Towards a Corpus of South African English: Corralling the Sub-varieties[*]

Leela Pienaar, *Extended Studies Unit (l.pienaar@ru.ac.za)*
and
Vivian de Klerk, *Dean of Students (v.deklerk@ru.ac.za),*
Rhodes University, Grahamstown, Republic of South Africa

Abstract: Within the last twenty years, the use of a corpus for language research has become the *sine qua non* in many areas of linguistic enquiry. This trend is particularly evident in lexicography, a discipline which has become increasingly and overtly 'corpus-driven'. This article draws on research from a Master's project which involved the collection of a small corpus of Indian South African English (ISAE), an acknowledged component or sub-variety of South African English (SAE). The discussion highlights the importance of aiming for a balanced representation of the known sub-varieties of a language when compiling corpora for lexicographic and linguistic investigation. Since ISAE is primarily an oral dialect, specific focus is given to the methodological challenges involved in compiling a spoken corpus. Methodological insights from local as well as international corpus research were used to guide and inform the process. These include the Xhosa English Corpus, the New Zealand Corpus of Spoken English and the Hong Kong Corpus of Conversational English. The various stages in the research process are described, together with explanations of how problems such as ways of corpus design, the selection of corpus contributors, the data-collection process and developing guidelines for consistency during the corpus compilation were addressed. The article provides a keyhole view of the main lexical and syntactic features of ISAE exemplified in the corpus and juxtaposes these against the backdrop of general SAE and trends in World English. The article concludes with a proposal for the collection of parallel corpora of other sub-varieties of SAE which will provide an objectively compiled repository of language in use to enable researchers to discern the linguistic features at the core and periphery of SAE. It is argued that the establishment of corpora of the various known sub-varieties of SAE could constitute an important step towards the creation of a truly representative large corpus of SAE and ultimately towards a better definition and understanding of SAE.

Keywords: CORPUS, SPOKEN CORPUS, DESIGN, SOUTH AFRICAN ENGLISH, SUB-VARIETIES, INDIAN SOUTH AFRICAN ENGLISH, LEXICOGRAPHY

Opsomming Op weg na 'n korpus van Suid-Afrikaanse Engels: Die bymekaarbring van die subvariëteite. Gedurende die laaste twintig jaar het die gebruik van 'n

[*] This article is a revised version of a paper presented at the Thirteenth International Conference of the African Association for Lexicography, organized by the Bureau of the *Woordeboek van die Afrikaanse Taal*, Stellenbosch, Republic of South Africa, 30 June–2 July 2008.

korpus vir taalnavorsing die *sine qua non* op baie gebiede van taalondersoek geword. Hierdie neiging is veral te sien in die leksikografie, 'n vakgebied wat toenemend en merkbaar "korpusgedrewe" geword het. Hierdie artikel maak veral gebruik van navorsing vir 'n Meestersprojek wat die versameling behels het van 'n klein korpus Indiese Suid-Afrikaanse Engels (ISAE), 'n erkende komponent of subvariëteit van Suid-Afrikaanse Engels (SAE). Die bespreking beklemtoon die belangrikheid om na 'n gebalanseerde weergawe van die bekende subvariëteite van 'n taal te streef wanneer korpusse vir leksikografiese en linguistiese ondersoek saamgestel word. Omdat ISAE primêr 'n mondelinge dialek is, word spesifiek gefokus op die metodologiese uitdagings gepaardgaande met die samestelling van 'n gesproke korpus. Metodologiese insigte van sowel plaaslike as internasionale korpusnavorsing is gebruik om leiding en vorm aan die proses te gee. Dit sluit die Xhosa English Corpus, die New Zealand Corpus of Spoken English en die Hong Kong Corpus of Conventional English in. Die verskillende stadiums in die navorsingsproses word beskryf, saam met verduidelikings van hoe probleme soos maniere van korpusontwerp, die keuse van korpusbydraers, die dataversamelingsproses en die ontwikkeling van riglyne vir konsekwentheid gedurende die korpussamestelling gehanteer is. Die artikel verskaf 'n intieme blik op die belangrikste leksikale en sintaktiese eienskappe van ISAE soos beliggaam in die korpus en plaas dit teen die agtergrond van algemene SAE en neigings in Wêreldengels. Die artikel sluit af met 'n motivering vir die versameling van parallelle korpusse van ander subvariëteite van SAE wat 'n objektief saamgestelde bron van taal in gebruik sal verskaf om navorsers in staat te stel om taalkundige eienskappe in die kern en periferie van SAE te onderskei. Daar word geredeneer dat die totstandbrenging van korpusse van die verskillende bekende subvariëteite van SAE 'n belangrike trap kan vorm tot die skep van 'n werklik verteenwoordigende groot korpus van SAE en uiteindelik tot 'n beter omskrywing en begrip van SAE.

Sleutelwoorde: KORPUS, GESPROKE KORPUS, ONTWERP, SUID-AFRIKAANSE ENGELS, SUBVARIËTEITE, INDIESE SUID-AFRIKAANSE ENGELS, LEKSIKOGRAFIE

South African English in the canon of World Englishes

Recent thinking on the global uses of English has acknowledged that there is not just one English language but rather a family of 'World Englishes'. The umbrella-term 'World Englishes' provides a conceptual framework to accommodate the different varieties of English which have evolved from linguistic cross-fertilization caused by colonization, migration and trade, which resulted in the transplantation of the original 'strain' or variety. Linguistic models such as Kachru's concentric circle model (1985) have done much to legitimize varieties of English around the world. Although the model has been lauded for acknowledging the pluralistic nature of English, it has also been criticized: firstly, for entrenching an essentially Anglocentric view of English by situating historically-native varieties of English (e.g. British and American) at the centre while relegating other varieties to the periphery, and secondly for failing to capture the dynamics of exchange that occur between varieties in the different circles.

In multilingual South Africa where English is one of 11 official languages, the model treats South African English as a unified whole and does not adequately take account of the socio-linguistic complexities of the South African situation and of the different varieties of English that have evolved as a result. A closer look at English in South Africa (henceforth 'South African English' or 'SAE') reveals a complex situation where the English used here is anything but monolithic. Its speakers include those for whom it is a first language, those for whom it is an additional language and those for whom it is a replacement language. Census data which elicit information about home language do not tell the whole story as they fail to capture such complexities. For example, data derived from Census 2001 which merely indicate that 8.2% (roughly 3.7 million people) of the population of 44 million are English mother-tongue speakers, do not acknowledge that roughly 45% of the South African population (with different home languages) also use English to varying degrees in the domains of government, education, commerce, industry and in the media.

Linguistic research into SAE has identified various sub-varieties: Afrikaans English (Watermeyer 1996), Black South African English (Gough 1996), Cape Flats English (Malan 1996), Xhosa English (De Klerk 2002a, 2006) and Indian South African English (Mesthrie 1996), and has hinted at the existence of others. However to date there has been no real attempt to explain the dynamics either between the constituent sub-varieties themselves or between them and the variety of English spoken as a mother tongue in South Africa. There are ideological problems too with marking the sub-varieties of SAE with racial qualifiers (such as Black, Indian or Coloured) as it entrenches their status as ethnolects that are 'other' while leaving the 'colonial' variety of English unmarked. It has been argued that this practice actually affirms the position of a variety that is spoken by less than 10% of the population and sets it up as the standard against which all other English varieties in South Africa are measured (De Kadt 2001, cited in Coetzee-Van Rooy and Van Rooy 2005). In South Africa's rapidly-changing linguistic environment it is therefore useful to take a more realistic and inclusive view of SAE. Therefore any attempt to document and define SAE should strive to incorporate all known sub-varieties in a balanced way in order to provide a reliable representation of this Southern Hemisphere variety of English.

Corpora

One of the ways of creating a balanced and representative sampling of a language is through the establishment of a corpus. The term *corpus* is used here to mean an organized 'collection of pieces of language text in electronic form, selected according to external criteria to represent as far as possible a language variety as a source for linguistic research' (Sinclair 2005: 16). Thus defined, the use of a corpus for language research has become the *sine qua non* in many areas of linguistic enquiry (including lexicography) within the last twenty years and there are corpora for a host of major national as well as minority languages, ranging from Arabic to Walloon. In addition, there are corpora for different varieties of English: the British National Corpus (BNC) and the Bank of

English (BOE) in Britain; the American National Corpus; the Wellington Corpora of Spoken and Written New Zealand English; the Australian and the Macquarie Corpora; the Kolhapur Corpus of Written Indian English; and the enormous International Corpus of English (ICE) which, when completed, will comprise parallel corpora of regional varieties of World English (Meyer 2002). Yet, to date, there is no large corpus to represent South African English (SAE). A South African component of the ICE project (henceforth ICE-SA) has been in preparation since the early 1990s (Jeffery 2003) but the corpus has not yet been completed or released to the research community. Constructed according to ICE specifications, ICE-SA will be a relatively small contemporary corpus of 1 million words (500 texts of 2 000 words each) and is unlikely to be able to provide a balanced representation of the different varieties of English in South Africa. A corpus which is to serve as standard reference for a language would need to be much larger in order to reflect its sub-varieties and the domains in which these are used. In tacit acknowledgment that SAE is a constellation of various sub-varieties, De Klerk (2002a: 35) has argued for differentiated corpora of SAE on the grounds that 'linguists need … a database which carefully distinguishes speakers of English on the basis of their background MT [mother tongue], ethnicity and geographical location'. To this end, she pioneered the development of a 500 000-word spoken corpus of Xhosa-English which was completed in 2005.

The idea of creating a corpus of Indian South African English (ISAE) was conceived in the wake of the development of the afore-mentioned corpus of Xhosa English. The corpus of ISAE, which involved the collection of conversational data from a narrow age band, is intended as the first building block towards a full corpus of ISAE that will ultimately be demographically and contextually balanced. When completed, the full corpus of ISAE could serve a dual purpose: firstly as a standard for referencing research into the sub-variety; and secondly as a component in a truly representative large corpus of SAE. The collection of spoken ISAE is envisaged as the foundational section towards the construction of a fuller corpus of ISAE, and despite its modest size, it could provide useful initial data for comparison with earlier studies of ISAE and Indian English worldwide. When complete, the full corpus of ISAE would constitute an important building block in a comprehensive corpus of SAE, with the sub-varieties represented in equitable ratios. A SAE mega-corpus thus constituted could represent a valuable standard reference for determining the salient features of this important variety of world English. Sub-corpora such as the ISAE data at the centre of this study would provide ready linguistic repositories for testing theories of language variety and for assessing the effect(s) of the country's official policy of desegregation since 1994 on ethnically-based taxonomies of SAE.

It is hoped that such a corpus, which encompasses all the known sub-varieties of SAE, would create a more nuanced understanding and definition of this important variety of World English. This enterprise will obviously depend on collaboration amongst South African researchers.

What is ISAE?

Indian South African English or South African Indian English (henceforth ISAE) refers to a variety of English spoken by the 1.1 million (or 2.5% of the population) South Africans of Indian descent (Census 2001). The majority (71.6%) of this self-classified group reside in KwaZulu-Natal, descendants mainly of indentured labourers, together with some traders and missionaries who arrived in the former province of Natal between 1860 and 1911. The original immigrants were a linguistically diverse group from different geographical regions in India, and very few had a command of English. However, just over 100 years after the arrival of the first Indian immigrants to South Africa, English has ousted the ancestral Indian languages and become a replacement first language or mother tongue for this community. Mesthrie describes ISAE as a complex example of a 'language-shift' variety of English where English replaced the Indian languages 'as the main (and often sole) language of daily interaction' (Mesthrie 1992b: 3). According to the 2001 census 95.8% of Indian South Africans listed English as their first language or home language. The shift to English by this community is linked to various factors such as the lack of a common Indian language, diminished contact opportunities with India during the apartheid regime, the desire for economic advancement through proficiency in English, and finally, an education system that did not support the vernacular needs of minority communities (Mesthrie 1992b: 32-33).

ISAE as a sub-variety of SAE is well-documented (Bughwan 1970, Crossley 1987 and Mesthrie 1992a, 1992b, 1996, 2002a, 2002b). It is an ethnolect of SAE spoken by South Africans of Indian extraction. It is largely distinct from 'Indian English' as spoken on the Asian sub-continent, the latter being broadly characterized by ornate lexis and stylistically formal constructions (Kachru 1994). ISAE has retained residues of lexis and syntax rooted in the ancestral Indian languages, but it shares several features with other sub-varieties of SAE. ISAE has absorbed lexical items such as *robot* (traffic light), *dagha* (mud), *babalaas* (a hangover) and *tickey-line* (cheap or of poor quality) from general SAE; and in turn ISAE has enriched the lexis of general SAE with contributions such as *bunny-chow* (a hollowed out half-loaf of bread filled with curry), *char-ou* (Indian person), *larney/lahnee* (one's boss or a wealthy person), *ballie* (an old man or person) (Silva et al. 1996). ISAE also features additional senses for general English words (Mesthrie 1992a), as the following sentences illustrate:

> My uncle's got *sugar* (= diabetes).
> I went to visit my *future* (= fiancé or fiancée).
> She's so *independent* (= haughty or aloof).

In the absence of sufficient contextual information to prime comprehension, an outsider to the ISAE speech community would be challenged clearly to discern the sense of the lexical examples *sugar*, *future* and *independent* quoted above.

For the most part, the lexis of general ISAE does not feature the numerous

Afrikaans-based items such as *handlanger* (an untrained assistant), *lappie* (a rag) and *skelm* (a rascal), which are common in several other sub-varieties of SAE, as its geographical base has been KwaZulu-Natal, where English rather than Afrikaans has dominated in official and public domains. The notable exception to this generalization is the slang register in ISAE where the shift away from community-based norms is discernible in the liberal use of Afrikaans-based lexis such as *ou* ('chap'), *graaf* ('work'), *lakker* (from *lekker*, 'nice'), and *vaai* (from *waai*, 'go'), alongside Zulu-based words such as *mache* ('money' from *amatshe* meaning 'stones'), *chebe* ('a beard', from *intshebe*), *gane* ('a child', from *ingane*), *skatul* ('a shoe', from *isiscathulo*), and *pozi* ('a house', from English army slang 'pozzie', a shortening of 'position', which could conceivably have been a dug-out or shelter). There are also a few slang lexical items traceable to Indian languages such as *mota* ('rich', from Hindi *mota* meaning big or fat) and *ballie* ('old man', from Hindi *balig* meaning an adult). As with many other forms of slang (Burchfield 1985), ISAE slang is governed by gender and age boundaries with usage prolific in the speech of young males (under 25 years of age), but extending to older males from socio-economic groups and occupations which favour a very informal style of speech.

While the lexis of ISAE was influenced to some extent by contact with local languages such as English, to a lesser extent Afrikaans, and the pidginized Fanagalo used on the mines, it was preserved and fossilized by the social isolation caused by the South African government's apartheid policies enforced between 1948 and 1994.

Methodology used to create the corpus of ISAE

Structure and design

There is no 'one-size-fits-all' corpus design, as each corpus is determined by socio-linguistic factors relating to the population under consideration and by the purpose which the corpus will serve. In planning this module of the ISAE corpus, the design features of significant earlier corpora such as the BNC and ICE, as well as corpora of the spoken language such as the LLC, the New Zealand Spoken Component of ICE (ICE-NZ), the Hong Kong Corpus of Conversational English (HKCCE) and the Xhosa-English Corpus (XE Corpus) were explored to find a suitable framework.

Size and boundedness of the corpus

There is also no ideal corpus size, only an optimum corpus size determined by the research needs and pragmatic considerations such as the availability of resources. In terms of size, the corpus for this research did not strive to be in the same league as the mega-corpora of hundreds of millions of words such as the BNC (100 million words), the Oxford English Corpus (1 billion words) or the continually growing BOE Corpus (450 million words). It is a small corpus of finite length, rather than a large unconstrained or continually-growing monitor corpus. As a sample corpus collected within a narrow age and education band, any inferences drawn from the results would need to be interpreted with those parameters in mind.

Since the research aim was to collect a corpus of conversations, the ICE specification of 2 000 words for private, direct conversations between two people was taken as a useful benchmark of size (Nelson 1996: 29). With one researcher carrying sole responsibility for all the conceptual and labour-intensive aspects of supervising corpus collection and transcription, a scaled-down version of the ICE conversational component of 180 000 words was realistic. Thus a feasible target for this module of the ISAE corpus was a third of that in size or 60 000 words. Apart from the time- and labour-intensive aspects of converting spoken data into machine-readable form, there are other arguments in favour of modest-sized, well-balanced corpora. De Klerk (2003: 467), quoting McCarthy (1998), argues for smaller, well-designed corpora 'of spoken material which contain authentic and reliable representative data, [that] can be analyzed exhaustively in a variety of ways'. It is tempting to accumulate vast quantities of data on the assumption that corpus analysis is largely computerized, but Kilgarriff et al. (2004: 106) caution against accumulation of a welter of data that makes even the analysis of simple features like word occurrence difficult and time-consuming: 'If there are five hundred [occurrences of a word], [analysis] is still a possibility but might well take longer than the editorial schedule permits. Where there are five thousand, it is no longer viable. Having more data is good — but the data then needs summarizing.' (It must nonetheless be acknowledged that software is available which sifts and organizes data semantically, thus simplifying the searching of huge quantities of data.)

The 60 000-word corpus of ISAE is made up of thirty 'texts' or speech samples of approximately 2 000 running words or 'tokens'.[1] Following the ICE model, texts of 2 000 running words constitute the building blocks of the ISAE corpus, as they provide reliable linguistic samples for analysis, while being manageable in size. In fact, Biber and Finegan (1991: 212-213) maintain that a component of even half that size (1 000 words) is adequate to deliver data that will reveal the main linguistic characteristics in a text. Each ISAE text segment is a self-contained unit of roughly 2 000 words extracted from one thirty-minute dialogue, rather than a composite constructed from several short verbal exchanges.

Type of corpus: spoken rather than written

This research selected spoken English as the starting point because ISAE is 'primarily [an] oral dialect' (Mesthrie 1992b: 35). Previous research into ISAE has also focused on the spoken variety of the language (Bughwan 1970, Crossley 1987, Mesthrie 1992, 1996). In the broader South African linguistic context, there has been a proposal to collect spoken corpora for nine of the official languages of South Africa (Allwood and Hendrickse 2003) and in terms of SAE in particular, much research has already been done towards a corpus of spoken Xhosa-English (De Klerk 2002a, 2003, 2006) which, it is hoped, will ultimately form part of a larger corpus of Black South African Englishes. Viewed against these national linguistic research initiatives, this first building block towards a corpus of ISAE could facilitate comparative studies of different sub-varieties of spoken English in the South African context and Indian English worldwide.

Internationally there are more examples of written than spoken corpora and even in the BNC the bias towards written data as opposed to spoken data is in the ratio of 9:1. Leech et al. (2001: 1) explain that this imbalance was a result of practical considerations. They acknowledge that the spoken language is 'the primary channel of communication', and that on these grounds it should have been allocated a greater proportional share of the corpus. However, they explain that this was not done because 'it is a skilled and very time-consuming task to transcribe speech into the computer-readable orthographic text that can be processed to extract linguistic information'. Compiling a corpus of spoken language is comparatively more difficult, labour-intensive and expensive than compiling a similarly-sized corpus of written language. The reasons for this are located in the basic differences between speech and writing. Writing is already in a mode visible for study, but speech (an audio medium) has to be converted to writing (a visible medium) before it can be studied and analyzed. Casual spontaneous speech is also 'messy' and not well-behaved syntactically: incomplete sentences are the norm, as are false starts, latched (or simultaneous) utterances and hesitations.

For the corpus of ISAE, the transfer of the spoken data to the written mode involved listening to the recording several times and manual word-for-word transcription. Although speech-recognition technology has been developed to handle the automated transcription of formal, clearly articulated speech such as broadcast monologues and dialogues, as yet there is no reliable program to deal with the unpredictable nature of spontaneous speech. Transcribing the recordings therefore was the most arduous part of the research project with a 2 000-word text segment taking roughly fifteen hours to transcribe, mark up with simple annotations and proof-read. The ICE-USA team have reported similar experiences of 15 to 20 hours from transcription to proof-reading of a 2 000-word multi-party conversation (Meyer 2002: 71).

Classifying 'text types' or genres for spoken corpora

In addition to written corpora being more numerous, there are also established systems for classifying written data. Although there is no agreed taxonomy for categorizing genres of spoken language, two broad approaches to classifying spoken data for corpora exist: one demographically-motivated and the other context-governed or task-oriented. The BNC, for example, distinguishes between private conversation (40%) and public, task-oriented aspects of speech (60%), and classifies the latter in four domain-specific areas, designated educational and informative, public and institutional, business, and leisure. Within each domain, verbal interactions are identified as being either monologues (such as lectures, speeches, sermons) or multi-party activities (such as classroom interactions, meetings, chat shows) (Leech et al. 2001: 2-3). In the conversational component, the speech interactions are all spontaneous and informal with a demographically-motivated approach controlling variables such as social class, gender, age and geographical distribution across samples. From the BNC experience it would appear that both approaches are valuable for determining text types for spoken corpora. De Klerk (2002b: 27) recommends the context-governed approach for classifying spoken text types, on the grounds that it strives for a balance 'between speaker, environment, context and recurrent features' and because it facilitates subsequent analysis from 'different [speaker and contextual] perspectives'. This argument holds, provided that demographic considerations are also accommodated within the defined linguistic contexts.

Since such detailed sub-types would have been impractical in a small corpus, as they would not have generated sufficient data for generalizable linguistic patterns or the formulation of reliable conclusions, a decision was taken to confine data-collection to one demographic band, namely 'young adults'. The spoken data comprised only casual face-to-face conversations between two people. The research took casual conversation as a starting point because firstly, spontaneous informal dialogues in private settings exemplify the kind of naturally-occurring language that everyone engages in daily. Casual conversation has also been described as the quintessence of language, a kind of 'pre-genre' in the development of language since all other forms of language, whether spoken or written, trace their genesis to this genre (Swales 1990). Cheng and Warren (1999: 6–7), in their study of inter-cultural conversations of Hong Kong English, argue that 'conversations are a benchmark for other spoken discourses, and that by more fully describing conversational English … we will better understand the ways in which other spoken discourses differ from it'. Secondly, in the case of ISAE, it is in informal, private settings, rather than in public speech situations that the features of ISAE are most observable. In this regard Mesthrie observes that 'in public it is the ISAE accent which is its clearest marker; but in private situations or informal situations involving ISAE speakers mainly, the lexical carry-over and use of basilectal syntax increases' (Mesthrie 1992: xviii).

The contributors to the corpus

There were 49 South African-born contributors to the corpus, all of whom were of Indian extraction. Contributors supplied biographical details and information about their linguistic background on the *Personal Details and Consent Form* (Appendix A). Although the questionnaire did not use the racial labels designed by the apartheid government (Black, White, Coloured and Indian), respondents indicated their alignment with the group 'Indian' by selecting the substrate Indian language or cultural group with which they identified (Question 8 Appendix A). The corpus excluded anyone who had not been born or raised in South Africa, such as Indian nationals and members of the Indian diaspora in general. The corpus also excluded anyone who had spent more than 12 months outside South Africa within the last three years in order to eliminate linguistic features which might be the result of recent contact with other languages or other varieties of World English (Question 5 Appendix A).

The first small group of data collectors and contributors to the corpus comprised family members and friends of the researcher, all of whom were students at Rhodes University. Since ISAE is not generally used in public discourse and members of this speech community tend to adopt 'more careful and formal styles in public interactions' (Mesthrie 2002a: 341), the fieldworkers used their access to existing social networks to identify other contributors to the corpus. This measure was guided by significant earlier research (Gumperz 1970, Milroy 1987) who recommend that a researcher who might be perceived as an outsider should avoid interaction with the targeted social group, and who also found that using a member of the 'in-group' was effective in securing access to a range of vernacular and non-standard codes which are often eschewed in groups specifically constituted for research and observation. In other research, Schmied (1996: 186) refers to the 'famous sociolinguistic paradox' of the observer effect where the presence of a researcher who is perceived as an outsider to the 'in-group' causes the participants in an observed conversation to speak in ways that are not natural, in a bid for standard or prestigious forms. In a further attempt to secure naturalistic data, field-workers were instructed not to structure the social interactions as interviews as they could potentially control the discourse and determine the elicitation topics and techniques. Instead, field-workers maintained their status as members of the 'in-group' by functioning as active participants with equal speaker rights. It is hoped that all these precautions assisted in modifying the observer effect and contributed to the procurement of representative linguistic data.

Substrate language groups represented in the corpus

In order to yield a reliable sample, the corpus was structured to be proportionally representative of the five main Indian language groups found in South Africa. Although many terms are shared by all groups, other terms especially those from culinary, kinship, clothing and religious domains are specific to different linguistic or cultural groups. There are no recent national statistics for ancestral language affiliation to Indian language groups in South Africa but by using the 1960 census records as a reference point it was possible to establish very broad guidelines for symbolic attachment to the main language groups. The year 1960 appears to have been a linguistic watershed for the Indian community in South Africa as there was a marked decline in the use of Indian languages from that date (Mesthrie 2002b: 165). Thus the ancestral language distribution in the corpus was closely aligned to language data recorded for Indian South Africans in the 1960 census (Table 1).

Table 1: A comparison between the 1960 census and the substrate Indian language groups represented in the ISAE Corpus

Language	1960 census %	ISAE corpus %	ISAE corpus actual number
Hindi	32%	37%	18
Tamil	36%	27%	13
Gujarati	14%	14%	7
Telugu	9%	10%	5
Urdu	9%	4%	2
Other	0.5%	8%	4

Gender

In order to avoid a gender bias in the corpus and facilitate future comparisons of language use in equally-weighted gender group configurations, attention was given to achieving a 50:50 gender distribution. In addition, conversations between same-sex dyads and mixed-sex pairs were distributed as follows: ten conversations between women only, ten between men only and ten in mixed gender groups.

Age and geographical distribution

The contributors to the corpus were all young adults ranging in age from 18 to under 29, making it a highly-focused collection of ISAE speech with potential value for comparison with similarly-profiled corpora, such as the locally-collected Xhosa English Corpus (De Klerk 2002a, 2006) and the Corpus of London Teenagers (COLT) abroad (Stenström et al. 2002). All contributors to the corpus were born and educated in KwaZulu-Natal, the province with the largest concentration of South African Indians (Census 2002). Previous significant studies on the use of English by Indian South Africans have all used population samples from KwaZulu-Natal (Bughwan 1970, Crossley 1987 and Mesthrie 1992, 1996). This research has the potential, therefore, to provide useful data for comparison with these studies.

Time frame for data collection

With the research focus on *contemporary* ISAE, the time frame for data collection was limited to eighteen months (October 2004 to April 2006) in order to build a reliable synchronic corpus and thus exclude, or at best minimize, variables related to language change. This is even narrower than the time-frame of five to ten years for synchronic corpora recommended by Meyer (2002: 46).

Equipment and recordings

The conversations were recorded on small, battery-powered analogue tape recorders with built-in flat microphones which were unobtrusive, unintimidating and manageable in informal, private settings. The decision to make analogue recordings rather than digital ones was a pragmatic one, influenced by financial constraints and technological availability. If required, at a later stage the analogue recordings could be digitized for preservation using specially designed software, such as Syntrillium's 'Cool Edit' program. [2] Out of a total of 37 recorded conversations, only 30 were eligible for inclusion in the corpus and in retrospect, the audio quality would have been enhanced if participants had been wired up with small lapel microphones, as has been successfully implemented in other research, notably COLT (Stenström et al. 2002).

Transcription and storage

Contributors to the corpus were anonymous, so each speaker was assigned a core identity number from 1–49 (labelled $01–$49) and biographical details such as gender, age group and ancestral language group were encoded together with the speaker numbers. Thus $10M1H would be interpreted as follows: $10 = core speaker identity number; M = male; 1 = age group 16–19; H = Hindi. In line with the practice followed by the LLC (Svartvik 1990), pseudonyms of equivalent gender and number of syllables were substituted for personal names of third parties, addresses, telephone numbers and names of clubs or groups mentioned in the conversations. However, it was not deemed necessary to protect the identities of figures which exist in the public domain. The principle guiding these decisions was to disguise only details traceable to private individuals. The thirty recordings used in the corpus were given file numbers ranging from #01 to #37, and each file was prefaced by additional header information with encoded details about the material in the file:

FILENAME: (e.g. #21)
RECORDING: (e.g. 10B)
DATE RECORDED: (e.g. 25/09/2005)
DATE TRANSCRIBED: (the date when the transcription was completed e.g. 7/10/2005)
NO. OF WORDS: (e.g. 2 014)

This header information is stored separately, as it is not part of the speech text itself, but it does provide a useful 'handle' for identification and retrieval of material.

In line with other corpora such as the XE Corpus, ICE-SA and the proposed spoken language corpora for the nine official African languages of South Africa, data was orthographically transcribed with no prosodic or phonetic mark-up. Transcription was done in lower case with no punctuation apart from question marks, apostrophes for enclitic forms and possessives and hyphens for hyphenated terms. The initial transcription used word-processing software (MS Word) and converted the data to plain text to ensure electronic compatibility between different operating systems and different programs. There was minimal annotation to ensure that the corpus files are not 'bloated' and that the raw corpus is available in a simple form. The following notation conventions were used:

= incomplete words e.g. *wed* for 'wedding'
⟨,⟩ short one-second pause
⟨,,,⟩ pause over three seconds
[sniffs] or *[doorbell rings]* non-verbal features essential to making sense of the conversation
⟨*??*⟩ unfamiliar words for which an approximate spelling was used
[unclear] undecipherable utterances
21 numbers transcribed in full e.g. 'twenty one'

The conversation was laid out like the script of a play with each speaker-turn on a new line and changes in speaker-turns sequentially numbered in multiples of *5* to allow for corrections or insertions during the checking phase. Latched utterances were enclosed between brackets thus: <{> </}> and each speaker's words within the overlapping segment were enclosed by brackets <[></]>. This system of mark-up does not strive for iconicity in transcription but preserves each speaker-turn as a unit, while using the mark-up to indicate the position and extent of the overlap (Nelson 1996: 41).

The following excerpt is a typical extract of a transcription taken from the ISAE Corpus:

⟨#03:$07F2G:930⟩ but like, I was gonna wear it today, but look at the weather it's like, <{><[>warm.</[>
⟨#03:$09F2T:935⟩ <[>tomorrow's twenty one.</[></{>

Analysis and findings

Wordsmith Tools was used to produce alphabetical and frequency-based word lists of the corpus data. Despite the absence of lemmatization, it was possible to discern inflected forms of words from the alphabetically-arranged lists. The frequency-based lists yielded information on the scope and range of vocabulary within the ISAE corpus, while at the same time delivering data to facilitate objective comparisons with other spoken corpora. Mesthrie's extensive research into ISAE (1992a, 1992b, 2002 etc.) has established it as a distinct variety of SAE, so features selected for analysis were referenced firstly against prior research into the sub-variety (Mesthrie 1992a, 1992b, Crossley 1987 and Bughwan 1970) and then against the *DSAE Hist*. (Silva et al. 1996). Recurrent features in the corpus for which no explanation or discussion could be found in ISAE- or SAE-related literature were investigated in terms of the context in which they manifested themselves and according to the type of spoken language they typified. These investigations involved research into slang, discourse markers, the language of adolescents and briefly, and admittedly, only superficially, diachronic linguistics.

Although small (approximately 60 000 words), and representing a narrow age band of young adults, the resulting corpus of spoken data confirmed the existence of robust features identified in prior research into the sub-variety. These features include the use of *y'all* as a second person plural pronoun, of *but* in a sentence-final position, and of *lakker* /'lʌkə/ as a pronunciation variant of *lekker* (meaning 'good', 'nice' or 'great'). An examination of lexical frequency lists revealed examples of general South African English such as the colloquially pervasive *ja*, *bladdy* (for bloody) and *jol(ling)* (for partying or enjoying oneself) together with neologisms such as *eish*, the latter previously associated with speakers of Black South African English. An extraction of frequency lists facilitated cross-corpora comparisons with data from the BNC and the Corpus of London Teenage Language and similarities and differences were noted. The study also used discourse analysis frameworks to investigate the role of high-frequency lexical items such as *like* in the data. In recent times, *like* has emerged globally as a lexicalized discourse marker, and its appearance in the corpus of Indian South African English confirms this trend.

Future developments

In considering what the ideal speech corpus should be like, Williams (1996) makes several important recommendations which have relevance for the addition of 'building blocks' to extend this modest initiative towards a corpus of ISAE. In an ideal world, a fundamental consideration is that the speech corpus should include a range of speech forms (e.g. monologues as well as dialogues) contributed by a demographically representative range of speakers (taking account of age, gender, geographical location and occupation) across a range of styles and functions. In addition, it is vital that future spoken language corpora ought to be available in at least two forms: audio and written, to take account of the fact that the transcription of speech in orthographic form results in the loss of much essential information. The recordings should be publicly available, together with different versions of the transcript (all in electronic form): orthographic, grammatically tagged, as well as one with prosodic mark-up. In addition, the corpus should be phonetically segmented and labelled. Naturally, developing such a corpus is a formidable task and 'a major undertaking' (Williams 1996: 19), and for this reason, building a corpus is usually a team endeavour requiring the kind of investment in terms of human resources and capital outlay that is beyond the scope of single individuals, and even well-resourced university research departments. The scale of such an undertaking requires the support of large agencies with appropriate financial, human and technological resources. In addition, since the development of a corpus is usually a lengthy process, it requires a long-term commitment to language research and development. Successful international precedents for this type of collaboration are the BNC, created by an academic/industrial consortium which included Lancaster University, Oxford University Press, Longmans, the British Library and the British Academy, and the BOE which is jointly owned by the University of Birmingham and the publishing house HarperCollins. In the South African context, such an initiative could involve government structures such as the Pan South African Language Board (PanSALB) in combination with university-based language research centres and business enterprises such as publishing houses.

In the case of SAE, what is needed is a system that will acknowledge the permeability of the boundaries between what has been defined as 'Inner Circle' and 'Outer Circle' varieties. Lee (2001) and Nelson (2006) have successfully demonstrated the value of conceptualizing parallel corpora of varieties of World English as a set of overlapping Venn diagrams in order to discern the essential items at 'the core' of World English and the items which radiate out towards 'the periphery'. This model has potential relevance for SAE, which is composed of several sub-varieties and where the notion of what constitutes the standard is constantly under review. It could also provide useful methodology for editors and lexicographers whose job it is to decide on usage norms and the degree of assimilation of various lexical and syntactic options. However, the key to harnessing that methodology is the establishment of parallel corpora of the existing and emerging sub-varieties of SAE. The idea of establishing exactly what constitutes 'the core' of SAE and noting the degree of closeness or distance of different lexical items and grammatical features from this core, seems to offer a really objective method of classifying constituent sub-varieties of the language.

Notes

1. A token is 'an individual occurrence of any word form' (Barnbrook 1996: 53).
2. Share-ware freely available from http://www.syntrillium.com/cooledit/index.html

References

Allwood, J. and A.P. Hendrickse. 2003. Spoken Language Corpora for the Nine Official African Languages of South Africa. *Southern African Linguistics and Applied Language Studies* 21(4): 189-201.

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to Computer Analysis of Language.* Edinburgh: Edinburgh University Press.

Bughwan, D. 1970. *An Investigation into the Use of English by the Indians in South Africa with Special Reference to Natal.* Unpublished Ph.D. Thesis. Pretoria: University of South Africa (UNISA).

Burchfield, R. 1985. *The English Language.* Oxford: Oxford University Press.

Census 2001: http://www.info.gov.za/otherdocs/2003/census01_key.pdf

Cheng W. and M. Warren. 1999. Facilitating a Description of Intercultural Conversations: The Hong Kong Corpus of Conversational English. *ICAME Journal* 23: 5-20.

Coetzee-Van Rooy, S. and A. van Rooy. 2005. South African English: Labels, Comprehensibility and Status. *World Englishes: Journal of English as an International and Intranational Language* 24 (1): 1-19.

Crossley, S. 1987. *The Syntactic Features of South African Indian English among Students in Natal, with regard to Use and Attitudes towards Usage*. Unpublished M.A. Thesis. Durban: University of Durban-Westville.

De Kadt, E. 2001. *What's in a Name? Labelling English in South Africa*. Paper presented at the 8th International Association for World Englishes Conference, Potchefstroom, Republic of South Africa, 1 December 2001.

De Klerk, V. 2002a. Starting with Xhosa English ... Towards a Spoken Corpus. *International Journal of Corpus Linguistics* 7(1): 21-42.

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

De Klerk, V. 2003. Towards a Norm in South African Englishes: The Case for Xhosa English. *World Englishes. Journal of English as an International and Intranational Language* 22(4): 463-481.

De Klerk, V. 2006. *Corpus Linguistics and World Englishes: An Analysis of Xhosa English*. London/New York: Continuum.

De Klerk, V. (Ed.). 1996. *Focus on South Africa*. Varieties of English around the World 15. Amsterdam/Philadelphia: John Benjamins.

Gough, D. 1996. Black English in South Africa. De Klerk, V. (Ed.). 1996: 53-77.

Greenbaum, S. (Ed.). 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gumperz, J.J. 1970. Sociolinguistics and Communication in Small Groups. Pride, J.B. and J. Holmes (Eds.). 1972. *Sociolinguistics: Selected Readings*: 203-224. Harmondsworth: Penguin.

Jeffery, C. 2003. On Compiling a Corpus of South African English. *Southern African Journal of Linguistics and Applied Language Studies* 21(4): 341-344.

Kachru, B. 1985. Standards, Codification and Sociolinguistic Realism: The English Language in the Outer Circle. Quirk, R. and H.G. Widdowson (Eds.). 1985. *English in the World: Teaching of Learning of Language and Literature*: 11-16. Cambridge: Cambridge University Press.

Kachru, B. 1994. English in South Asia. Bolton, K. and B. Kachru. 2006 *World Englishes: Critical Concepts in Linguistics*. *Volume 2*: 255-310. London: Routledge.

Kachru, B.J., Y. Kachru and C.L. Nelson (Eds.). 2006. *The Handbook of World Englishes*. Malden, MA: Blackwell.

Kilgarriff, A., P. Richly, P. Smrz and D. Tugwell. 2004. The Sketch Engine. Williams, J. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 105-116. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Lee, D.Y.W. 2001. Defining Core Vocabulary and Tracking its Distribution across Spoken and Written Genres. *Journal of English Linguistics* 29(3): 250-278.

Leech, G., P. Rayson and A. Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow/London: Pearson Education.

Malan, K. 1996. Cape Flats English. De Klerk, V. (Ed.). 1996: 125-148.

McCarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Mesthrie, R. 1992a. *A Lexicon of South African Indian English*. Leeds: Peepal Tree Press.

Mesthrie, R. 1992b. *English in Language Shift*. Johannesburg: Witwatersrand University Press.

Mesthrie, R. 1996. Language Contact, Transmission, Shift: South African Indian English. De Klerk, V. (Ed.). 1996: 79-98.

Mesthrie, R. 2002a. From Second Language to First Language: Indian South African English. Mesthrie, R. (Ed.). 2002: 339-355.

Mesthrie, R. 2002b. Language Change, Survival and Decline: Indian Languages in South Africa. Mesthrie, R. (Ed.). 2002: 161-176.

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

Meyer, C.F. 2002. *English Corpus Linguistics: An Introduction*. New York: Cambridge University Press.

Milroy, L. 1987. *Observing and Analyzing Natural Language: A Critical Account of Sociolinguistic Method*. Oxford: Blackwell.

Nelson, G. 1996. The Design of the Corpus. Greenbaum, S. (Ed.). 1996: 27-53.

Nelson, G. 2006. The Core and Periphery of World Englishes: A Corpus-based Exploration. *World Englishes. Journal of English as an International and Intranational Language* 25(1): 115-129.

Schmied, J. 1996. Second-Language Corpora. Greenbaum, S. (Ed.). 1996: 182-196.

Silva, P., W. Dore, D. Mantzel, C. Muller and M. Wright (Eds.). 1996. *A Dictionary of South African English on Historical Principles*. Oxford: Oxford University Press.

Sinclair, J.M. 2005. Corpus and Text: Basic Principles. Wynne, M. (Ed.). *Developing Linguistic Corpora: A Guide to Good Practice*: 1-16. Oxford: Oxbow Books. http://ahds.ac.uk/linguistic-corpora/ [Accessed 18 May 2008].

Stenström, A., G. Andersen and I.K. Hasund. 2002. *Trends in Teenage Talk*. Amsterdam/Philadelphia: John Benjamins.

Svartvik, J. (Ed.). 1990. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.

Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Watermeyer, S. 1996. Afrikaans English. De Klerk, V. (Ed.). 1996: 99-124.

Williams, B. 1996. The Status of Corpora as Linguistic Data. Knowles, G., A. Wichmann and P. Alderson (Eds.) 1996. *Working with Speech: Perspectives on Research into the Lancaster/IBM Spoken English Corpus*: 3-19. London/New York: Longman.

Appendix A

Personal Details and Consent

1. Gender

□ Male □ Female

2. Age group:

□16–19 □20–24 □25–29 □30–34 □over 35

3. Were you born in South Africa?

□Yes □No

Province:... Town/City:..

Where did you grow up? ....

4. Where did you go to school ?

Province:...

Name of school(s):..

Highest standard or grade passed: ....

5. Have you spent more than 12 months in total overseas in the last 3 years?

□Yes □No

If yes, please state which country ....

6. Which language does/did your <u>mother</u> use <u>most often</u> in your home?

□Tamil □Telugu □Hindi □Gujarati □Urdu □English

□Afrikaans □Other (please specify)..

7. Which language does/did your <u>father</u> use <u>most often</u> in your home?

□Tamil □Telugu □Hindi □Gujarati □Urdu □English

□Afrikaans □Other (please specify)..

8. Which language/cultural group do you identify with?

□Tamil □Telugu □Hindi □Gujarati □Urdu

□Other (please specify) ....

9. Which language(s) did <u>you</u> first speak at home? (You may tick more than one).

□Tamil □Telugu □Hindi □Gujarati □Urdu □English

□Afrikaans □Other (please specify)..

10. Which language(s) do you <u>still speak</u> at home?

□Tamil □Telugu □Hindi □Gujarati □Urdu □English

□Afrikaans □Other (please specify)..

11. Apart from English and Afrikaans which languages can you <u>read and write</u>?

□Tamil □Telugu □Hindi □Gujarati □Urdu

□Other (please specify) ....

I give permission for the recording of my voice to be included in a corpus of South African English (which may be released on CD) to be used for linguistic research.

Signed: .... Date:..