mother tongue; South African English.

 ISSN 0256-5986 Language Matters 37 (2) 2006 pp. 206–226

Overview

Lass (1990) makes the point that South African English (SAE) is an enormously complex and grossly under-described dialect cluster, which includes mother-tongue and second language (L2) varieties, and is severely lacking in detailed descriptions of single, localised examples of particular varieties. Over 200 years have passed since the first English speakers (admittedly few in number) set foot on South African soil, and scholars still need detailed descriptions of the systems which underlie the various varieties of English that have evolved since then. Admittedly, there have been a few isolated attempts to describe selected varieties over the past 40 years, including Lanham and Macdonald's (1979) descriptions of so-called 'conservative' 'respectable' and 'extreme' South African English, and Lass's (1990) paper, which describes one local standardised variety, namely 'respectable upper middle class white Cape Town English'. In addition, we have descriptions of Indian English (Mesthrie 1992) Afrikaans English (Watermeyer 1996), and Xhosa English (de Klerk 2002).

The classification of South African English alluded to above was based on observations of speech patterns that prevailed in the 1970s and 1980s (and probably earlier). In all likelihood these varieties now exhibit different characteristics which need to be systematically studied and described. Our study seeks to provide a description of current South African English, that is, the variety spoken at the beginning of the twenty-first century. In embarking on such research, one is faced with the choice of providing a very detailed account of a restricted linguistic community – a microcosm of sorts (such as Lass's (1990) description) or of trying to describe a broader, and inevitably 'messier' and less distinctively identifiable variety, which people intuitively recognise as being spoken by a wider range of speakers. As a starting point, we seek to describe a variety of English which we characterise as *Educated South African English*, since level of education (as Seppe, Maxim and Wells (2000) demonstrate) significantly influences speech patterns and creates identifiable dialect groups in the wider speech community.

Without intending to criticise unduly, we need to acknowledge that many of the earlier studies cited above were heavily reliant on limited observations of speech patterns (given the constraints on the quality of tape-recordings that could be made at the time) and the informed guesses and intuitions of the phonetic experts at the time, who did the best they could with the limited resources available to them. But with ongoing recent advances on the technological front, which make high-quality recording and phonetic analysis much easier to achieve, one can now aim to describe such varieties once more, basing descriptions on a greater amount and higher quality of data as well as more efficient analytical tools.

As part of a wider project, this article sets out to describe the methodological

issues involved in describing this variety. Such research involves numerous thorny technical, theoretical and ideological issues, each of which had to be 'resolved' (insofar as such questions can ever be resolved) prior to embarking on the collection and analysis of the data.

Historical background

A description of any variety of South African English needs to take into account the socio-historical background of the linguistic communities involved, and the nature of movement and diffusion between various social groups. Inevitably, this involves taking into account the old racial classification systems of the twentieth century, and as Lass (1995, 89) points out, 'however unpalatable its socio-political implications and however unsavoury its origins', these factors are unavoidable and deeply significant. It was white, English-speaking people who came from the south of Britain in the early 1800s, and these people tended to mix with and marry their own kind. The enforcement of apartheid until 1994 entrenched racial divisions, and as a consequence, the variety which we seek to describe is, essentially the variety of a white linguistic community.

While further detail and descriptions of the historical origins of the English speakers who came to South Africa are available in Branford (1994), Lass (1995) and Lanham (1996), a brief summary overview of the pertinent historical background follows.

South African English is an extra-territorial, or transported, variety which shares features common to English spoken in the Southern Hemisphere, including Australia and New Zealand. (Although the distinction is loosely made, Northern Hemisphere Englishes such as American English are varieties that evolved from people who left the British Isles during the 1600s, while Southern Hemisphere Englishes are varieties based on emigrations after 1795.) These Southern Hemisphere varieties represent the dialects of Southern British English, and all share basic characteristics which include a raised [æ] in TRAP, distinctive vowels in STRUT ([ɐ]) and FOOT ([ʊ]), a long [a:] in BATH, and lengthened [æ:] before voiced nasals and stops (CAT vs. CAD). (Descriptions are from Lass (1995, 90) which is based on Wells's (1982) standard lexical sets.)

Both input and history are vitally important to the phonetic features which subsequently evolved: apart from the different timing of waves of emigration north and south, another reason for considerable dialectal differences relates to the fact that, historically, the Southern Hemisphere varieties kept up their ties to Britain throughout the nineteenth and twentieth centuries, while Northern Hemisphere varieties generally did not.

The history of English in South Africa goes back to 1795, when the British took

over the Cape from the Dutch (who had been settled there since 1652). The English speakers between 1795 and 1820 were relatively impermanent, mainly engaged in a military 'holding operation', but this changed in 1820, with the large influx of around 5 000 settlers from the Home Counties (Lanham 1996, 20) to the Eastern Cape, encouraged to immigrate in order to form a human barrier against the indigenous people on the boundaries of the British colony. It was these speakers, predominantly rural and working-class, who formed the true 'seeds' of South African English, forced by harsh frontier circumstances to work closely together in order to survive and prosper in their new environment. English was declared the sole official language in 1822, and a second wave of immigration followed in 1840–1850, but these people came from higher social classes, and settled in Natal. The discovery of diamonds and gold led to further influxes of English speakers after 1875, this time to Kimberley and the Witwatersrand, along with speakers of several European languages. The Natal settlement was not without influence, since these people served to reinforce the nostalgic, positive attitudes to British norms of speech (which has ensured the survival of these norms over the next 100 years or so), but it is the initial settlement in 1820 which had the most lasting effect on the South African English accent.

The mother tongue — varieties of English which slowly evolved in South Africa lie on a continuum of standardness, ranging from a dialect very close to the Southern British dialectal source, to a second, increasingly prevalent, local standard dialect (Lass 1995, 90). According to Lass (1995) and Lanham and Macdonald (1979), by the twilight of the twentieth century, three lectal types had emerged:

1. A conservative, or 'cultivated' variety, typical of the upper classes (especially females older than 45), and strongly similar to British RP.
2. A general, or 'respectable' local standard variety, largely middle-class yet prestigious in its own right.
3. Local vernaculars, described as broad or 'extreme', stigmatised by speakers of the other two lectal types, as being strongly associated with masculine speech, and showing the influence of Afrikaans.

Social fluidity has subsequently had a complicating effect on varieties in recent years. Since the demise of apartheid in 1994 there have been fundamental and rapid changes in access to education, and political power has shifted into black hands, with racial boundaries no longer holding as much force as they did formerly. With these dramatic shifts, those occupying the most powerful positions in South Africa, politically, socially and economically, are now more likely to be mother-tongue speakers of indigenous African languages. Steady emigration since the mid 1980s of white English speakers has also depleted the

number of mother-tongue speakers in the country (recent statistics (Census 2001) report a total of 1 687 661, which is 3.8 per cent of the population of 44.8 million). In addition, the national broadcaster, the SABC, has changed its criteria in terms of which announcers are selected, so that the full range of South Africa's accents are given exposure, and the official media are dominated by L2 varieties of English. The lack of an 'officially' approved overt, prestige variety of English leaves the way open for the emergence of new standards, which are likely to evolve rapidly in the new socio-economic climate of South Africa. Given this sudden impetus for rapid change, this research project aims to capture and describe (insofar as this is feasible) one of the many varieties of English that coexist in South Africa, the (possibly threatened) variety of English spoken by educated mother-tongue white South Africans, which we have chosen to call *Educated Mother-Tongue South African English*.

The aims of the project

The primary aim of the project is to collect a corpus of recordings of Educated Mother-tongue South African English, which includes the following samples: the reading of numbers and isolated words selected from Wells's (1982) lexical sets,[1] the reading of phonetically rich sentences extracted from the SCRIBE corpus, the reading of extracts from the Lancaster–Oslo–Bergen (LOB) Corpus (since this corpus is already fully tagged, and is available for *bona fide* research purposes) and casual conversation with an interviewer. The objective of including this range of data is to ensure the collection of a continuum of speech styles from formal and highly monitored to informal, spontaneous and unmonitored. A range of hypotheses may be tested using this corpus. Other important aims of the study are:

1. to establish whether there is indeed a variety that can be called *Educated Mother-tongue South African English*, and what the phonetic characteristics of this variety are
2. by describing this educated variety, to achieve an understanding of the ways in which it might be used to restrict social and economic mobility in the context of a country which has a commitment to socio-economic change
3. to develop large-vocabulary continuous speech recognition systems for South African English using recordings and the accompanying tagged text transcriptions
4. to allow focused research into the use of part-of-speech (POS) information in the statistical language models employed by the large-vocabulary recognition systems.

Secondary aims which we plan to explore in the corpus at a later time include the following:

1. to investigate how this variety differs, phonetically, from other varieties of South African English and colonial Englishes (e.g. Australian or New Zealand Englishes)
2. to describe selected aspects of the discourse of the interview sections of the data, especially in terms of the use of discourse markers, and other narrative strategies
3. to compare formal reading styles (e.g. used during reading of word lists and texts) and informal spontaneous discourse (characterising the interviews)
4. to analyse the reading strategies of informants in the study, and explore points where errors tend to occur
5. to compare (both in terms of phonetics and discourse) selected sub-groups of speakers, e.g. females versus males, old versus young.

Methodology

The informants

The sample of informants whose speech was collected for the corpus comprises 50 adult first-language speakers of English aged between 30 and 70 years. The lower age limit was set at 30 years of age in order to focus on a 'stabilised accent', since it has been shown (Romaine 1984) that the speech and accents of young people is highly variable during adolescence and during their early twenties, when speakers are more vulnerable to adjusting their accents because of peer pressure and other social influences. Another cogent reason for choosing 30 as a cut-off point relates to the fact that all potential subjects would have been 18 years or older in 1994, the time when schools became multiracial. All our potential informants did not experience that linguistic 'melting pot', and were likely to have accents modelled fairly closely on the mother-tongue English we are seeking to describe.

The focus of data collection in this study was deliberately kept fairly narrow, namely, focusing only on mother-tongue speakers of English, so as to form as homogenous a starting point as possible from which to expand later, when comparisons with other varieties could be made. Another advantage from our point of view was the accessibility of suitable subjects from the staff of our English-medium University (Rhodes) and several other English-medium schools in Grahamstown. (The population of the town is transient to a high degree, as is evidenced by the geographical origins of our subjects (see Figure 1).)

In order to ensure that informants were 'educated' and had stable South African

mother-tongue accents of English, the following additional criteria were applied in selecting informants:

- subjects had to have 12 years schooling plus a minimum of 1 year of tertiary study at English institutions
- all education had to have been completed in South Africa;
- both parents had to be mother-tongue speakers of English (with at least one parent speaking SAE);
- subjects had to have spent most of their lives in South Africa;
- subjects could not have spent more than 12 months outside South Africa in the previous three years.

Initially, criteria were more stringent: both parents, as well as the spouse (if applicable) of subjects, had to be mother-tongue speakers of SAE and be born in Southern Africa. In practice, we found that that this excluded a large proportion of our potential subjects, as well as those whom we felt intuitively to be exemplar members of this linguistic community (for example, those who had an Afrikaans parent or spouse).

As a start, potential subjects were identified drawing on the acquaintances and personal networks of 5 members of the research team. A surprisingly high number of these potential subjects actually did not qualify in terms of the listed criteria, or were unwilling to be interviewed, and at that stage the second-phase of selection was utilised: they were asked to identify suitable people from their networks whom we could approach. Thus, although the sample was not randomly selected (which would have been highly impractical, given the low numbers of potential subjects) we felt it was sufficiently widespread to claim some representivity.

The spoken data

The primary texts selected for extended reading in the corpus described in this study were drawn from the well-known Lancaster-Oslo Bergen corpus (LOB) of written British English, which comprises approximately 1 million-words of English text that have been carefully annotated with a detailed set of POS tags. By using this LOB corpus (already tagged and available to the research community) and overlaying it with our Educated South African English accents, we are able to produce a grammatically tagged speech corpus without the intensive and laborious exercise of transcribing and tagging and checking the speech ourselves.

The speech database for Educated South African English comprises recorded

recitations of word lists, phonetically rich sentences and recorded readings of extracts from the LOB Corpus (Garside et al. 1987), as well as short interviews to capture spontaneous speech. These were elicited in this order so as to move from the most formal, self-monitored reading style to less formal more spontaneous speech at the end. Each of these data types will now be described in turn.

The word list

Wells's (1982) standard lexical set represents the range of words that share a particular vowel found in RP: 28 words were selected from this list to cover all English vowels in South African English (see Appendix 1). These were followed by two 10-digit numbers covering the full numerical range, the first of which resembled a South African mobile telephone number (0829513647) (this was done to test the effect of numerical phrasing on intonation and pronunciation) (All speakers recited the same numbers). The reason for including numbers was because of their obvious importance in the voice recognition field, in terms of dates, flight numbers and prices etc.

The phonetically rich sentences

Subjects were asked to read six pre-selected phonetically rich sentences taken from the SCRIBE Corpus, which contains a range of sentences consisting of words displaying a wide range of permissible syllables in English: the sentences for the Educated South African English corpus were constructed to cover all possible onsets and possible codas associated with a particular vowel nucleus. The sentences are phonemically balanced, in that the words consist of a wide variety of English phonemes and phoneme clusters (see Appendix 2).

The LOB Corpus extracts

The LOB Corpus consists of 500 text samples, each of approximately 2 000 words, distributed over Fifteen text categories. In addition, these portions are themselves divided into paragraphs, which in news reportage are often delimited by sub-headings. These text categories include press reportage, editorials, reviews, skills, trades and hobbies, popular lore, general fiction, romance and love stories, etc. For our purposes we have thus far drawn only on the first three text categories mentioned, namely press reportage, editorials and reviews, all of which tend to relate to local British and world political and sporting events in the seventies (see Appendix 3 for two short sample texts). In order to follow the sequence exactly, no texts were omitted, despite the fact that some are rather unexciting, to say the least (thus each subject read a different extract, following the LOB sequence). The audio recordings were subsequently divided into the same paragraph-sized segments.

We presented each subject with four A4 pages of text, printed in 14 point Times New Roman. Subjects read the texts in sequence, usually taking approximately ten minutes to complete the task, though reading speed was variable. In order to render the texts reasonably reader-friendly and to assist South African readers, headings and sub-headings were set in bold font, with pages 2 and 4 ending at a sentence break. Capitalisation was restored, all encryptions and codings were removed (e.g. *+ = pound sign and *?2 = acute accent on preceding character) and punctuation and other markings (e.g. inverted commas and tildes) are provided. In the case of *+ we used the word *pound/s* rather than use the pound symbol. In addition, we replaced 's' and 'd' with the words *shillings* and *pence* and have replaced instances such as 3d with *thruppence* (or their equivalents). We also replaced county abbreviations (e.g. *Lancs.*) with the full form (e.g. Lancashire).

Spontaneous speech

After the reading, subjects were interviewed informally in order to generate spontaneous unmonitored speech. The same researcher who conducted the earlier recordings became the interviewer, in order to promote familiarity and to allow the subjects to relax. Subjects were first asked to relate their life stories, in as much detail as they chose, from the time they were born to their current life in the small university town. The initial plan was to follow up the life stories with Labovian triggers (i.e. incidents that had angered them or made them afraid) with the hope of generating anecdotes, but this failed in almost all of the first few cases, possibly due to the unfamiliarity of the researcher with most subjects, and was abandoned in favour of a less personal approach. Thus, in place of recounts of such incidents, the interviewer identified potentially interesting aspects from the initial life stories and asked follow-up questions relating to these. This proved more successful but was unfortunately not as productive as we had hoped. The life stories, however, will nonetheless prove a rich source of data in terms of unmonitored speech, as well as discoursal aspects.

The procedure

Subjects were allocated 45 minute slots for data collection. After a brief orientation with the administrator, who explained the procedure, they completed biographical information sheets and signed informed consent forms (see Appendix 4). They were then given time to familiarise themselves with the printed extract from the LOB Corpus and given a chance to clarify potential pronunciation problems. When they indicated that they were ready, subjects were escorted to the recording room.

The recording room has a covered window and is soundproofed with acoustic

tiles and a double door leading to the outside corridor. Potential sound interference was minimised by using an incandescent light bulb. The microphone (on a stand) was positioned eight to ten centimetres in front of the subject, who was seated facing a double-glazed window into the control room, through which the recording researcher was able to communicate (using gestures) with the subject. Pasted on the wall below this window were the phonetically rich sentences, which subjects were required to read during the second portion of the session. Subjects were first asked to read the 28 words from Wells's lexical set, which were presented individually on flash cards, to avoid the intonation typical of list-reading. The two numbers followed, after which they read the phonetically rich sentences. Subjects then read their four-page extracts from the LOB Corpus. Thereafter, they were joined in the recording room by the researcher, who proceeded with the short interview.

The recordings were made as 16-bit directly, using an AKG C1000 cardoid microphone a Behringer MX1602 mixing desk and a Vestax D90 hard disk recorder in four tracks. A pop filter was used as well. The four types of spoken data were saved as four separate tracks, enabling efficient subsequent separation and downloading using the Cubase software package.

Automatic speech recognition (ASR)

The availability of carefully annotated recorded speech corpora is essential to research into and development of automatic speech recognition technology. Furthermore, the language, dialect and even character of the speech must correspond to the intended research conditions or the application. At present there is no South African English speech database available with which to undertake such research. Hence, an immediate use of the data we are gathering will be to develop baseline South African English large vocabulary speech recognition systems. The phonetic pronunciation dictionary describing the words in the corpus will make use as far as possible of the IPA-based phonetic inventory developed as part of the African Speech Technology (AST) initiative (Roux et al. 2004; Niesler et al. 2005).[2] Baseline acoustic models can then be determined from the audio data, the associated orthographic transcription, and the pronunciation dictionary using the HTK toolkit (Young 2002).[3]

A sub-component of modern automatic speech recognition systems that is known to be critical to overall system performance is the language model. This is a statistical model describing the word sequences that occur in a language. It is able to compare the likelihoods of competing sequences, and thereby guide the speech recogniser in its search for the best transcription of the input speech.

Language models are produced by extracting patterns and tendencies from large

text corpora. Current state-of-the-art language models are trained using plain text corpora, and do not make explicit use of underlying grammatical information. A very small number of corpora are available that have been marked up with *Part-Of-Speech* POS information. Here every word is accompanied by a classification describing its grammatical function (e.g. article or adjective) in the sentence. The production of such tagged corpora is extremely laborious since it requires extensive manual intervention and checking. Some work has been carried out into the use of POS information in language models. However this has been hampered by the lack of audio recordings to accompany the tagged text, with which acoustic models would be built and a speech recogniser could be tested.

Hence, in recording recitations of LOB material, we will be in a position to develop the acoustic models required for South African English, and also to determine baseline n-gram language models from the LOB orthography. Furthermore, we will be able to carry out research into the use of POS by language models and to test the effectiveness of such approaches by means of large-vocabulary speech recognition experiments.

Problems

Several problems which arose during the preparatory phases and the collection of the corpus deserve a brief discussion here. One set of problems revolved around the theoretical issues related to selecting the informants for the study. A second set of problems related to the reading 'behaviour' of informants, and a third set concerned administrative and technical issues.

Why Educated South African English?

Critics might suggest that the decision to describe this particular variety of English was an easy way out, since the variety has already been described and there are other, more 'vibrant' post-1994 varieties emerging from the melting pot of South Africa. Our response is two-fold: (1) there is no database of this variety and it has not yet been fully described using the new techniques which will enable a more nuanced and detailed description, and (2) since it forms the original core from which all other South African English varieties have subsequently evolved, and from which a pan-South African English is likely to develop, it is an excellent place to begin.

The criteria

Branford, writing in 1994, stressed the importance of recognising South African English as an autonomous variety which, although frequently described in terms of its variance from received pronunciation and, he claims, approximately isophonic with received pronunciation (at least in its more conservative forms) is

nonetheless a separate variety. He acknowledges 'traces' (p. 472) of variation within this variety, particularly as regards sex and region, but maintains that 'variation tends to be more obviously a matter of class and upbringing than region' (p. 472) and is modified by the formality of the context as well as the degree of accommodation to the speech of the listener.

In planning who should serve as subjects in this study, it soon became clear that we would indeed need to delimit informants in terms of age, geographical location, education and ethnicity. While it is relatively uncontroversial to select 30 years as a lower age limit, given the recognised variability in accents in the early phases of life, selection of the other criteria are significantly more sensitive, and demand justification. Firstly, geographical location: it occurred to us that our readers might (logically) tend to assume that a sample of 50 speakers, all current residents of Grahamstown, a small university town in the Eastern Cape Province, probably represent Grahamstown English, and not Educatied South African English at all. To counter such an argument, we made an effort to ensure a wide range of geographical backgrounds among our informants (see Table 1). In a sense, they all just 'happened' to be living in Grahamstown at the time of recording, and many had spent many years in other parts of the country. The biographical data of the speakers recorded to date indicate that they tended to come from a fairly diverse spread of backgrounds, with many changing schools several times during their adolescence, and many completing schooling in one place and tertiary study elsewhere. For this reason, points were allocated for each main 'phase' of life (primary school, secondary school and tertiary), and Figure 1 reflects overall proportions at the time of writing (remaining recordings will aim to supplement under-represented areas).

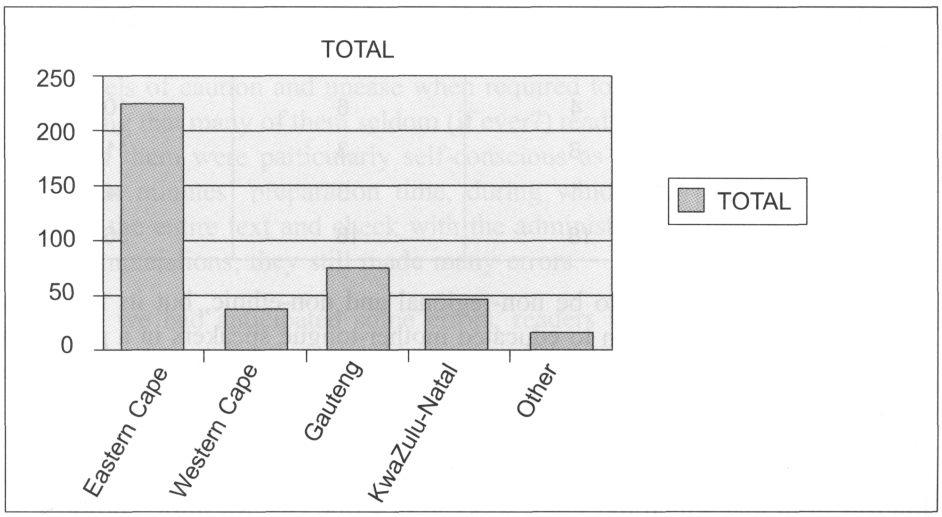

FIGURE 1: Geographical distribution of informants

In addition, we made a special effort to exclude long-term members of the local farming community in the area, many from local 1820 Settler stock, who bear strong traces of what is known as a 'lower Albany' accent (Gough 1995), and in so doing ensured that we sampled a broader, more representative community, possibly not so firmly rooted in a particular part of South Africa. For this reason we made a point of checking which of our subjects had grown up on farms (Q5, Appendix 4), and we excluded those who had.

The second problematic criterion related to education: all informants were required to have completed all 12 years of their schooling and have at least one year of post-school study, all in South Africa, and all through the medium of English. This decision was made in the light of the assumption that it is during schooling that the foundations of any variety are laid, and that some post-school education (minimally a diploma) would enable us to describe the variety as 'educated'. This automatically excluded a huge number of potential informants who have matric certificates, and would, no doubt, see themselves as educated. While we intend no disrespect to them, we were mindful of the fact that a high proportion of South Africans have a matric,[4] and we were aiming for the upper end of the continuum. At the time of writing (36 recorded interviews) 74 per cent (= 28) of all informants had more than one year of post-school education, and 58 per cent (= 22) had postgraduate degrees. Inadvertently, this requirement of tertiary study also automatically excluded a fair proportion of the farming community. Table 1 reflects the age and gender balance of the 36 people recorded so far.

TABLE 1: Age and Gender

Age group	Female	Male	Total
30–39	5	6	11
40–49	4	6	10
50–59	8	3	11
60–69	1	3	4
Total	18	18	36

The corpus was designed to be non-regional and non-ethnic, but its relatively small size and its restriction to educated mother-tongue speakers in a particular age bracket whose parents also had to be mother-tongue speakers resulted in a sample primarily of white individuals. This is therefore an artefact of these criteria but is not inherent to the sample. The fact that Indian speakers were excluded rests on the fact that these communities use a different variety of English from that which the study seeks to describe (see Mesthrie 2002).

Another potential problem related to setting up exclusive criteria to describe

what might seem to be an 'exclusive' variety was the possibility that elitism might result. We acknowledge that this variety already enjoys considerable social prestige (given that its speakers have traditionally held privileged social positions) and, in addition, that our work with this variety may well, unintentionally, result in it being further entrenched, through the use of the label *educated*. It will be interesting, nonetheless, to explore the extent to which this particular variety may be used as a gatekeeper, possibly restricting social and economic mobility in South Africa, despite the country's commitment to socio-economic change. In contrast, it may equally well function as a gate-keeper in the opposite direction, if it is used to identify previously and currently 'advantaged' individuals and exclude them. It must nonetheless be emphasised that the selection of this variety is not intended to strengthen its status or prestige as a norm or benchmark; rather, as stated earlier, we aim to contribute a fuller description of this variety, since it does form the core of all other South African English varieties. We therefore aimed to capture it for posterity and provide a rich description of the way it is currently used: a linguistic snapshot in time, as it were.

Despite these stringent criteria, we still expected to encounter some speakers with accents which might not conform to the overall group accent which we had set out to describe, and we hoped that such idiolects would not skew the data too much, and would simply exemplify the natural 'outliers' in any statistical sample. It remains to be seen whether our analyses uphold this assumption.

Reading problems

Our aim has been to collect fluent, read speech samples corresponding as closely as possible to the LOB text, but it soon became evident that these educated people did not read as fluently as we had expected. Several of them manifested high levels of caution and unease when required to read while being observed, suggesting that many of them seldom (if ever?) read aloud for any purpose at all. Many of them were particularly self-conscious as a result, and, despite being given ten minutes' preparation time, during which they were asked to read through the entire text and check with the administrator if they were unsure of odd pronunciations, they still made many errors.

Because we had anticipated some errors, readers were instructed that, should they make a mistake, they should pause and then restart from the beginning of the sentence. Such restarts could then subsequently be easily removed from the audio without affecting the integrity of the overall recording. However, readers did not always follow this instruction, usually because they were not aware that they had made an error, or possibly did not regard their dysfluency as a true 'error'. There were several departures from the original text, and each of these had to be marked by the recording researcher so that they could be removed by

editing the audio files. In addition to the sentence restarts, the following types of reader dysfluency were encountered:

- Word corrections, when the reader repeated the most recent word or two, in order to correct a pronunciation or misreading (this occurred frequently despite the explicit instruction to restart at the beginning of the current sentence).
- Hesitations, when a pause or silence was inserted at an unexpected point in the sentence, usually because the reader was about to pronounce a difficult word or was anticipating what came next.
- Deletions, insertions and substitutions, where the reader did not read the LOB text precisely, and did not attempt a correction.

Table 2 shows the number of times each of these errors was encountered per 1 000 words of LOB texts in the first 2.3 hours of recorded speech.

TABLE 2: Average number of hesitations, sentence restarts, corrections, and word deletions, insertions or substitutions per 1 000 words of LOB text

Sentence restart	1.9
Word correction	2.7
Hesitations	9.8
Deletion, insertion or substitution	3.3

Hesitations, sentence restarts and word corrections were removed by audio editing. Word insertions could not always be corrected in this way due to strong cross-word coarticulation (the table indicates only insertions that could not be removed). Deletions and substitutions could never be removed from the audio and therefore required the LOB text to be adjusted correspondingly. Often such (incorrect) read text remained syntactically permissible, as illustrated by the following examples (the second version in each case is the read version).

Substitution	labour_NN MPs_NPTS cheered_VBD
	labour_NN MPs_NPTS *jeered*_VBD
Insertion	and_CC not_XNOT attending_VBG
	and_CC *are*_BER not_XNOT attending_VBG
Deletion	placing_VBG particular_JJ stress_NN on_IN
	placing_VBG stress_NN on_IN

Insertions required the correct POS tag to be supplied in the tagged LOB

transcription for the new word, while substitutions sometimes required the word's POS tag to be changed.

Finally, numbers, as found in amounts, dates and quantities, were usually expanded by speakers into their spoken equivalents, and in order to ensure that the LOB transcriptions always matched exactly the words spoken, these required careful checking, as there were several variants,

for example

> on_IN July_NR 17_CD 1953_CD

could be rendered as

> on_IN July_NR the_ATI seventeenth_OD nineteen_CD fifty_CD three_CD

or as

> on_IN July_NR seventeen_CD nineteen_CD fifty_CD three_CD

Even though the LOB text was changed in places as described above, the line numbering of the original corpus was preserved throughout, for ease of reference.

The content of the LOB corpus

Further unanticipated problems relate to the dated content of the LOB Corpus. Firstly, there were many proper names in some of the texts which posed difficulties to our readers by virtue of their 'foreignness': place names such as Leicester and Gloucester, for example, tended to be pronounced as spelt; names of politicians in the 1960s (e.g. Mr David Ormsby-Gore) were not as familiar as they must have been at the time, and there was a surprisingly high occurrence of French words in the corpus. Even in a thoroughly 'South African' text, however, African words (such as *Nqhurha*) would not necessarily have been easier to read.

Secondly, the passages were, with few exceptions, dull and pedatic, more than a little dated (being based on writing in the 1960s) with a strong 'British' overlay which had not been fully anticipated. While we acknowledge that the texts were never written to be read aloud (indeed, some include stock prices and steeplechase results), and some of them are far-removed from the experiences of our informants, the value of keeping the texts the same far outweighed the problems this caused. Nonetheless, although our researcher prepared each informant for this aspect of the procedure prior to recording, and explained the

technical reasons why we were obliged to stick to the texts in the first place, this may well have had an unfortunate effect on the overall reading performance of the subjects.

Final remark

We anticipate completing the collection of the corpus by mid year and publishing the results of initial analyses by year-end. Further papers should follow in due course, hopefully supplemented by the collection of additional corpora of other varieties of South African English for comparative purposes (either by our research group or by fellow researchers in South Africa). In this way, we hope to have stimulated some interest in this avenue of research, and ultimately to make some modest contribution to the understanding of the variety we have chosen to call Educated South African English and the role it plays in South African society.

Appendix 1: Selections from Wells's standard lexical set

PIN	DRESS	TRAP
LOT	STRUT	FOOT
BATH	CLOTH	NURSE
FLEECE	FACE	PALM
THOUGHT	GOAT	GOOSE
PRICE	CHOICE	MOUTH
NEAR	SQUARE	START
NORTH	KIN	FORCE
CURE	HAPPY	LETTER
COMMA		

Appendix 2: Phonetically rich sentences

1. The smell of the freshly ground coffee never fails to entice me into the shop.
2. The government triumphed four years ago and we have every reason to believe that it will triumph again.
3. She flicks through a magazine when she gets a chance.
4. Bright sunshine shimmers on the ocean.
5. Where were you while we were away?
6. Aluminium cutlery can often be flimsy.

Appendix 3: Samples of text for reading from the LOB Corpus

What Labour is Lacking by James Beecroft

Labour's cash problems were discussed last night by Mr. Len Williams, the

party's national agent and deputy general secretary. The party's national executive, he said, was considering ways of increasing Labour's income. But whatever was done, the party would never have funds on the Tory scale. Mr. Williams was talking to more than 300 young socialists attending their organisation's national rally at Skegness, Lincolnshire. He stressed that last year 213,000 pounds of the Labour Party's 250,000 pounds income was contributed by the trades unions. The average contribution from individual party members, he said, was only 4 shillings a year. 'Even with the support of the unions,' he went on, 'the amount of money we have today is not sufficient for our party to do its job adequately. Most of the constituency parties are always short of cash. Many of them are in debt for the last election.' The trades unions, said Mr. Williams, had not only been the main financial support of the Labour Party – they had been, through their steadiness, 'the ballast which has kept the ship upright in heavy seas.'

12 Minutes of the Duke on TV

The Duke of Edinburgh made a twelve-minute appearance on BBC television last night – and looked more relaxed than his interviewer, Richard Dimbleby. It was the Duke's first interview on British TV and he came across like an unflurried man having a cosy fireside chat. This pre-recorded interview was for the weekly programme 'Panorama'. It was concerned with the commonwealth technical training week which opened yesterday. The aim of the week's campaign is to draw attention to the need for technical training. Twenty-eight Commonwealth countries are taking part and in this country 188 local councils have helped to arrange special events to boost the campaign. In his TV interview the Duke was obviously enthusiastic about the whole project.

Four men accused of bank robbery

Cardiff magistrates yesterday rejected an application that two of four men in the dock should be allowed to have their hands free and not handcuffed to one another. The four men were charged jointly with breaking and entering Lloyds Bank in Cardiff between January 14 and January 16 and stealing 9,465 pounds and other property including watches and jewelry. Before the court were: Colin David Baldwin, aged 26, of Braunton Avenue, Llanrumney, Cardiff; Albert Augustus King, aged 32, of Southmead, Bristol; Maurice Charles Harry, aged 32, of Northam Avenue, Llanrumney, Cardiff; and James Bernard Powell, aged 32, of Penarth Road, Cardiff.

Appendix 4: Biographical data form

Please fill in the missing information, or circle, where appropriate:

1. **Your parents:** **First language** **Place of birth**
   Mother ....
   Father ....

2. **Gender:** Female Male

3. **Your age group** under 30 30–39 40–49
   50–59 60–69 over 70

4. Where did you go to **school**? If you attended more than one school, please indicate the number of years in each case:

	School name	**Town/city**	**Number of years**
School 1	.....	.....	.....
School 2	.....	.....	.....
School 3	.....	.....	.....

5. Have you ever lived on a farm? Yes No
   If yes, for which years? (e.g. 1976–1986) .....

6. What is your highest education qualification? .....

7. Where did you complete **your tertiary training**?

	Name of Institution	**Number of years**
Institution 1	.....	.....
Institution 2	.....	.....

8. Have you ever lived for longer than 1 year outside South Africa?
   Yes No

9. If you answered yes to Question 8:
   Name of country ..... Which year(s)? .....

10. What is your current occupation? .....

Notes

1 We have selected from a range of the various lexical sets proposed by Wells. The standard lexical sets consist of 'groups of words which tend to share the same vowel' (Wells 1982, xviii), and reference to a particular set enables one to refer to the particular vowel shared by the members of the lexical set.

2 The African Speech Technology Project represents the only substantial set of databases (i.e., speech recordings and their transcriptions) that is useful for speech recognition work. However it is telephone (and not wideband) speech, and consists of

isolated, unrelated sentences. Our database will consist of wideband speech in coherent paragraphs of substantial length.

3 Hidden Markov model Toolkit. An open-source collection of software used to develop speech recognition systems (available at htk.eng.cam.ac.uk). Its use is widespread and hence systems we develop will be familiar to other researchers around the world.

4 In South Africa learners typically spend 12 years at school, and a 'matriculation certificate' is the standard school-leaving qualification.

References

Branford, W. B. 1994. English in South Africa. In *The Cambridge history of the English language*. Vol. V, *English in Britain and overseas: Origins and development*, ed. R. Birchfield, Cambridge: Cambridge University Press.

Census 2001: http://www.statssa.gov.za/publications/populationstats.asp (accessed 7 December 2005).

de Klerk, V. 2003. Towards a norm in South African Englishes: The case for Xhosa English. *World Englishes* 22 (4): 463–481.

Garside, R., G. Leech, and G. Sampson, eds. 1987. *The computational analysis of English: A corpus-based approach*. London: Longman.

Lanham, L. 1996. A history of English in South Africa. In *Focus on South Africa*, ed. V. de Klerk, 19–34. Amsterdam: John Benjamins.

Lass, R. 1990. A 'standard' South African vowel system. In *Studies in the pronunciation of English: A Commemorative volume in honour of A. C. Gimson*, ed. S. Ramsaran, 272–285. London: Routledge.

——. 1995. South African English. In *Language and Social History: Studies in South African Sociolinguistics*, ed. R. Mesthrie, 89–106. Cape Town: David Phillip.

Mesthrie, R. 1992. *English in language shift: The history, structure and sociolinguistics of South African Indian English*. Cambridge: Cambridge University Press.

——. 2002. From second language to first language: Indian South African English. In *Language in Social History* (2nd edition), ed. R. Mesthrie, 339–355. Cape Town: David Phillip.

Romaine, S. 1984. *The language of children and adolescents*. Oxford: Blackwell.

Roux, J. C., P. H. Louw and T. R. Niesler. 2004. The African Speech Technology Project: An assessment. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC). Lisbon.

SCRIBE http://www.phon.ucl.ac.uk/resources/scribe/scribe-manual.htm (accessed 8 December 2005).

Seppe, S., J. Maxim and B. Wells. 2000. Prosodic variation in Southern British English. *Language and Speech* 43:309–334.

Niesler, T. R., P. H. Louw and J. C. Roux. 2005. Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. *Southern African Linguistics and Applied Language Studies* 23 (4): 459–474.

Watermeyer, S. 1996. Afrikaans English. In *Focus on South Africa*, ed. V. de Klerk, 125–148. Amsterdam: John Benjamins.

Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland. 2002. *The HTK Book* (version 3.2.1). Cambridge: Cambridge University Press.