Towards a corpus of Indian South African English (ISAE): an investigation of lexical and syntactic features in a spoken corpus of contemporary ISAE.

A thesis submitted in fulfillment of the requirements for the degree of
Master of Arts

Rhodes University

by

Cheryl Leelavathie Pienaar

December 2007

Dedication

For my grandmother Celestine who graduated cum laude from the university of life

Acknowledgments

I wish to acknowledge the part played by my team of fieldworkers who exploited their social networks and persuaded their friends to have their privacy invaded by the tape-recorder. For ethical reasons the fieldworkers and conversation participants must remain anonymous, but I am fully cognizant of the great debt of gratitude that is rightfully theirs. Without their co-operation, there would be no corpus of ISAE and no thesis to describe any findings.

This thesis developed tentacles which drew in unsuspecting family members and friends. My niece Nadine, secured an entrée for me into a social group that would normally be out of bounds for someone of my vintage. My children Kiran and Ashwin, tolerated their mother’s abstracted ways with astonishing good humour, and provided encouragement with strategically-placed chocolates, messages and warm hugs. My WASA mentors and friends gave practical advice, academic support, and loyalty when these were most needed. Dana in particular, switched seamlessly between the roles of consultant, associate and friend and helped me ride through the bumpy existential crisis that beset me midway through the journey. Peter gave his time to read closely and make objective yet affirming comments. My friends in the ISEA rallied round to give desperately-needed collegial support at a crucial point in the production of this thesis.

Finally, my deepest thanks go to my supervisor, Prof. Vivian de Klerk, and to my ‘second reader’, Prof. Ron Simango. Prof. Simango was painstakingly thorough in his reading and gave focused feedback on a range of issues. Prof. de Klerk asked the questions essential for academic rigour, with unflinching clarity of mind and candour. I am privileged to have had my work supervised by such wise minds and generous hearts.

Abstract

There is consensus among scholars that there is not just one English language but a family of “World Englishes”. The umbrella-term “World Englishes” provides a conceptual framework to accommodate the different varieties of English that have evolved as a result of the linguistic cross-fertilization attendant upon colonization, migration, trade and transplantation of the original “strain” or variety. Various theoretical models have emerged in an attempt to understand and classify the extant and emerging varieties of this global language. The hierarchically based model of English, which classifies world English as “First Language”, “Second Language” and “Foreign Language”, has been challenged by more equitably-conceived models which refer to the emerging varieties as New Englishes. The situation in a country such as multi-lingual South Africa is a complex one: there are 11 official languages, one of which is English. However the English used in South Africa (or “South African English”), is not a homogeneous variety, since 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. The Indian population in South Africa are amongst the latter group, as theirs is a case where English has ousted the traditional Indian languages and become a *de facto* first language, which has retained strong community resonances.

This study was undertaken using the methodology of corpus linguistics to initiate the creation of a repository of linguistic evidence (or corpus), of Indian South African English, a sub-variety of South African English (Mesthrie 1992*b*, 1996, 2002). 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, the use of *but* in a sentence-final position, and ‘*lakker*’ /ˈlʌkə/ 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. The frequency lists facilitated cross-corpora comparisons with data from the British National Corpus and the Corpus of London Teenage Language and similarities and differences were noted and discussed. 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.

The corpus built as part of this study is intended as the first building block towards a full corpus of Indian South African English which could serve as a standard for referencing research into the sub-variety. Ultimately, it is argued that the establishment of similar corpora of other known sub-varieties of South African English could contribute towards the creation of a truly representative large corpus of South African English and a more nuanced understanding and definition of this important variety of World English.

Table of contents

Dedication ii

Acknowledgements iii

Abstract iv

List of Tables and Figures xi

Select List of Abbreviations and Corpus Resources xii

Chapter 1: Introduction **1**

1.0 Overview 1

1.1 Corpus Linguistics 2

1.2 Indian South African English 5

1.3 Research goals 6

1.4 Outline of thesis 7

Chapter 2: Corpus Linguistics **9**

2.0 Overview 9

2.1 Corpus linguistics: honing the definition 9

2.1.1 A corpus and other text collections 9

2.2 Precursors of the corpus movement 11

2.3 Chomsky's rational approach vs. the corpus linguist's empirical approach 12

2.4 Corpora in the 20th century 16

2.4.1 Electronic corpora 17

2.4.2 Second-generation machine-readable corpora 19

2.4.3 Recent developments 20

2.4.4 Spoken corpora 21

2.5 The applications of corpus linguistics 23

2.5.1 Corpus linguistics and grammatical studies 24

2.5.2 Corpus linguistics and lexicography 24

2.6 Assessing the corpus linguistic approach 27

2.6.1 Advantages 26

2.6.2 Limitations 27

Chapter 3: Indian South African English 29

3.0 Overview 29

3.1 The family of World Englishes

3.2 A new model: the core and periphery of world Englishes 30

3.3 South African English and its sub-varieties 32

3.4 Indian South African English: the history 33

3.4.1 The arrival of Indian immigrants to South Africa 34

3.4.2 The language situation in South Africa in the late 19th and early 20th century 34

3.4.3 The move towards English 35

3.5 A new sub-variety of South African English: Indian South African English 37

3.5.1 Theoretical classification of ISAE 38

3.6 Features of ISAE 40

3.6.1 Phonetic features 40

3.6.2 Syntactic features 42

3.6.2.1 Topicalisation or topic formation 43

3.6.2.2 Use of conjunctions such as ‘*but*’ in a clause-final position 43

3.6.3 Lexical features 44

Chapter 4: Methodology 48

4.0 Overview 48

4.1 Structure and design 48

4.1.1 Size and boundedness of corpus 49

4.1.2 Type of corpus: spoken rather than written 50

4.1.3 Classifying ‘text types’ or genres for spoken corpora 52

4.2 The contributors to the corpus: who counts as a speaker of Indian South African English? 54
4.2.1 How the contributors were recruited 54
4.2.2 Substrate language groups represented in the corpus 56
4.2.3 Actual language practice 58
4.2.4 Gender 61
4.2.5 Age 61
4.2.6 Geographical distribution 62
4.3 Data collection 62
4.3.1 Time frame for data collection 62
4.3.2 Equipment 62
4.3.3 The recordings 64
4.4 Transcription and storage 65
4.4.1 Speaker profiles 65
4.4.2 Classifying and storing the data collected 66
4.4.3 Transcribing and computerizing 67
4.4.4 Speaker identity and speaker turns 69
4.4.5 Punctuation 69
4.4.6 Pauses 70
4.4.7 Letter abbreviations and numbers 70
4.4.8 Non-verbal sounds 70
4.4.9 Non-fluencies 70
4.4.10 Unfamiliar words and inaudible speech 71
4.4.11 Overlapping speech 71
4.5 Analysis of data 72

Chapter 5: Findings and Discussion 74
5.0 Overview 74
5.1 Conversation trends 75
5.1.1 Speakers and word counts 75

5.1.2 Conversation topics 77

5.1.3 Word frequency 83

5.1.4 Syllable length 87

5.2 ISAE features observed in the corpus 88

5.2.1 The use of y'all as a plural pronoun 89

5.2.2 Copula attraction to *wh*- in indirect questions 92

5.2.3 The use of *of* in partitive genitive constructions 92

5.2.4 Topicalisation 93

5.2.5 '*But*' and '*like*' as reinforcing tags in clause-final positions to mark contrary pre-supposition 94

5.2.6 Associative plural – '*and them*' 95

5.2.7 The use of '*never*' 95

5.2.8 The use of '*it*' with plural noun phrases 96

5.2.9 Near misses 97

5.3 Features of general English 99

5.3.1 The uses of '*like*' 99

5.3.2 Taboo and swear words 102

5.3.3 Innovative swearing 104

5.4 Slang 104

5.5 South African English lexis 107

5.6 Concluding remarks 110

Chapter 6: Conclusion 111

6.0 Overview 111

6.1 Review and summary of the findings 111

6.1.1 The corpus design 113

6.1.2 Reviewing the method of data collection 114

6.2 Limitations of the study 115

6.2.1 Social variables 115

6.2.2 Conversational variables 116

6.2.3 Telephone conversations 116

6.2.4 Collecting data from formal contexts 117
6.2.5 Length of extracts 117
6.2.6 Transcription and mark-up 118
6.3 Suggestions for further research 118
6.3.1 Transcription conventions 118
6.3.2 Email correspondence 119
6.3.3 Written data 119
6.3.4 Speakers' attitudes to their own usage 120
6.3.5 Communities of practice 120
6.3.6 Examining "the core" and "periphery" of SAE 121
6.3.7 PanSALB or business sector involvement 121
6.4 Final remarks 122

References 124

Appendices 134

Appendix A: Personal Details and Consent

Appendix B: Guidelines for Fieldworkers

Appendix C: Background to Recordings

LIST OF TABLES AND FIGURES

TABLES **Page**

Table 1 : Stress placement in polysyllabic words 41

Table 2 : Examples of ISAE lexis of Indian origin 44

Table 3 : Everyday ISAE lexical items with 'restricted usage' 56

Table 4 : A comparison between the 1960 census and the substrate Indian language groups represented in the ISAE Corpus 57

Table 5 : Literacy in Indian languages 60

Table 6 : Corpus contributors according to age and gender 62

Table 7 : Individual speaker word counts 76

Table 8 : Summary of popular conversation topics and associated lexical items 79

Table 9 : Top 100 words in COLT and ISAE corpus 84

Table 10: Comparative frequencies of top-ranking words in successive syllable groups in the ISAE corpus 88

Table 11: Comparison of use of quotative *like* by male and female speakers in the ISAE corpus 101

Table 12: A comparison of the most frequently-occurring swearwords in the ISAE corpus and the BNC 103

Table 13: A comparison of the most frequently-occurring swearwords in the ISAE corpus and COLT 104

Table 14: The five most common slang words in the ISAE corpus 106

FIGURES

Figure 1 : Substrate Language Groups in the ISAE Corpus 58

Figure 2 : ISAE Conversation Topics 78

Select List of Abbreviations and Corpus Resources

Every effort has been made to ensure that the website information provided for the corpus resources cited in this thesis were correct and active at the time of printing. However the researcher cannot guarantee that the sites will remain active or that their contents will remain unchanged and appropriate.

ARCHER	A Representative Corpus of Historical English Registers (copyrighted in-house corpus)
BNC	British National Corpus http://www.natcorp.ox.ac.uk/
BOE	Bank of English Corpus http://titania.cobuild.collins.co.uk/boe_info.html
BSAE	Black South African English
CANCODE	Cambridge and Nottingham Corpus of Discourse English http://www.cambridge.org/elt/corpus/cancode.htm
CHILDES	Child Language Data Exchange Database http://childes.psy.cmu.edu/
COBUILD	Collins Birmingham University International Language Database Project http://www.collins.co.uk/Corpus/CorpusSearch.aspx
COLT	Corpus of London Teenage Language http://www.hf.uib.no/i/Engelsk/COLT/index.html
CSAE	Corpus of Spoken American English http://projects.ldc.upenn.edu/SBCSAE/
DARPA	Defense Advanced Research Projects Agency http://www.darpa.mil/
HKCCE	Hong Kong Corpus of Conversational English http://www.engl.polyu.edu.hk/department/academicstaff/Personal/ChengWinnie/HKCorpus_SpokenEnglish.htm
ICAME	International Computer Archive of Modern and Medieval English http://icame.uib.no/cd/
ICE	International Corpus of English http://www.ucl.ac.uk/english-usage/ice/

ICECUP	International Corpus of English Utility Program http://www.ucl.ac.uk/english-usage/resources/icecup/
ISAE	Indian South African English
Kolhapur Corpus:	http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM#ABI
LLC	London-Lund Corpus http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM
LOB	Lancaster-Oslo/Bergen http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM
MICASE	Michigan Corpus of Academic Spoken English http://quod.lib.umich.edu/m/micase/
OED	Oxford English Dictionary http://dictionary.oed.com/
SAE	South African English
SCRIBE	Spoken Corpus Recordings in British English www.phon.ucl.ac.uk/resource/scribe/
SEU	Survey of English Usage http://www.ucl.ac.uk/english-usage/about/index.htm
TEI	Text Encoding Initiative http://www.tei-c.org/
TLEC	Tswana Learner English Corpus http://ctext.p.nwu.ac.za/ProductsCorporaTLEC.html
XE	Corpus of Xhosa-English http://www.ru.ac.za/academic/departments/linguistics/research.html

Chapter 1: INTRODUCTION

1.0 Overview

For many scholars, the use of a *corpus* for language research has become the *sine qua non* in linguistic enquiry within the last twenty years. By way of introduction to a more finely-tuned definition, a *corpus* can be described as a collection of texts representing either “the language as a whole or … some linguistic genre” and in that way constituting a “source of evidence for research on the language” (Sampson and McCarthy 2004: 1). This specially constructed body of evidence is called a corpus, deriving directly from the Latin *corpus* (plural: *corpora*) which means “body”. In linguistics, the practice of using a corpus to study language is called “corpus linguistics”.

Corpus linguistics is not a separate paradigm in the way that sociolinguistics and psycholinguistics can be regarded as paradigms within the discipline of linguistics. Rather, it is an empirically-based methodology that involves the processing of vast amounts of linguistic evidence in a specially constructed corpus. Although not intended to supplant other systems of linguistic enquiry, the growth in the popularity of corpus linguistics has nevertheless constituted a challenge to theoretically-based linguistic methodologies. From a purely ideological perspective, the empiricist approach associated with corpus linguistics has been regarded as being diametrically opposed to methodologies that are the outcomes of rationalist constructions. However, this polarization of the linguistic terrain is a simplification that obscures the interdependence of both positions. As an empirically-based methodology, corpus linguistics has the potential for application in a variety of linguistic spheres since it does not delimit its own area of operations. It is capable of providing the evidence to complement and enhance theories that have been arrived at through introspective and descriptive methodologies (Chapter 2).

The rest of this introductory chapter will enlarge on corpus linguistics by indicating the range and variety of corpora that exist worldwide. I shall then introduce the variety of

English known as “Indian South African English” and motivate the value of corpus linguistics as a method of inquiry suitable for its investigation. Finally, I will articulate my research goals and give an outline of the plan of this thesis.

1.1 Corpus Linguistics

To expand on the earlier definition of corpus linguistics, it is useful to cite Kennedy (1998: 7), who describes corpus linguistics as an empirical approach that is “based on bodies of text as the domain of study and as the source of evidence for linguistic description and argumentation”. The “bodies of texts” to which Kennedy refers may comprise either spoken or written language, or a combination of both. There is nothing inherently novel about the use of empirical evidence as a source of “description or argumentation” in the study of language. It is well known that Samuel Johnson used “corpus-type” techniques as early as 1755 to collect language data for lexicographical purposes. What distinguishes modern corpus linguistics from earlier efforts such as Johnson’s is the manner in which the corpus is compiled, the fact that the modern corpus is designed to serve ultimately as a model of reference for the language under investigation and finally, the use of technology to explore the vast volume of data collected. In order for the corpus to provide reliable evidence for broader linguistic description and generalization, it is necessary that the data constitute a balanced and representative sample of the language or language variety; that the corpus be a pre-determined size; that the data contained within it be capable of electronic processing; and that ultimately the corpus should constitute a standard reference of the language or language variety under consideration (McEnery and Wilson 1996:21-24). In order to meet the criterion of representivity, corpora are therefore typically vast, with one million words making up a corpus of modest size. This in turn explains why the data have to be processed electronically.

In historical terms, the compilation of the first significant corpus in the twentieth century was pioneered in 1959 by Randolph Quirk who initiated the establishment of the Survey of English Usage (SEU) Corpus of spoken and written texts. This was followed shortly afterwards by the Brown corpus (1961) of written American English and its trans-

Atlantic counterpart, the Lancaster-Oslo/Bergen (LOB) corpus (1978) of written British English texts. Each of these corpora consisted of one million words of running text selected as representative samples of American and British English respectively (Kennedy 1998). Lexicographic needs for corpora that would deliver a greater quantity and spread of lexical items spurred the development of ever-larger corpora. In response to this demand, the 20-million word Birmingham corpus (1987) was developed by Sinclair and the 30-million-word Longman Lancaster corpus was developed by Summers and Leech (Meyer 2002).

Technological progress in the electronic industry in terms of enhanced computer capacity and the speed of data processing have impacted positively on corpus linguistics, advancing its growth as a methodological approach in linguistics. As a result, there are corpora for a host of major national as well as minority languages ranging from Arabic to Walloon. Corpora are constantly being developed and shared in electronic forums such as the website aptly titled http://devoted.to/corpora/. In addition there are corpora for different varieties of English. [1]These include 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 (Greenbaum 1996; Meyer 2002). In addition to the plethora of corpora devoted to geographically-based varieties of English, there are also corpora of specific genres of English, such as the Michigan Corpus of Academic Spoken English, the Corpus of Middle English Prose and Verse, and the ARCHER corpus of American and British English Historical Registers from 1650-1990. Yet, to date, there is no large corpus to represent an important variety of world English *viz.* South African English (SAE). For this reason, even the most scholarly documentation of the history of SAE lexis, *The Dictionary of South African English on Historical Principles* (*DSAE Hist*.) (Silva et al. 1996), had to be undertaken without recourse to an electronic corpus, with both simple and complex queries being directed to a manually-constructed citations

[1] See p xi-xii for the corpora websites

database. The difference between a citations database and an electronic corpus will be explained more fully in Chapter 2 of this thesis. While the observation on the methods used to construct the *DSAE Hist.* does not detract from the quality of the collective scholarship harnessed to produce this dictionary, it does indicate the kinds of technological choices available to this team and how these influenced the time span of 25 years that was devoted to its compilation.

Although no large corpus of SAE exists, the South African component of the ICE project has been in preparation since the early 1990s. However this locally-relevant sub-corpus of ICE is not yet freely available and there is some debate about its adequacy in terms of size and representivity (see 2.4.3). The reservations with regard to representivity are related to arguments that SAE is not a monolithic variety of English, but that it is a dynamic composition of various sub-varieties: for example Indian South African English (Mesthrie 1996), Cape Flats English (Malan 1996), Afrikaans English (Watermeyer 1996), Xhosa English (de Klerk 2002*a*, 2002*b*, 2006) and possibly other varieties yet to be identified. Research outputs from the team responsible for the compilation of the SAE component of ICE have been rather limited and those that have appeared in the public domain have not indicated the extent to which the design of this corpus addresses the crucial question of representivity.

Two recent local initiatives spurred by the need to explore differences and similarities in Black South African English (BSAE) have resulted in the creation of the 200 000-word Tswana Learner English Corpus (TLEC) (van Rooy and Terblanche 2006) and the 500 000-word corpus of Xhosa-English (XE) (de Klerk 2006). The TLEC, which is part of the International Corpus of Learner English, consists of samples of student writing, while the XE corpus comprises spoken data. According to de Klerk, the rationale behind her research is the need for several corpora of BSAE, each designed to represent the English of speakers of the major indigenous South African languages. She proposes that these specialized corpora could be used together to build a balanced corpus of BSAE. She argues that a corpus constituted in this manner would yield valuable information to nicely define the “characteristics of each indigenous Black variety of South African English (e.g. Xhosa English from Sotho English)" without obscuring the "salient differences which might exist" (de Klerk 2002: 26).

Following de Klerk's argument above and noting the developments towards mother-tongue-aligned corpora for BSAE, this thesis motivates for the construction of a corpus to represent Indian South African English, an acknowledged sub-variety of South African English. The thesis will describe collection and distinguishing features in the first building block of this corpus, which has been restricted to the 18-29 year age range. I believe that the establishment of corpora of the various known sub-varieties of South African English could constitute an important step towards the creation of a truly representative large corpus of South African English (SAE) and ultimately towards a better definition and understanding of SAE.

1.2 Indian South African English

Indian South African English (ISAE), as a constituent sub-variety of SAE, is characterized by its own distinctive lexical, syntactical features and phonological features which will be more fully described in Chapter 3. The existence of ISAE as a recognized variety of SAE is well documented (Bughwan, 1970; Crossley, 1987; Mesthrie, 1992*a*, 1992*b*, 1996, 2002*a*). Some earlier research (notably Bughwan 1970) used deficit theories to explain features of ISAE with the focus on a comparison of ISAE with external (usually Standard British) norms of pronunciation and syntax. However, research and insights by Mesthrie (1992*b*, 1996, 2002*a*) which are grounded in sociolinguistic theories of creole studies, represent a significant departure from the deficit view mentioned earlier. Briefly stated, Mesthrie's contention is that ISAE is "a social dialect in its own right, not a substandard variety or a 'bad' approximation of (good) English" (Mesthrie 1992*b*: 220). This view is informed by a study of the historical and social complexities attendant upon the evolution of ISAE and an acknowledgment of the kind of "language-shift" that developed when English replaced the language of a community as "the main (and often sole) language of daily interaction" (*ibid*: 3). It will be clear throughout this study that I have drawn on the insights from Mesthrie's ground-breaking explorations into ISAE as a consolidated reference point for this sub-variety of SAE, and to guide investigations of the corpus data.

Yet despite the various research studies (cited above) into the use of English by Indian South Africans, there has been no effort to establish a systematized searchable collection of the variety of English used by Indians in South Africa: in other words there is no known corpus of ISAE. This thesis has been undertaken with the intention of compiling a collection of spoken data within a selected age range, as the first step towards the establishment of a corpus of contemporary ISAE, which could serve as a standard reference for this sub-variety of SAE. This study will initiate the process of developing a corpus of ISAE by establishing a corpus of spoken texts, because ISAE is "primarily [an] oral dialect" (*ibid*: 35).

The size of the corpus of conversational ISAE underpinning this thesis is 60 583 words, made up of individual speech samples averaging 2 000 running words each. The contributors to the corpus are South African-born Indians who are between 18-29 years of age, who have had 12 years formal schooling through the medium of English in South Africa. The speakers have been selected to represent the major Indian language groups in South Africa *viz*. Tamil, Telugu, Hindi, Gujarati and Urdu, either by virtue of their claim to a working knowledge of these languages or because they have indicated their affiliation to these ancestral languages and cultural groups. Data was collected from speakers residing in KwaZulu-Natal, where the majority of Indian South Africans are located (Census 2001).

1.3 Research goals

The overarching aim of this study is to use the methodology of corpus linguistics to design and establish a 60 000-word corpus of ISAE as exemplified in the informal speech of 18-29 year olds. The corpus is attached in electronic form on the accompanying CD.[2]

[2] The CD is for the examiners, but the corpus will be available through the Department of English Language and Linguistics at Rhodes University: http://www.ru.ac.za/academic/departments/linguistics/

In this respect, the corpus is itself both an outcome of this thesis and the vehicle for possible further outcomes. Although modest in size, the corpus will be used as follows:

1. to identify and describe a selection of lexical and syntactical features in the speech of the young adults represented in the corpus
2. to inform the design of a larger corpus of ISAE that should ideally encompass a wider demographic and contextual range and
3. to motivate for research into the establishment of a series of parallel corpora of varieties of SAE that could facilitate corpus-based explorations into the lexis and syntax at the core and periphery of SAE.

The third goal, *viz.* the idea of using corpus-based methodology to determine the lexical items that constitute the "core and periphery" of a language, is indebted to recent research initiated by Lee (2001) and Nelson (2006) who have demonstrated the value of using parallel corpora to investigate "the core and periphery of world Englishes". This lexically-focused approach has potential value towards a deeper analysis of SAE which exists not in isolation, but in a dynamic relationship with ten other official languages in the country.

It is hoped that the description of current patterns in a corpus of ISAE collected from young adults, who have all had 12 years schooling through the medium of English, will help to confirm or deny the existence of those features which have been identified as typical of ISAE by prior research, together with any other emerging language trends. By anchoring the essential robust features of sub-varieties in this manner, such research could hopefully contribute towards a deeper understanding of the different sub-varieties, and also to an understanding of their shared features.

1.4 Outline of thesis

It will be immediately apparent that this study is not girded by a single theoretical framework. The reasons for this are intrinsically linked to the way corpus linguistics functions as a methodology. The analysis of the data will demonstrate the usefulness of corpus linguistics to a variety of disciplines in linguistics. It is the methodology of *corpus linguistics* that unifies the different threads explored in this study.

Chapter 2 reviews the relevant literature pertaining to corpus linguistics. It sketches the history and development of corpus linguistics and evaluates it as a methodology for different types of linguistic investigation, but particularly for the study of language variety. Chapter 3 describes Indian South African English (ISAE) as a sub-variety of South African English, with its own distinctive lexical and syntactic features and phonology. Chapter 4 discusses the methodology used to create the corpus of ISAE. It also comments on problems encountered at different stages of the research, and explains how these were addressed. Chapter 5 discusses a selection of features identified in the corpus of ISAE. In the last chapter, I assess the study in terms of its contribution to our understanding of ISAE and I also attempt to analyze the study's weaknesses and limitations. In light of these, I hope that the corpus that was produced and which is attached in electronic form, will have some value to the academic community, and that it may go some way towards the establishment of a more comprehensive corpus of ISAE. Ultimately, I hope that this research towards a corpus of ISAE will inspire the creation of parallel corpora of the different linguistic sub-varieties that are present, but not always acknowledged, in the composite that is South African English.

Chapter 2: CORPUS LINGUISTICS

2.0 Overview

This chapter reviews the history and development of corpus linguistics, evaluates it as a methodology for linguistic investigation, and finally it explores its relevance to the study of language variety, in particular the variety known as Indian South African English (ISAE).

2.1 Corpus Linguistics: honing the definition

Corpus linguistics is an empirical methodology in linguistics which uses a large systematically collected, organized body of text (the corpus) to investigate language structure and patterns of usage. As indicated in Chapter 1, the use of a specifically designed corpus intended to "embody" or represent the language variety under investigation is central to this approach. Sinclair (2004*a*) takes the earlier definitions of a corpus further, in his description of a corpus as "a 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". Two points are key in this definition: firstly, the application of external criteria to build the corpus and secondly, the storage of the materials in electronic form. Both points will be discussed more fully later in this chapter and in Chapter 4 of this thesis.

2.1.1 A corpus and other text collections

Considering the first of Sinclair's points *viz*. the corpus as a collection of electronic texts, it is important to consider other electronic text collections in order to determine whether they meet the necessary criteria to be regarded as corpora or not.

The World Wide Web for example, is a convenient electronic information resource and as much of the material it offers involves words, it is tempting to regard it as an off-the-shelf resource for corpus linguistic investigation. However, in terms of linguistic representivity the Web would not qualify as a corpus since it "is not representative of anything other than itself" (Kilgariff and Grefenstette 2003: 1): its exact parameters are unknown and there is no indication of the linguistic population that it represents. In fact even a brief glance at the websites available will reveal that the information posted on the World Wide Web has not been compiled by linguists to serve the goals of their specific discipline (Sinclair 2004*a*). Secondly, a closer look at the contents of the websites will indicate that there is much duplication of material posted on the web. Finally there is no stability to the material posted on the Web, as websites, unless cached, are often all too ephemeral. For all these reasons it would be linguistically naïve to regard the World Wide Web as a ready-made corpus and attempt to extract and apply linguistic generalizations from the welter of material delivered by web searches. However, what the World Wide Web can offer is abundant source material, a veritable "linguists' playground" (Kilgariff and Grefenstette 2003: 13) for the compilation of a linguistic corpus. Existing search engines such as the popular Google, while very powerful, are designed for information retrieval, not for the extraction of linguistic data. To this end, software such as BootCaT[3] (Baroni and Bernadini 2004), are being developed and tested to facilitate the extraction of material to build corpora from the web. The BootCaT designers admit that one of the drawbacks of the program is that "users must possess relatively advanced UNIX command line skills" *(ibid*: 4). However the abiding caveat and one that is conceded by even the most ardent proponents of the "Web as Corpus" argument (Kilgariff and Grefenstette 2003; de Schryver 2002), is that the raw results of Web searches should be cautiously interpreted for linguistic purposes, with the data mined from Web searches sifted according to the corpus compilers' specifically-articulated design criteria.

Similarly, other text collections such as an archive of texts or a quotations database should not be regarded as ready-made corpora either, without further investigating the motivating principles behind the collections. The corpus linguist would need to establish whether the aims and criteria for such collections match the constraints of corpus linguistics. In this regard, an explanatory comment about a quotations or citations

[3] Acronym for ***Boot****strapping* ***C****orpora* ***a****nd* ***T****erms* from the Web.

database is necessary. A quotations database of the sort compiled by historical lexicographers (for example Samuel Johnson's collection of 150 000 quotations on which his *Dictionary of the English Language* (1755) is based, or the *Oxford English Dictionary* database), typically consists of excerpts of materials from larger texts, gathered according to *internal* criteria. Although the universe of texts from which data is excerpted may be specified in advance, the length of the excerpts usually varies from a single isolated sentence to a longer chunk of text, depending on how much context is required to illuminate the word under the lexicographer's microscope (see also 2.5.2). Thus the focus is set at the micro-level of the lexical item, and "word-hunting" for lexical profiling is the lexicographer's object. The categories of text from which the excerpts are taken is determined by the sources in which the lexical items appear and not the other way around. This approach contrasts with corpus linguistic methodology which requires that *external* criteria involve decisions about the text categories that will be used and then ensure that a representative balance of those categories is applied at all data entry points into the corpus. The criteria usually also stipulate exactly which sections of texts will be excerpted and the number of words that will be extracted uniformly from each text. This is done to minimize bias and to ensure that objective methods are used, as the corpus so constituted is required to represent a specific linguistic population. From the above, it will be clear therefore, that not every electronic text collection qualifies as a corpus.

2.2 Precursors of the corpus movement

Although the term "corpus linguistics" is relatively new (just under 40 years), studies based on analysis of real-life collections of linguistic data are not. Roughly contemporaneous with Johnson's lexicographic quotations database, was Alexander Cruden's (1737) Biblical Concordance which used the Bible as a lexical database. The resulting concordance enabled researchers to trace occurrences of any given word in Christian scripture. Following on historically from these early 18th century lexically-related investigations, McEnery and Wilson (1996) quote a variety of empirically-based linguistic studies in the 19th and 20th centuries which can be regarded as precursors of modern corpora. Amongst these are parental diary collections (1876-1926) of child language acquisition, Kading's (1897) collection of 11 million German words to determine spelling conventions, Thorndike's (1921) and Palmer's (1933) vocabulary lists for second language teaching, Eaton's (1940) word frequency lists for comparative linguistic studies and Fries's (1952) descriptive grammar text, based on a collection of transcribed telephone conversations. There are two important points to be made about these early studies. Firstly, they represent a variety of linguistic disciplines: lexicography, biblical studies, language acquisition, lexicology, comparative linguistics and reference grammar. Secondly, all these studies involved the collection and analysis of authentic language, put together in large pre-electronic corpora. This was the accepted methodology for the empirically-based linguistics of the time. For example, in the 1950s the linguists Harris and Hill acknowledged that working with a large body of naturally-occurring or authentic language was both essential and of primary importance, and they regarded reliance on theories or "intuitive evidence [as] a poor second…" (Leech 1991: 8).

2.3 Chomsky's rational approach vs. the corpus linguist's empirical approach

However at the same time, there existed a school of thinking represented by theoretical linguists such as Chomsky (1957), who rejected the empirically-based methodology which used collections of language data to abstract and formulate generalizable linguistic concepts. By contrast, the theoretical or structural linguistic school that Chomsky represented advocated the use of introspection-based rather than empirical methods to generate plausible theories about language structure. In an extension of Sausser's model (1915) of *langue* and *parole*, Chomsky distinguished between linguistic "competence" on the one hand, and "performance" on the other. Chomsky's definition of "competence" is an abstract notion of the native speaker's ideal internal model of language, consisting of a set of systematic relational structures between sound and meaning. His definition of "performance" is the realization or expression of language that is observable and manifest in real-life situations: ordinary language use. Chomsky argued that because linguistic competence was essentially a theoretical or abstract construct, the linguist could not study it using empirical means. Furthermore, he argued that since performance data was only the external manifestation of the larger and deeper concept of linguistic competence, at best it could offer was a partial explanation of the abstract model that the linguist was seeking to identify and define.

Chomsky’s approach, while extremely rational, does not take adequate account of the associative relationship between linguistic variation and factors such as geographical distribution, ethnic background, socio-economic class, age group or gender differences. It relegates these variables to the periphery of linguistic investigation, on the basis that they have no major relevance to the core features of language or “universal grammar”, which was the focus of inquiry for theoretical linguists. This is an essentially minimalist view of language that treats observable linguistic variables dismissively, since it regards them as interesting, peripheral factors that just happened to co-occur at the time of “performance”. It does not attempt to deal with them further.

The corpus linguist’s view, on the other hand, strives to encompass all the variables which might potentially exercise influence on language in order to understand language variety better. Furthermore, in contrast to the introspection- and intuition-based methods of the theoretical linguist, the corpus linguist’s approach to the study of language is chiefly empirical, involving as it does, the collection and analysis of authentic language data in a corpus. Fillmore (1992) gives a satirical summary of the opposing viewpoints represented by these two schools of thought, in the following imaginary exchange:

> “Corpus linguist: Why should I think that what you tell me is true?
>
> Theoretical linguist: Why should I think that what you tell me is interesting?” (cited in Meyer 2002: 4).

This fictional conversation illustrates the oppositional stances of the two schools of thought, while illustrating the essential issues for each: the centrality of authentic empirical data and suspicion towards arguments based on intuitively-generated examples on the part of the corpus linguist; and the disdain for linguistic theory drawn from empirically-collected evidence on the part of the theoretical linguist. Furthermore, the school of theoretical linguistics argued that corpora, because of their essentially finite nature, could not handle the infinite generative capacity of human language. Chomsky maintained that because of the boundless creativity of human language, any corpus is necessarily incomplete. Hence he said, “Some sentences won’t occur because they are obvious, others because they are false, still others because they are impolite” (Chomsky: 1962: 159). According to Chomsky it follows therefore, that any corpus will offer an incomplete and limited picture of the language it purports to represent. While it is a truism that no corpus can match the language entirely, this is still a valid challenge. My brief response at this point is that much depends on the sampling criteria used to compile the corpus, as well as on the manner in which the data is exploited. This issue relates to the whole question of balance and representivity in corpus design, which will be explored more fully in Chapter 4.

Chomsky also challenged the value of raw frequency analyses extracted from corpus data. He is reputed to have commented that the sentence *I live in New York* is bound to occur more frequently in a corpus than *I live in Dayton, Ohio* by virtue of the fact that there are more people in New York than in Dayton Ohio (McEnery and Wilson 1996: 12). The counter to that argument is firstly, that corpus linguistics does not concern itself with the frequency of individual sentences, but rather with the frequency of sentence *patterns*. Secondly, in order to yield a worthwhile interpretation of the frequency of the two sentence structures quoted above, corpus linguistic methodology requires that observed frequency results are evaluated within the framework of the expected frequency for the population size of the real world (Stefanowitsch 2005: 295-301). Apart from such obvious considerations, frequency analyses of corpus data are in themselves interesting. In linguistics generally, they can help focus the researcher to investigate precisely *why* some structures occur in a language or language variety and others do not. For example the plural construction *y’all* occurred frequently in the corpus of ISAE (Chapter 5) as illustrated in, “Oh, *y'all* you stayed in Parlock”. The frequency of the linguistic structure *y’all* in the corpus confirms earlier research findings of this as a typical feature of ISAE (Mesthrie: 1992*a*). The endurance of this plural construction in the speech data of the university students who contributed to the corpus, pointed to the need for an exploration of the grammatical and discourse-related functions of the structure (Chapter 5).

In disciplines such as lexicography, frequency counts provide crucial evidence on which to base a host of decisions. At the macro-structural level, the frequency counts influence decisions about which words to include in the dictionary since they indicate the most commonly-used words in a language or in a domain of the language. At the micro-structural level, they assist in determining what specific information should be included in the entry and how it should be ordered. For example, these counts would usefully reveal the most common plural form (*computer mouses* or *computer mice*) or the most frequently occurring spelling form used in writing (*kombi* or *combi* for a minibus in SAE). Frequency studies interpreted together with chronological data reveal the currency trends of lexemes so that information such as *historical*, *obsolete*, or *rare* can be reliably signposted with appropriate labels. At a deeper level, the study of word frequency combined with a study of meaning and context could assist the lexicographer in separating out the senses (or meanings) of polysemous words. This has implications for which senses will be selected to feature in the dictionaries and how these senses will be ordered. Depending on the size and scope of the dictionary, senses with very low frequency counts in a corpus might be excluded from certain dictionaries, as only a very large dictionary such as the *Oxford English Dictionary* (*OED*), can hope to be really comprehensive. The application of frequency studies and the ordering of meanings is illustrated by comparing the way in which the different meanings of the same word are sequenced in general purpose dictionaries e.g. *The Chambers Dictionary* (*Chambers*) and dictionaries organized on historical principles e.g. the *OED*. *Chambers* places the most frequently occurring current popular meaning first, while the *OED* places the oldest meaning first.

These lexicographic decisions rely heavily on objectively-compiled corpus evidence, to the extent that lexicography can be said to be a "corpus-driven" discipline. Landau asserts in unequivocal terms that lexicographers place great value on corpus-based evidence:

> "There is scarcely any area of dictionary work where a corpus cannot provide important evidence for the lexicographer. The corpus is of great value in deciding on the word list and on the form of each entry. It is vitally important in defining,

> first in determining the sense breakdown if the word is polysemous and then in discovering more particularly how best to define each sense" (Landau 2001: 305).

The lexicographer's corpus-driven dependence on corpus data is radically empirical and as such can be said to represent one end of the extreme on the scale of corpora reliance. According to this system of classification, Gast (2006) describes the relationship between corpus data and the practice of linguistics as a continuum which ranges from *corpus-driven* to *corpus-informed.* Most linguistic specialisms which concede the value of exploiting information from corpus data in combination with other methods of investigation and theoretical insights, in order to structure linguistic argument, would be regarded as occupying the large middle ground termed *corpus-based* approaches. Gast makes a further distinction by noting that those approaches which use corpus methodology solely to test linguistic hypotheses are "located even further towards the centre of the empiricist/rational scale" *(ibid*: 115). Gast's classification, while useful in general terms, could lead to needless casuistry if one tried to apply it rigidly, especially to distinguish nicely between last two approaches that he defines. This study, advocating the collection of a corpus of ISAE, acknowledges the theories about language variety which have contributed to the establishment of ISAE as a variety. It therefore positions itself as a *corpus-based* approach to the study of language variety.

Finally, the corpus linguist's theories are based on an examination of naturally-occurring language data which are at once both observable and verifiable, in a way that intuition-based theories are not. Advances in computer technology, particularly the improvements in computer storage capacity and its efficiency in processing vast quantities of complex data have contributed to a resurgence of interest in empirically-based evidence for the study of linguistics.

2.4 Corpora in the 20th century

In 1959 Randolph Quirk initiated the establishment of the SEU Corpus consisting of 50% spoken and 50% written British English data, at the University College London. The million-word corpus (made up of 200 texts of 5 000 words each) was drawn from printed and manuscript sources and included dialogue as well as monologues. The stated goal of the corpus was “no less than to describe the grammatical repertoire of adult educated native speakers of British English” (Svartik and Quirk 1980: 9). Originally compiled on paper and only computerized later, it is correctly termed a *pre-electronic* corpus. As one of the world’s first English language corpora constructed between 1955 and 1985, it represents important foundational work in corpus compilation. It has also provided a reference database for significant linguistic studies such as *A Grammar of Contemporary English* (1972) and *A Comprehensive Grammar of the English Language* (1985) by Quirk et al. The two major publications cited above, arising out of this research, were reference grammars. The research goal driving the compilation of the SEU was clearly normatively intended, in that it was seeking to establish norms on which to base grammatical descriptions of the language. The corpus confined itself to the speech and writing of “educated adults”, and the language samples selected in terms of this criterion are therefore fairly limited in scope as they are all formal and academic in nature. Subsequent research has questioned the notion of the “educated speaker” as a criterion for corpus building on the grounds that it is both vague and subjective (de Klerk 2002). It could also be viewed as educationally elitist as it would exclude sections of the population that do not comply with the prescribed educational norms. Such considerations have influenced later corpus design.

2.4.1 Electronic corpora[4]

The SEU was followed in 1961 by a corpus of “tagged” or linguistically annotated written American English developed by Nelson Francis and Henry Kučera at Brown University (the Brown corpus). This was a significant endeavour in many ways. Being the first machine-readable corpus produced specifically for linguistic research, it was tagged with analytic information about grammar (especially parts of speech) and sentence structure. It also introduced the principle of free access to corpora for the first time. Its trans-Atlantic counterpart the Lancaster-Oslo-Bergen (LOB) corpus of tagged written British English texts followed in 1978. Each of these corpora consisted of one million words of running text samples of American and British English respectively. Although more than a decade separates these corpora, they were compiled along similar lines using

[4] See p. xii for website information

similar text categories. The Brown and LOB parallel corpora have been valuable for enabling comparative studies of American and British varieties of English, through the assessment of features such as high frequency lexical items and grammar in each (Kennedy 1998).

Other corpora belonging to this generation of one-million word corpus studies are the Kolhapur Corpus (1978) of written Indian English, the Wellington Corpus of written New Zealand English (1986) and the Australian/Macquarie Corpus of written Australian English (1986). Apart from being of similar size, they have all been compiled along similar lines to the Brown and LOB corpora, with slight adaptations to accommodate regional availability of data within text genres. The Kolhapur Corpus for example, does not accord the same weight to imaginative prose as the Brown and LOB corpora, which were collected in countries where English is a first language. The compilers of the Kolhapur Corpus have commented that "Inspite [*sic*] of including samples from all the available full length novels, the proportion could not come anywhere near those of the LOB or the Brown Corpus"

In 1975, a team led by Jan Svartik at the University of Lund, began converting the speech components of the SEU into machine-readable form. This corpus, known as the London-Lund Corpus (LLC), comprises 100 text samples of 5 000 words each (87 from the original SEU, supplemented with 13 others). As the texts were carefully marked up for grammar and prosody, they constitute a rich source of data collected over approximately 15-years (1959-1975/6). However, with more than 80% of the texts drawn from the SEU, which represented the speech of academics or people working in an academic environment, they are neither linguistically nor demographically representative (see 2.4 above). Bearing this criticism in mind, the corpus of ISAE which is based on the speech of 18-29 year olds does not claim to be a reference standard for ISAE as it stands. It is intended as but the first step *towards* a fuller corpus of ISAE.

Various specialized corpora have been compiled to facilitate focused linguistic research into selected areas. There are corpora of learner English, such as the Longman Corpus of

Learner English, which represents the English of second and foreign language learners; corpora representing the language of specific age groups, such as the Corpus of London Teenage Language (COLT); and corpora of specific dialects, such as the Corpus of York English (Meyer 2002: 142-150). There are diachronic corpora compiled to facilitate research into language variation across time and genre, such as the Helsinki corpus of Early Modern English (750 to c.1700) and the Representative Corpus of Historical English Registers (ARCHER) (1650-1990) (Kennedy 1998: 40). Finally, the 20 million-word Child Language Data Exchange Database (CHILDES) is a multilingual database of spoken language from children and adults engaged in first or second language acquisition. Although in its entirety the whole database cannot be regarded as a corpus, it is nevertheless a repository from which corpora of language acquisition can be extracted.

2.4.2 Second generation machine-readable corpora

It is customary to conceive of the development and evolution of corpus linguistics in terms of "generations". Thus one can say that the "second generation" of machine-readable corpora became larger and more comprehensive than their pro-genitors the original standard one-million word corpora. The first major mega-corpus was the British National Corpus (BNC) of more than 100 million words collected between 1991 and 1994. This huge undertaking was the result of collaboration between the British government, academic institutions, publishing houses and public institutions. It is a corpus of present-day British English made up of 90% written and 10% spoken data. Despite the immediately apparent imbalance between written and spoken data, the corpus has followed a careful, principled approach in the selection of text categories. For example, the written sources are made up of 75% informative prose and 25% imaginative writing from a variety of contexts, ranging from the formal or technical areas of language use through to the informal. Furthermore, unlike the SEU, the BNC has been demographically balanced, particularly in the speech component (Kennedy 1998:50-54). The selection of contributors to the corpus was guided by the need to reflect the country's demographic features, such as age, social class, gender and to a lesser extent, geographical location (Leech et al. 2001). Work on building the BNC was completed in 1994 and it was released as a finite corpus to researchers in 1995.

In contrast with the BNC, the Bank of English (BOE) Corpus, jointly owned by the University of Birmingham and publishing house HarperCollins, is an ongoing project. Described as a “monitor corpus”, this dynamic corpus is constantly being updated “to reflect the fact that new words and meanings are always being added to English” (Meyer 2002:15). The BOE, which totalled 524 million words in 2004, has been the basis for various dictionaries, e.g. the general-purpose Collins Cobuild and the BBC English Dictionary. For the lexicographer who needs to keep track of subtle shifts in meaning and usage, the dynamic and unconstrained nature of a monitor corpus is an invaluable source of data with a live link to language trends.

2.4.3 Recent developments

In 1990 the International Corpus of English (ICE) project was launched at University College London under the direction of Greenbaum to enable comparative studies of some twenty different world varieties of English. It includes countries where English is either a majority first language (e.g. Canada, New Zealand and Australia), or an official additional language (e.g. India, Nigeria, Hong Kong and South Africa). The project involves the compilation of parallel corpora in specific text categories of both spoken and written English. To date, there are completed corpora available in ICE from East Africa, Great Britain, Hong Kong, India, New Zealand, the Philippines and Singapore.

Schmied (1996:185) notes that while the ICE design categories provide a useful framework for the compilation of a corpus, they need to be adapted for multilingual countries where English is one of many languages used. He points out that in such countries, the use of English might be restricted to formal or public domains, with other languages preferred in more private settings. He also observes that it would not be correct to give the same weight to the various text categories, owing to economic, political and cultural differences. This is certainly true of a country like South Africa, where the costs of book production make books a luxury beyond the reach of the average person. In South Africa, newspapers and periodicals fulfill a whole range of functions, from providing hard news, to being sources of recreational reading.

The South African component of the ICE project, in preparation since the early 1990s, is not yet freely available. If it has followed the ICE text categories without making allowances for local conditions, the above criticisms are worth bearing in mind. There are additional concerns about whether it provides a balanced representation of the different varieties of English in South Africa (de Klerk 2002). In this regard, it has been argued that South African English (SAE) is not a monolithic variety of English, but is made up of various sub-varieties e.g. Cape Flats English (Malan 1996), ISAE (Mesthrie 1996), Afrikaans English (Watermeyer 1996), Xhosa English (de Klerk 2002) and possibly other varieties yet to be identified. de Klerk, in arguing for differentiated corpora of SAE maintains that "What linguists need is a database which carefully distinguishes speakers of English on the basis of their background MT [mother tongue], ethnicity and geographical location" (de Klerk 2002: 35). To this end she pioneered the development of the spoken component of a corpus of Xhosa-English which was completed in 2005.

2.4.4 Spoken corpora

From the brief history of corpora described above, it is clear that all the pre-electronic corpora collections comprised pieces of written text. However the SEU, initiated in 1959, represents a major departure from the tradition of favouring writing over speech, as it consisted of 50% written language and 50% spoken language. This balance was not immediately taken up as a model for the corpus collections that followed. All the significant million-word corpora that were developed in its wake were collections of written language, and even today there are still more written corpora available than speech corpora. Apart from the historical reasons for this imbalance, there are other reasons that relate directly to the difficulty associated with making collections of spoken data. Speech, especially spontaneous speech, with its incomplete utterances, indistinct words and phrases, and simultaneous speech or "latched utterances", is not tractable, making it difficult (possibly also undesirable?) to corral into neat sentences and nicely alternating speech turns. In addition, transcribing speech is a skilled, labour-intensive and expensive undertaking. The developers of speech corpora such the BNC, the Corpus of Spoken Israeli Hebrew, the Wellington Corpus of New Zealand English, the Hong Kong

Corpus of Conversational English (HKCCE) and locally, the Xhosa-English Corpus have all confirmed that these problems have presented huge challenges in their own undertakings. (See also 4.1.2).

Amongst the publicly available spoken corpora that do exist, very few are examples of spontaneous speech. Of these, the Scottish MapTask Database (1992) collected at Edinburgh University, which consists of task-oriented dialogues where one participant was required to reproduce a route on a map by following instructions given by the other participant, is a notable exception. It is cited as an example of "truly spontaneous speech" containing "many examples of hesitation pauses, fillers, false starts and other disfluencies" (Williams: 1996: 11). It is distinguished from other speech corpora of that period such as those produced by the Defense Advanced Research Projects Agency (DARPA) in the USA and the Spoken Corpus Recordings in British English (SCRIBE), which were compiled in order to study phonetic features of speech. The speech databases originating in DARPA and the SCRIBE used either scripted monologues, or isolated sentences and lists of individual words, all of which were perfectly adequate for the kind of accoustic-related language investigations they were compiled to serve.

The London-Lund Corpus (LLC), which was extracted as a sub-set of data from the SEU, is a valuable early example of naturally-occurring language, as it is made up of conversations which were recorded without the participants' knowledge or consent (Svartik and Quirk 1980). It is no longer considered acceptable ethical practice to make such surreptitious recordings, even in the interests of collecting naturalistic data. Current methods of data-collection have had to explore other methods of procuring the desired data (see Chapter 4). However, the LLC is useful as a source of prosodically marked-up conversation data that is available in electronic form. It has been criticized for not including monologues and consequently for featuring a limited range of intonational patterns. Furthermore the recording quality is generally poor, and there are large sections of unclear speech (Williams 1996: 14). By contrast, the Lancaster/IBM Spoken English Corpus (a collaborative project between Lancaster University and IBM technology) meets very high standards of acoustic quality. The aim of the project was to study a range of intonational patterns in British English in order to produce models for artificially synthesized speech. To meet these specific needs, the corpus used carefully-elocuted, rehearsed monologues (from drama, poetry and various radio broadcasts) produced under strictly controlled recording conditions (Taylor: 1996: 21).

An American initiative to build a Corpus of Spoken American English (CSAE) began in the 1990s (Chafe et al. 1991). When completed, it will comprise spontaneous as well as planned speech in the form of monologues and dialogues, in a range of demographic and contextual categories. It promises to be a potentially useful resource, as both the transcribed and the auditory versions will be available for researchers. Furthermore, in order to customize accessibility for different kinds of research interests, the broad as well as the narrow transcriptions (see 4.4.3) will be available. (Chafe et al. 1991: 77).

The Cambridge and Nottingham Corpus of Discourse English (CANCODE), built as a result of co-operation between the two universities, comprises 5 million words of spontaneous speech collected between 1995 and 2000. This corpus is not generally available to the public but it has a unique category which encodes the relationship between the speakers, to indicate whether they are intimates (living together), casual acquaintances, colleagues or strangers. It is believed that this will facilitate an analysis of the influence of level of familiarity on discourse.

2.5 The applications of corpus linguistics

Although all the corpus examples up to this point have made reference to those that have been compiled with the study of English as the focus, other languages have also used the corpus linguistic approach. A specific forum organized in Lancaster in 2001 for “non-English corpus linguistics” showcased corpus-based research in a variety of ancient and modern international languages, such as medieval Irish, Korean, French, Abidjanee French, and Swedish (Wilson et al. 2003).

The popularity of corpus-based research in various fields of linguistic endeavour is undisputed. It has the potential to enrich theoretically-based linguistic approaches through the quantity and quality of authentic data that it can deliver. The following examples will provide an overview of how corpora are being harnessed in different areas of linguistic specialization.

2.5.1. Corpus linguistics and grammatical studies

The link between pre-electronic corpora and reference grammars has already been mentioned (see 2.2). More recently the following projects have drawn extensively on corpora to create reference grammars for learners of English: the Collins COBUILD Project which used the BOE corpus; Greenbaum's *Oxford English Grammar* (1996) which is based on the ICE-GB; and Biber et al.'s *Longman Grammar of Spoken and Written English* (1999) (Meyer 2002: 13-14). While these are all examples of corpus-based general grammar studies, there is also a substantial body of research that has drawn on corpus data to isolate and explore specific grammatical constructions. The International Computer Archive of Modern and Modern and Medieval English (ICAME) based at the University of Bergen, Norway acts as a repository for different corpora. Research generated from these corpora is shared with the scientific community at the annual ICAME conference and in the variety of papers published in the related ICAME journal (http://icame.uib.no/journal.html).

2.5.2 Corpus linguistics and lexicography

For lexicographers, the much larger Birmingham Corpus (1987) developed by John Sinclair and the Longman Lancaster Corpus developed by Della Summers and Geoffrey Leech have proved indispensable in providing a greater quantity and spread of data and lexical items.

Early lexicography dating from Samuel Johnson to James Murray used sets of index cards as their "corpus". However this was essentially a system of 'wordwatching' which involved the collection of citations for words under surveillance. James Murray, in the preparation of the Oxford English Dictionary, employed teams of workers (including his own children) to collect and sort through citation slips to determine usage and frequency (Green 1997). The quotations that were selected often reflected the readers' own bias or interests and did not always result in an objective card collection or a balanced "corpus". Thus Murray's card collection of citations cannot be regarded as a corpus by today's standards. The selection of citations was determined by internal subjective criteria, as opposed to objectively selected external criteria. The criteria for eligibility were not clearly articulated and it seems as though often it was the abstruse words which contributors found striking, that found their way into the card collection and Murray (1879) is quoted as complaining that "The editor or his assistants have to search for precious hours for examples of common words, which readers passed by…Thus, of *abusion*, we found in the slips about 50 instances: of *abuse* not five" (Krishnamurthy 2002).

It was the Collins Birmingham University International Language Database Project (1970) (COBUILD), led by John Sinclair, that ushered in the era of corpus lexicography proper. This project harnessed the computer's capacity in three key areas for lexicography: the storage of vast amounts of data, efficient data retrieval and objectivity in the selection of data for storage. At the time of its completion, with 20 million words, it was regarded as a large corpus. The importance of a large corpus for lexicographical purposes is immediately obvious if one considers that in the LOB corpus of 1 million words, the five most common words are function words: *the*, *of*, *and*, *to* and *a* (Meyer 2002:14). The lexicographer needs to draw on a very large reservoir of words in order to find and analyze sufficient examples of content words, especially rare words.

In terms of efficient data retrieval, the computer has made it possible for the lexicographer to extract concordance lines with ease from vast quantities of text so that word collocations can be analyzed and words can be studied in context. This is important for the lexicographer, who relies on context to discern, disambiguate and define the various "senses" or meanings of a single word. In this connection Moon (1987:87) confirms that the "context provided by the concordance line gives clear signals of meaning in most cases, in particular through syntax and collocation, and an interplay of these permits disambiguation."

In order to explain the need for objectivity in data selection, a distinction must be made between a citation database and a corpus. The difference lies in the manner in which the evidence is collected. In the case of a corpus, entire texts or parts of texts are collected according to strict external design criteria and stored electronically. Brief reference has already been made to examples of the kinds of external corpus design criteria that can be applied in corpus construction (see 2.1.1), but each corpus will require the formulation of its own set of external criteria. This is described and exemplified in terms of the methodology underpinning the Corpus of ISAE (Chapter 4). Early lexicography was based on databases of extracts often subjectively excerpted from larger texts. This system has the potential to yield a biased, unrepresentative database. Working within the discipline of frame semantics, Fillmore’s (1992) survey of 10 monolingual dictionaries revealed discrepancies in the meaning of the word *risk* for example. Upon consulting a 25-million word corpus he found examples of meaning unaccounted for in the dictionaries. His conclusion was that “the citation slips the lexicographers observed were largely limited to examples that somebody happened to notice” (cited in Meyer 2002:17). Fillmore’s comment clearly highlights the internal subjectivity of the material selected for a citations database and reveals its limitations in representing the language or a genre of the language adequately. It is true that the lexicographer still has to interpret the evidence from the concordance lines, but the increased amount of evidence delivered by a corpus diminishes the reliance on subjective intuition. In this respect, Rundell and Stock (1992: 29) maintain that the corpus “enhances our understanding of the workings of language and forces us to rethink our own subjective perceptions”.

In the case of *A Lexicon of South African Indian English* (Mesthrie 1992*a*), the author’s selection is based largely on native-speaker experiences, supplemented by interviews and samples of written material intended for Indian readership. The proposed corpus of conversational ISAE could constitute the first step towards the establishment of a larger corpus of spoken ISAE that is demographically and contextually balanced.

2.6 Assessing the corpus linguistic approach

2.6.1 Advantages

It has already been mentioned that corpus linguistics is not a new branch of linguistics: it is really "descriptive linguistics aided by new technology" (Kennedy 1998: 268). As a methodological approach which does not confine its area of operations to any particular linguistic paradigm, the corpus-based approach can be applied to fit a range of specializations within the discipline of linguistics (see 2.5). Its great value must surely lie in the fact that it investigates authentic language data (see 2.3) and uses the results of these inquiries to define or confirm theoretical constructs. Secondly, since corpora are vast structured repositories of linguistic information, extrapolations drawn from an examination of the material they contain are more likely to be generalizable than those drawn from small, specialized case studies. The value of corpora for studies of language variety is evidenced by the corpora which have already been constructed to study varieties of English (Brown, LOB, Kolhapur, BNC, CSAE, Wellington, COLT) and the ICE (Greenbaum 1996), which will house parallel corpora under one umbrella. However, in order for a corpus to be maximally effective, it is essential that it is constructed with the needs of the specific users in mind (Meyer 2002: 28). The issues relating to the planning and construction of the corpus will be discussed and illustrated in terms of the ISAE corpus in Chapter 4 of this study.

2.6.2 Limitations

Interestingly, one of the greatest strengths of the corpus, *viz*. its use of authentic language data, is also a potential weakness, but not through an inherent fault in the concept of authenticity itself. The problem lies in a potential for confusion between authenticity and typicality: every instance of authentic data is not necessarily a typical example (Hunston 2002: 107). Hunston illustrates the point with reference to several dictionaries (Longman, Cambridge and Cobuild) where lexicographers occasionally actually prefer to *invent* sentences to reflect typical usage instead of using authentic corpus examples if there are none that exactly match the needs of the dictionary being compiled. However, in such cases, it could be argued that the value of the corpus data lies in the delivery of a range of examples to enable generalizations about recurring patterns or constructs for the synthesis of a typical sentence or phrase.

On a technological level, Barnbrook (1996) identifies additional basic problems associated with the electronic analysis of linguistic data. The most obvious one is that linguistic data does not usually come in a ready-to-use electronic format. Almost all language data requires some form of preparation or processing to make it compatible for the computer. This is especially true of the spoken language, which has first to be transferred from the spoken medium into a visual medium (writing) and then rendered in a format that is suitable for electronic processing. In the absence of any reliable computer program to automate the transfer, transcription of speech is highly labour-intensive and fraught with possible inconsistencies, unless the highest standards of quality control are adhered to (see also Chapter 4). A related difficulty is the need to design and test specific programs that will accommodate the kinds of linguistic analysis required. Finally, despite the speed and power of computers, Barnbrook's (1996: 12) caution that the computer is still a machine and as such lacks "normal human background knowledge" is not as trite or obvious as it seems. The computer can only be effective in processing the complexities of language if it has been properly programmed to do so. And even then, it cannot interpret the results that it delivers in terms of established linguistic theory. It requires human intelligence, intuition and sensitivity to construct meaning from the large volume of descriptive data that the computer can deliver so efficiently and obligingly.

Chapter 3: INDIAN SOUTH AFRICAN ENGLISH IN THE WORLD ENGLISHES PARADIGM

3.0 Overview

This chapter situates Indian South African English as a sub-variety of South African English. It describes the development and main features of this sub-variety. It is structured as follows:

3.1 The family of World Englishes.

3.2 South African English and its sub-varieties.

3.3 Indian South African English: who speaks it?

3.4 Socio-historical background to the arrival of Indian immigrants to South Africa.

3.5 Features of Indian South African English: phonological, lexical and syntactic.

3.1 The family of World Englishes

The theoretical model of "Englishes" initially proposed by Kachru (1985) argues that there is not just one English language, but a "family of World Englishes" made up of several regional or national varieties, each with its own integrity and cultural identity. Thus Kachru's seminal model of three concentric circles seeks to account for the spread, patterns of acquisition, and functions of English in different cultural contexts. Briefly, the three concentric circles that he proposes are made up of the "Inner Circle", where the majority of the population are monolingual English speakers (UK, USA, Canada, Australia, New Zealand), the "Outer Circle", where English is an institutionalized relic of colonial policy and a vibrant additional language (e.g. India, Pakistan, Sri Lanka, Malaysia, Singapore, the Philippines, Nigeria, Ghana, Kenya, Zimbabwe, Zambia, Tanzania, South Africa) and the "Expanding Circle" where English is essentially a foreign language (e.g. China, Japan, Korea, Russia, Brazil). Thus the difference between the role of English in the "Outer" and "Expanding" circles is fundamentally linked to the functions that English serves in these contexts. In "Outer Circle" countries English is encountered in official domains such as government or commerce and might be regarded as a “second or additional language” while in the “Expanding Circle” countries English is essentially a foreign language. These differences are expressed in terms of lexis, idiom and syntax.

On the one hand, the model has been lauded for legitimizing varieties of English around the world by taking account of the cultural contexts in which these varieties emerged and where they continue to develop. But on the other hand, it has been criticized for entrenching an essentially ethnocentric view of English by placing historically native varieties of English (British and American) at the centre and relegating other varieties to the periphery. It has also been argued that it fails to capture the fluidity and exchange that occurs between the circles (Yoneoka 2001). In response to these shortcomings in the Kachruvian model, Yoneoka proposes an “umbrella-shaped” model. According to Yoneoka’s umbrella-shaped model, the handle would represent the core of English, the spokes would be the communication network and support systems, the tips would be the different varieties of English, the fabric covering would be the background socio-cultural systems, and the top would be an idealized “standard” English. Yoneoka argues that the umbrella-shaped model explains the genesis, development and relationship between the historically native and the non-native varieties of English in a more “egalitarian, flexible, generic and dynamic” way. Equally critical of the Kachruvian concentric circles model, Bruthiaux (2003) talks of “squaring” the circles, in an attempt to highlight the inequalities implied by Inner (native) vs. Outer (non-native) varieties. The debate about the actual shape and form of the model is ongoing, but the pluricentric view reflected in the term “Englishes” remains unchallenged.

3.2 A new model: the core and periphery of world Englishes

More recently, Nelson (2006) has proposed a method of investigating the “core and periphery of world Englishes” using statistical analyses of lexical items in the corpora of six varieties of English in the ICE project: British, Hong Kong, Indian, New Zealand, Phillippine and Singaporean. Nelson’s view of the varieties, which is conceptualized as a series of overlapping Venn diagrams, is based on a belief in the principle that there must be a common core of mutually intelligible lexical items and syntactical constructions which all the varieties of world English share. Nelson's investigation seeks to establish the dimensions and contents of the core, from which there radiates a graded periphery. Basically this involves the identification and comparison of lexical items in the different corpora. The analyses of lexical items ultimately reveal grammatical features as well, since, as the author point out, "lexis is ultimately grammatical" (Nelson 2006: 116). The value of this approach is that it seems to address exactly the concerns raised by earlier research (Yoneoka, Bruthiaux) about the lack of egalitarianism and the failure to account for the exchange that occurs between varieties of English in the Kachruvian concentric circles model. In a world where there are more non-native speakers of English than historically native speakers (Crystal 1995), Nelson's corpus-based exploration is valuable in that it does not make use of *a priori* assumptions about a hierarchy of different varieties of world Englishes, or about the directions of diffusion between the varieties. Briefly stated, the findings from this investigation were that the "absolute core" of items common to all the varieties of English was very small (only 11% of types), but these occurred with very high frequencies and consequently made up 91% of all the tokens in the corpora. The "absolute periphery", consisting of items which occurred in one corpus only, was by contrast made up of 63% of all types, but as these occurred very infrequently, it made up only 3% of all tokens in the corpora. It would appear that this investigation has potential as a model for usefully comparing and conceptualizing varieties of the same language without placing an explicit or implicit value judgment on any one variety. It acknowledges the shared nucleus common to all varieties of the same language, and by establishing a cline of sharing for items outside the core, it indicates the dynamic exchange that exists among language varieties. This is relevant to a country like South Africa where a rigid classification of English into native and non-native varieties has the potential to raise socio-political problems and possibly also real linguistic ones.

Currently, "South African English" is the overarching name given to the constellation of varieties of English that are used in the country. In general terms, it can be said to include the English of those for whom it is a mother tongue, as well as those for whom it is an additional language. In this respect, and staying with the Kachruvian terminology and classification scheme just a little longer, one could describe the linguistic situation in

South Africa as one where “Englishes of the Inner Circle and Outer circles co-exist in close proximity” (Coetzee-Van Rooy and Van Rooy 2005:3). It is a linguistic fact that languages or varieties of a language cannot exist in “close proximity” without some cross-fertilization taking place.

3.3 South African English and its sub-varieties

The current situation with regard to classification in SAE acknowledges the existence of various sub-varieties but does not really attempt to accommodate or explain the relationships between them. Brief reference has been made to these sub-varieties: Afrikaans English (Watermeyer 1996), Black South African English (Gough 1996), Cape Flats English (Malan 1996), Xhosa English (de Klerk 2002), Indian South African English (Mesthrie 1996), and possibly others (Chapter 1). While it may be appropriate to distinguish varieties of world English on the grounds of their national or sub-continental boundaries, the description of the sub-varieties of SAE by means of racial or ethnic labels is not undertaken without awareness that this kind of classification inherits ideological problems. In this regard, it has been argued that marking sub-varieties of SAE with racial qualifiers (such as Black, Indian or Coloured) entrenches their status as ethnolects that are “other”. Leaving the “colonial” variety of English unmarked could be interpreted as affirming the position of a variety that is spoken by less than 20% of the population and setting it up as the standard against which all other English varieties in South Africa are measured (De Kadt, cited in Coetzee-Van Rooy and Van Rooy 2005). Other researchers such as Lass defend the practice of using “white SAE as a kind of reference point for all other varieties” in South Africa, and leaving it unmarked, arguing that it “is simply a matter of history” (2002: 104). There is no point in arguing with a bald historical fact, but it seems that a case could be made for assessing a way forward from that point. In this regard, and in the rapidly-changing linguistic environment of South Africa, it might be useful to review the assumption that SAE should remain synonymous with the variety spoken by white English-speaking South Africans.

At the outset of this study I was aware of the subtle power dynamics inherent in the kinds of historically-determined naming practices described above. I therefore adopted a cautious approach, as I did not wish to align this study with systems that would wittingly or unwittingly perpetuate apartheid-type taxonomies and hierarchies. My efforts to find alternative, more recent labels in the relevant literature for these sub-varieties of SAE were unsuccessful and I had to concede realistically, that “the linguistic ecology” (Coetzee-Van Rooy and Van Rooy 2005) that has survived in post-apartheid South Africa is a relic of the policy of racial segregation that obtained in the country for over sixty years. As an official policy, it permeated every domain that could be legislated and regulated: housing, employment, schooling, transport, recreation and even worship. Against this background of social history it is hardly surprising that language developed along racially-determined lines as well.

3.4 Indian South African English: the history

The term “Indian South African English” (also called South African Indian English) is used here to describe a sub-variety of SAE which displays certain robust features and which is spoken by South Africans of Indian extraction. According to the South African census of 2001, just over 1.1 million people (2.5%) out of a total population of 44.8 million, classified themselves as “Indian or Asian South African”. Most people in this self-classified group are of Indian extraction, although there is a small group who are of Chinese extraction. Roughly 800 000 or 71.6% of Indian South Africans live in KwaZulu-Natal, with the remaining numbers geographically distributed in the provinces of Gauteng, the Western Cape and the Eastern Cape. The 2001 census also revealed that 95.8% of Indian South Africans listed English as their first language or home language. The figures relating to home language are interesting because they reveal the extent to which this community has “shifted” from the ancestral Indian languages to English as a first language. A comparison of the 1996 (94.4 %) and 2001 (95.8%) census figures reveals that the momentum of shift towards English is continuing steadily. In order to understand the factors associated with this shift, and the subsequent sub-variety of ISAE which evolved, it is necessary to give a brief historical background starting with the arrival of the Indian immigrants to South Africa.

3.4.1 The arrival of Indian immigrants to South Africa

Between 1860 and 1911 about 152 000 Indians arrived in South Africa under the system of indentured labour which had been designed to service the sugar cane plantations in the former province of Natal (Bhana and Pachai 1984). Subsequent waves of Indian immigration included traders (from India as well as Mauritius), merchants and Christian missionaries who were known as “free” or “passenger Indians”. This latter group, acquired the name “passenger Indians” because, unlike the indentured labourers, they had paid for their passages out to Natal (*ibid*: 2; Brain 1983). The Indian immigrants, who were drawn from different geographical regions in India, were a linguistically diverse group and their interactions with each other were not anchored by a common language. For example, those from the south-east spoke Dravidian-based languages such as Tamil and Telugu; those from the north-east spoke a variety of languages and dialects which eventually “coalesced” in South Africa under the name “Hindi”; those from the west were mainly Gujarati speakers, together with a few Marathi, Konkani and Meman speakers; and finally there were smaller numbers of Urdu, Bengali and Panjabi speakers (Mesthrie 1992*b*: 7). Only a few immigrants (mainly teachers, traders, Christian missionaries and interpreters) had a command of English. This linguistically mixed group of immigrants who lacked a common language, faced further linguistic challenges when they arrived in the former province of Natal, where the main languages were English and Zulu.

3.4.2 The language situation in South Africa in the late 19th and early 20th century

Just pre-dating the arrival of the Indian immigrants, Fanakalo (a pidgin comprising elements of the Nguni languages together with English and Afrikaans) (Adendorff 2002) had developed as a type of *lingua franca* in South Africa. With Fanakalo already in place and functioning as a *lingua franca* between the colonial and administrative authorities and the indigenous people of the province, there was no immediate linguistic imperative for the Indian immigrants to acquire English as such. Mesthrie (1992*b*: 23-27) cites an interesting mixture of documentary and anecdotal evidence to illustrate how the Indian immigrants employed Fanakalo for communicative purposes not only with Zulu, English and Afrikaans speakers, but with speakers of the other Indian languages as well. Apart from its use as a common everyday language, Fanakalo could be pressed into service in a range of situations, as the following example illustrates: (The incident, quoted in Mesthrie 1992*b*, was related through a scribe to the Protector of Indian Immigrants in 1903):

> The Calcutta man told me 1/- would be deducted from my wages for the sheet being torn – and I said ‘Sooga wina manga’ [‘Get away you’re lying’] and went away to my work – this was about four o’clock in the afternoon.
>
> I did not use the words ‘Sooga wina manga’ to the mistress, but she mistook me, and she gave me ten cuts with the riding whip.
>
> (Mesthrie 1992*b*: 24).

In the above example, a South Indian worker reports using the Fanakalo words “Sooga wina manga” to a North Indian co-worker (“the Calcutta man”). The English employer’s assumption that the remark “Sooga wina manga” had been addressed to her, suggests that Fanakalo may have been used as a medium of communication between employers and the Indian labourers as well.

3.4.3 The move towards English

However, a noticeable linguistic shift in the Indian immigrant community started to appear about the turn of the 20th century. Mesthrie (1992*b*: 26) quotes an observation by M.K. Gandhi in the newspaper *Indian Opinion* (dated 30/1/09) that voices concern over a growing preference for English amongst the “Indian youth”. In this newspaper, Gandhi remarked that “some Indian youth having acquired a smattering of English, use it even when it is not necessary to do so”, and he cites informal conversations and correspondence as examples of situations where he considered it unnecessary to use English. The significance of these observations is, firstly, that it highlights the encroachment of English into the domains of both speech and writing, and secondly, that it was the “youth” who were often responsible for initiating these linguistic moves.

Unlike Bughwan (1970) who contends that Indians in South Africa acquired English chiefly as a result of contact with native speakers of the language, evidence from other research (Brain 1983, Mesthrie 2002*b*) suggests that the models for English acquisition would have been more varied and would have included both native speakers (teachers and residents in Natal) as well as non-native speakers (teachers and missionaries, some of whom would have been Indians) (Mesthrie 2002*b*: 339). This is borne out by evidence on the activities of early missionaries, which often included attempts to establish schools where English would be included as a subject. Brain (1983: 179–202) identifies some of the key figures involved in the education of Indians in Natal: among these are Fr. Sabon (French), Madame Krovatchovik (language background not stated), Mr Rock (Anglo-Indian), Henry Nundoo (Indian), Rev. Stott (English), Balagdoroonada (Indian) and Mr Coster (English). This list indicates something of the range of native and non-native English input the Indian immigrants would have received formally from teachers, but also informally in religious and possibly social contexts.

The acquisition of English as an additional language at school and places of employment, continued as a pattern for the first half of the twentieth century, until educational provisions improved in the late 1950s (Mesthrie 2002*b*: 340). In the 1960s, with more children exposed to English-medium instruction, English was brought back into the homes and the shift to English gained momentum. A comparison of census figures for home languages before and after the 1960s reveals a dramatic shift away from the ancestral Indian languages towards English (see Chapter 4.).

In the latter part of the twentieth century, the widespread use of English by Indians in South Africa has been the subject of various research studies. Bughwan (1970), for example, found that school pupils claimed that they felt more confident using English than the ancestral Indian languages. Mesthrie (1991) has investigated in particular the decline of Bhojpuri-Hindi and its replacement by English. In the South African context, the displacement of Bhojpuri-Hindi by English reflects a typical scenario which affected, without exception, all the ancestral Indian languages. Mesthrie (1992*b*) summarizes the reasons for the general move to English as follows: the strong economic value of English; the lack of a common language in the Indian community; the lack of access to prestige or literary forms of the Indian languages because the South African education system did not accommodate the vernacular needs of minority communities; diminished contact opportunities with India during the apartheid regime, and finally the imposition of Afrikaans as a compulsory second language at school (1992*b*: 32-33).

The shift to English by Indians in South Africa is not unique to this expatriate community from the sub-continent: it has also been observed in other communities of the Indian diaspora. For example, Tent and Mugler (1996) have argued for the inclusion of a Fijian component to ICE based on their studies of language practices in Fiji which have revealed a surging transition to English by Indo-Fijians. Their summary of the language situation in that country concludes that the move towards English is "making inroads into the private lives of people…more so among Indo-Fijians than Fijians" (*ibid*: 281). The reasons they advance to explain the shift are largely similar to those that influenced the linguistic change in the Indian South African community: desire for economic advancement through proficiency in English, perceived prestige of English over the Indian languages and poor literacy in Hindi (the main Indian language of the Indo-Fijian community).

3.5 A new sub-variety of South African English: Indian South African English

The existence of ISAE as a recognized sub-variety of SAE is well documented (Bughwan 1970, Crossley 1987 and Mesthrie 1992*a*, 1992*b*, 1996, 2002). It is a separate ethnolect of South African English spoken by South Africans of Indian extraction. It is largely distinct from "Indian English" spoken on the Asian sub-continent, the latter being broadly characterized by ornate lexis and stylistically formal constructions (Kachru 1994). Obviously however, ISAE has residues of some lexical and syntactic features which are rooted in the ancestral Indian languages, but as a sub-variety of SAE it shares several features with other sub-varieties of SAE. ISAE in turn has contributed to the lexis of SAE.

Brief reference has already been made to the efforts of earlier research which used deficit-based theories to explain features of ISAE. In keeping with the linguistic and political trends of the time, such studies adopted an exonormative approach which involved a comparison of ISAE with Standard British norms of pronunciation and syntax. Revised linguistic thinking has favoured an endonormative approach, where language varieties that deviate from the standard variety are examined as rule-governed linguistic systems that fulfill valuable functions for the speech community which uses them (Schneider 2003). To this end, it has been suggested that ISAE should be regarded as a distinct ethnolect rather than “a substandard variety or a ‘bad’ approximation of (good) English” (Mesthrie 1992*b*: 220).

Within this framework and basing his theory in creolistics, Mesthrie uses ISAE as an umbrella term for a series of lects along a continuum from basilect through mesolect to acrolect. According to this theory, the basilect is the variety that is furthest away from Standard SAE, and the acrolect is that variety which is closest to it. According to research up to now, Mesthrie has discovered that most speakers fall into the mesolect category and regression to the mean that this lect represents, is the natural choice in most colloquial situations. He argues that this is a typical linguistic strategy in many creole societies, where it has been observed that the safe default is the mesolect as it does not carry the connotations of lack of education or sophistication of the basilect, or the affectation and pretension that is sometimes construed of speakers who use the acrolectal form (Mesthrie 2002*b*). However, the situation for speakers of ISAE is a dynamic one, with shift between lects on this continuum occurring in response to contextual demands. Movement in the opposite direction, or down-shifting towards the basilect, also occurs as part of a linguistic accommodation strategy in certain contexts, as it signals friendliness, informality and intimacy.

3.5.1 Theoretical classification of ISAE

Mesthrie argues that ISAE is 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 1992*b*: 3). At first glance, and with the Kachruvian model in mind, it might appear that ISAE developed from an immigrant language, to a “nativized” second language. However this does not take sufficient account of complexities posed by a linguistically diverse immigrant group needing to adapt to a linguistically diverse new environment. Studies by Mesthrie (1992*a*; 1992*b*; 2002) suggest that the process involved in the development of ISAE was one which more closely resembles "creolisation" than "nativisation" – the latter, as will be explained, is associated with second language acquisition. "Nativisation" involves lexical changes and adaptations to a person's *second language*, while the active use of another language as first language is maintained. This analysis clearly does not fit the picture of ISAE, where the use of the Indian languages gradually declined, as school-going children acquired English. In 1956, almost fifty years after Gandhi's observation about the growing use of English amongst the Indian youth in South Africa, fluency and literacy in English had developed to the extent that they could be adduced as reasons for advancing the status of Indians from "immigrants" to South African citizens. The following extract is from an address given by Drs. Cooppan and Lazarus at a symposium organized by the Institute of Race Relations to consider arguments around the theme "The Indian as a South African":

> "In an investigation now being carried out among 1,300 pupils in standards VI, VIII, and X, in twenty-five different Indian schools in Durban, nearly *every* pupil stated that he or she could read, write and speak *best* in English. Most of them could not read or write in their traditional Indian languages, but a few claimed to be able to speak them" (Bhana and Pachai 1984: 243).

Thus, scarcely 100 years after the arrival of the first Indian immigrants, English had all but ousted the ancestral Indian languages and had secured a foothold as a replacement first language in this community.

Mesthrie's arguments in support of the process of a creolization of English can be summarized as follows:

- Second language acquisition gives one a second language, whereas creolization gives one a first. Several studies (Bughwan, Crossley, Mesthrie) have confirmed that within fifty years of the last Indian immigrants arriving in South Africa, English became the first or replacement language for their South African-born offspring.

- This naturally leads to a further point *viz*. that children play a great part in creolization of languages. In the case of ISAE, it was the children who brought home the language they had learned in the classroom. Mention has already been made of Gandhi's observation that the "youth" appeared to be using English increasingly (see 3.4.3).
- Second language acquisition is usually achieved alone, while creolization occurs in groups. During the stage when English was still a second language for Indians in South Africa, it served as a useful *lingua franca* within the Indian community (which lacked a common Indian language) besides its more obvious function as a means of communications with mother-tongue English speakers.
- Whereas second language acquisition usually involves a normal or fairly homogeneous linguistic background, it is clear from the evidence cited above that there was nothing homogeneous about either the language heritage of the Indian immigrants, or about the linguistic situation which they found in multilingual South Africa.

(Mesthrie 1992*b*: 185-6)

Thus although he does not claim that ISAE represents a "classic" creolization situation, Mesthrie suggests that the creolization model is more useful than the classic second language acquisition model in explaining the origin and growth of the variety.

3.6 Features of ISAE

3.6.1 Phonetic features

Although this corpus of ISAE was not compiled with the analysis of acoustic features in mind, for the sake of completeness, a brief summary of the main phonetic features in ISAE will be given. At a superficial level, ISAE is most easily distinguishable by its phonetic and prosodic features. However, the degree of adherence to the phonetic patterns varies greatly according to where the speaker is located on the continuum of lects: very pronounced in basilectal speakers to almost completely absent in post-acrolectal speakers. Within that picture, many speakers who occupy positions in the mesolectal to acrolectal range are capable of shifting between phonetic styles according to interlocutor and contextual demands. The rhythm of ISAE is discernibly syllable-timed as opposed to stress-timed (Bughwan 1970: 308; Mesthrie 2002*b*: 342), and in that respect it is similar to Indian English. Stress placement within individual polysyllabic words is often located on the final or middle syllable, where it would be placed on the initial syllable in SAE, as the following examples in the Table 1 illustrate:

Table 1: Stress placement in polysyllabic words

ISAE	**SAE**
pe'nalty	'penalty
in'dustry	'industry
or'chestra	'orchestra

The main consonant differences between ISAE and SAE are a tendency of the former to replace the dental fricative /θ/ in the word 'thin' for example, with a more tapped /th/ sound; the occasional replacement of the /t/ with a /th/ sound in the words such as *tooth*/*teeth*, *tongue* and *tonsil*, and a reduction of the audible friction in the /v/ and /f/ sounds. While earlier research into ISAE reported a marked confusion of /v/ and /w/ amongst speakers with North Indian language ancestry and the non-aspiration of /h/ amongst those with Tamil language backgrounds (Bughwan 1970), these patterns do not appear to have endured. Subsequent research findings have shown a marked decline of both these features, particularly in the speech of the under-forty age group (Mesthrie 1992*b*: 139-140), together with diminished occurrences of the retroflex form of the consonants /t/ and /d/ (*idem* 2002*a*: 341). The picture with regard to vowels is slightly more complex, with the oft-caricatured rendering of the /o/ sound as /ɔ/ not universally observed. More prevalent is the shortening of certain long vowel sounds e.g. /ɔ/ in 'sport' and 'forty' to /ɒ/ and the /ʊː / in 'fruit' to the marginally shorter /ʊ/. However it is interesting to note that whereas the glide in the diphthongs /aɪ / in 'right' and /aʊ/ in 'south' are weak in SAE and are produced to sound /a:/, ISAE has been more 'conservative' in its retention of the full form of the diphthong (*idem*1992*b*: 137).

3.6.2 Syntactic features

Mesthrie (1992*b*) has identified the following robust syntactic features of ISAE that occur across all lects:

1. the use of "*y'all*" as a plural pronoun
2. copula attraction to *wh-* in indirect questions e.g. "I don't know *where's* the broom" (= where the broom is)
3. the use of *of* with the construction "too much" e.g. "There's too much *of* noise in this class".

These are three of the most enduring features and occur even in the acrolectal forms (see also Chapter 5). In the case of *ya'll* as a plural pronoun, various theories suggest that it is a grammatical and semantic strategy to compensate for the lack of a second person plural pronoun in English (Crystal 2004, Mesthrie 1992*b*). Indian languages have distinct and complex singular and plural forms for the second person pronoun. In the South African context the Hindi-Urdu colloquial dialect the plural (of all cases) is formed periphrastically with *log* meaning 'people', while in Tamil the second person pronoun is *ni* (sing.)/*ningkal* (pl.) On the level of discourse, Crystal (2004:449-450) notes that a similar strategy is adopted in informal colloquial contexts by American speakers. Both these functions are explored more fully in Chapter 5, alongside data drawn from the corpus of ISAE.

According to the theory of creolistics, the further the speaker is located along the polylectal continuum, the more discernible is the process of decreolisation. What this means is that the movement towards the acrolectal forms requires a reformulation of basilectal norms and rules while often retaining the earlier patterns and functions. It is a pattern which is also observed in aspects of first language acquisition such as child language in negation and question formation (Mesthrie 2002*a*). These insights from first language acquisition could be applied to ISAE, which, as has been illustrated, is a replacement *first language,* not a second or additional language for virtually all members of this speech community.

In addition to the three features described above, the following syntactic features may also be observed in ISAE:

3.6.2.1 Topicalisation or topic formation

According to Mesthrie, ISAE, like other new Englishes "has a predilection for topic formation" (Mesthrie 1992*b*: 112). He usefully calls this a "cross-linguistic" tendency, as topicalisation builds new syntactical structures that occur in response to discourse needs: focusing or "fronting" the topic and thereby highlighting it for attention in the discourse. The topic that is foregrounded could either be "old information" or "new information", as the following two examples from the ISAE corpus illustrate:

<#05:$04M3H:215> No, no the thing is Abi *writing up* for me is very hard. (The speakers were discussing the writing of reports on scientific experiments. "Writing up" which had already been mentioned in the discourse, is "old information".)

<#10:$08M3T:595> Baby, *the geyser* you put it off? (The two participants in this conversation were discussing a computer game. The question about "the geyser" was an aside to a third person. It thus represented a change of topic and is "new information").

Finally, topicalisation could also include a grammatical strategy called "left dislocation", where the focused information is situated to the left of the syntactic subject of the sentence, as in the sentence "*My mother*, she is a housewife". In this case, the actual subject of the sentence is "she", but the topic under focus is "my mother", so it is inserted ahead of, or in spatially-specific terms, "to the left" of the subject of the sentence. Although topicalisation does occur in other varieties of SAE, it is an abiding feature of ISAE, which is discernible across all lects and Mesthrie's assertion that ISAE speakers have a "predilection" for this type of sentence construction, is therefore apt.

3.6.2.2 Use of conjunctions such as *but* in a clause-final position

In the following sentence from the ISAE corpus: "Don't you love spinach *but*?", the conjunction *but* is placed at the end of the clause. It is generally the semantic equivalent of ‘though’ or ‘really’ but can occasionally serve to counter possible objections to the idea being expressed. Mesthrie, has commented that the use of *but* in a clause-final position is a feature of Indic languages, although it occurs in some Northern English dialects as well (1992*b*: 108). Given the range of English language models that the immigrant Indian encountered in South Africa (see 3.4.2), together with the lingering influence of substrate Indian languages, the endurance of this grammatical construction in contemporary ISAE is hardly surprising (See also 5.2.5).

3.6.3 Lexical features

The lexis of ISAE is a complex area, indicative of various points of linguistic contact in its development and the broader social history of its speakers. ISAE has obviously retained many lexical items of Indian origin, especially those reflecting traditional religious and cultural practices. The following table features a selection of lexical items from the broadly classified domains of religion, food, kinship and clothing:

Table 2: Examples of ISAE lexis of Indian origin

Domain	Word	Meaning
Religion		
	namaz	Islamic daily prayers
	hardhi/nalengu	Hindu pre-nuptial ceremony
Food		
	gram/channa	chick peas
	halim	broth containing lentils and meat
Kinship		
	ben/akka	older sister
	thatha/nana	grandfather
Clothing		
	choli	tight bodice worn under a sari
	ijar	long pants worn by Muslim women under a skirt

(adapted from Mesthrie 1992*a*).

Apart from words of Indian origin, ISAE has imported and adapted words from other South African languages, especially from Zulu, for example *mutton gullah* from *amatungulu* (num-num), *dugu* from *induku*, (a stick), *skoten* from *isikhoteni* (a scoundrel), *gwaai* from *ugwayi* (tobacco or snuff) (*ibid*; Silva et al. 1996). This group of words (which all have Zulu etymologies), have not been remarked upon in other sub-varieties of SAE. Their currency in ISAE suggests that they might be relics of earlier use of Nguni-based Fanakalo (or even Zulu itself) by the Indian immigrants to KwaZulu-Natal.

As a sub-variety of SAE, ISAE shares many words with general SAE such as *robot* (traffic light), *dagha* (mud), *babalaas* (a hangover) and *tickey-line* (cheap or of poor quality); 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) and most recently, *thunnee/thanni* (a card game). *Thunee/thanni* (http://www.thunee.com) a popular card game limited to the Indian community in South Africa, was recently acknowledged (September 2006) as one of several indigenous "sports" or pastimes in the country. With moves to formalize the rules of the game and extend its influence, it is conceivable that the terminology associated with the game may be adopted into broader SAE usage. In addition to the specific lexis exemplified above, ISAE also features additional senses for general English words (Mesthrie 1992*a*), as the following sentences illustrate:

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

In the absence of sufficient contextual information to prime comprehension, an outsider to the ISAE speech community would be challenged to work out the significance of these italicized words.

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 exception to this generalization is the case of ISAE slang where the tendency to shift away from community-based norms manifests itself 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* or 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’ originally meaning 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 2002), ISAE slang is governed by gender and generation boundaries and in the case of this sub-variety, usage is generally restricted to the speech of young males (usually those under 25 years of age), but it may extend to older males from socio-economic groups and occupations which typically 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, Fanakalo and to a lesser extent Afrikaans, it was preserved and fossilized by the social isolation caused by the South African government’s apartheid policies enforced between 1948 and 1994. The present collection of spoken ISAE can be regarded as the first stage towards the construction of a fuller corpus of ISAE. Despite its modest size, this foundational section of the ISAE corpus could provide useful initial data for comparison with earlier studies of ISAE. It is envisaged that a full ISAE corpus could constitute an important building block in a truly comprehensive or “mega-corpus” of SAE, in which equitable ratios of the identifiable sub-varieties are represented. When established, the SAE mega-corpus 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.

Chapter 4: METHODOLOGY

4.0 Overview

This chapter discusses the methodology used to create the corpus of ISAE. It looks firstly at the overall design of the corpus and then focuses on the people whose speech was used to build the corpus. Thereafter it describes the data-collection process and the guidelines for consistency that were developed and applied during the transcription stage. The chapter concludes with a brief overview of how the data was analyzed. Each stage of the research process had its own set of unique problems, and these are described, together with an explanation of how these were addressed.

4.1 Structure and design

There is no "one-size-fits-all" corpus design, as the elements of each corpus are determined by socio-linguistic factors relating to the population under consideration and by the purpose which the corpus is intended to serve. In planning the ISAE corpus, the design features of significant earlier corpora were evaluated in terms of their suitability for the framework of the corpus of ISAE. Significant corpora in this regard have included mixed corpora of speech and writing such as the BNC (2.4.2) and ICE (2.4.3), 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).

The problems and challenges that faced the compilers of ICE-NZ (Holmes 1996) have been helpful in informing the general design of the ISAE corpus. Although the parameters of the ICE-NZ corpus were laid down by ICE, their interpretation of the categories in terms of local contexts and availability of data were insightfully articulated and functioned as useful guidelines for this study. Issues such as "Whose speech should be included in the corpus?" and "No Surreptitious Recording" were referenced in particular when compiling the ISAE corpus.

4.1.1 Size and boundedness of the corpus

Just as there is no one-size-fits-all corpus design, there is no ideal corpus size. There is, rather, only an optimum corpus size, which is determined by the research needs and more pragmatic considerations such as the availability of resources. In terms of size, the corpus for this research does 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) (2.4.2). It is, instead, a small sample 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.

As the research aim was to collect a corpus of conversations, the ICE specification for private, direct conversations between two people was consulted in an attempt to find a useful benchmark of size. The conversational component of each variety of English in ICE is 180 000 words, made up of 90 texts of 2 000 words each (Nelson 1996: 29). For a small-scale research project like this one, with one researcher carrying sole responsibility for all the conceptual as well as the labour-intensive practical aspects of the corpus collection and transcription, a scaled-down version of the ICE conversational component was considered to be a more realistic target. A corpus one third of the size of the ICE component for private direct conversations, or 60 0000 words in total, was therefore selected as an achievable goal for a research initiative pitching itself as an exploratory step towards a larger corpus of ISAE. Another reason for such a modest goal was the enormous amount of time required to convert the data into machine-readable form (see 4.1.2). de Klerk (2003: 467), quoting McCarthy, also argues in favour of smaller, well-designed corpora “of spoken material which contain authentic and reliable representative data, [that] can be analyzed exhaustively in a variety of ways”. With computers to handle the mechanical aspects of corpus analysis, it is tempting to accumulate vast quantities of data. However Kilgariff et al. (2004) caution that too much data makes even simple features such as word occurrence difficult and time-consuming to analyze and interpret. “If there are five hundred [occurrences of a word], it 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" (Kilgariff et al. 2004: 106).

The 60 000 word corpus of ISAE is made up of thirty 'texts' or speech samples of approximately 2 000 running words or 'tokens'. A token is "an individual occurrence of any word form" (Barnbrook 1996: 53). The number of running words or tokens is a simple indication of the size of the text. For example in the sentence, "Fair is foul and foul is fair" there are <u>seven</u> tokens although there are only <u>four</u> different 'types', as some words are repeated. Texts of 2 000 running words constitute the building blocks of the ICE, 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. In the case of the ISAE corpus, the text segments are occasionally slightly longer than 2 000 words, as it was important not to truncate a conversation in the middle of a speaker's utterance, but rather to include coherent text samples. Each 2 000-word text is a self-contained unit in that it was extracted from one thirty minute dialogue only. It is not a composite constructed from several short verbal exchanges, for example.

4.1.2 Type of corpus: spoken rather than written

This research selected spoken English as the starting point because ISAE is 'primarily [an] oral dialect' (Mesthrie 1992*b*: 35). Previous research on ISAE has also focused on the spoken variety of the language (Bughwan 1970, Crossley 1987, Mesthrie 1992, 1996). An investigation of several newspapers and magazines such as the *Sunday Times Extra*, the *Post*, the *Stanger Herald* and *SA India*, which are aimed at South African Indian readership, revealed these publications are highly edited and "sanitized" to reflect standard (South African) English in their final published form. The result is that apart from a scattering of lexical items of an essentially cultural nature, these texts carry very little evidence of the range of syntactic and idiomatic features of ISAE described in the research literature (see also Chapter 3 and Chapter 5).

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). With regard to corpora for varieties of SAE in particular, much research has already been done towards a corpus of spoken Xhosa-English (de Klerk 2002*a*, 2002*b*, 2003, 2006) which, it is hoped, will form part of a larger corpus of Black South African Englishes. Viewed against these national linguistic research initiatives, this study could constitute the first building block towards a corpus of ISAE, which could facilitate comparative studies of different sub-varieties of spoken English in the South African context.

Internationally there are more examples of written corpora than spoken corpora (Chapter 2).with even a mixed corpus such as the BNC exhibiting a ratio of 90% written data to 10% spoken data. In defence of the bias towards written data in the BNC, Leech et al. (2001) explain that the ratio was dictated by practical considerations. They acknowledge that the spoken language is "the primary channel of communication", and that on those 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" (*ibid* 1). Compiling a corpus of the spoken language is comparatively more difficult, labour-intensive and expensive than compiling a similarly-sized corpus of the 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. Possibly the only specialized area of study which does not require speech to be converted into writing is the type of acoustic analysis which uses specifically designed software to trace sound patterns and represent these graphically – albeit in a visual form. Spontaneous spoken interaction is also "messy" and not well-behaved syntactically: incomplete sentences are the norm, as are false starts, simultaneous speech and hesitations.

In the case of the corpus of ISAE, the transfer of the spoken data to the written mode involved listening to the recording several times and, in the absence of appropriate automated software to handle this stage of the process, the speech had to be manually transcribed word for word. Speech-recognition technology has been developed to automate the transcription of formal, clearly articulated speech such as broa dcast monologues and dialogues, but as yet there is no utterly reliable program to deal with the unpredictable nature of spontaneous speech. Transcribing the recordings was therefore the most arduous part of the research project: to wit, a 2 000 word text segment took roughly fifteen hours to transcribe, then mark-up with fairly simple annotations and finally to proof-read. Meyer (2002:71) cites similar experiences by ICE-USA, where the team reported that average time for transcription, annotation and proof-reading of a 2 000 word multi-party conversation was between fifteen and twenty hours. Casual, spontaneous speech is particularly untidy or unconstrained and cannot be straitjacketed into perfectly formed sentences or clearly enunciated words during production. In addition, casual conversations usually contain latched utterances (simultaneous speech) which are difficult and time-consuming to separate out. These and other specific difficulties associated with transcription are described more fully in 4.4.3 below.

4.1.3 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, there are two broad approaches to classifying spoken data for corpora: one demographically-motivated and the other context-governed or task-oriented. In the BNC for example, a distinction is made between private conversation (40%) and the more public, task-oriented aspects of speech (60%). In the task-oriented component, activities are classified in four domain-specific areas designated as follows: educational and informative, public and institutional, business, and leisure. Within each domain, the type of verbal interaction is identified as being either a monologue (such as lectures, speeches, sermons) or a multi-party activity (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. But the data-collection for this component used a demographically-motivated approach where variables such as social class, gender, age and geographical distribution were controlled across conversation samples. From the BNC experience it would appear that both approaches are valuable for determining text types for spoken corpora. de Klerk (2002*b*: 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 demographic considerations are also accommodated within the defined linguistic contexts.

However, such detailed sub-types are impractical in a small corpus, as the small sub-groupings would not yield sufficient data to enable the researcher to discern generalizable linguistic patterns and formulate reliable conclusions. Hence a decision was taken that the conversational material should be collected from one demographic band, *viz*. "young adults" (see 4.2.5). All the conversations are spontaneous, casual, face-to-face dialogues. The research takes casual conversation as a starting point for two reasons: firstly, spontaneous informal dialogues in private settings exemplify naturally occurring language, as everyone engages in this activity daily. More precisely, casual conversation can be regarded as the quintessence of language, a kind of "pre-genre" (Swales 1990) in the development of language since all other forms of language, whether spoken or written, can trace their genesis to this genre. 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 discoursers 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).

4.2 The contributors to the corpus: who counts as an Indian South African speaker of English?

There were 49 contributors to the corpus, all of whom were South African-born individuals of Indian extraction. Each participant provided biographical details and information about their linguistic background by filling in a questionnaire on the *Personal Details and Consent Form* (Appendix A). While 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 (see 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. Finally, in order to eliminate linguistic features which might be the result of recent contact with other languages or other varieties of world English, the corpus also excluded anyone who had spent more than 12 months outside South Africa within the last three years (see Question 5 Appendix A).

4.2.1 How the contributors were recruited

The first small group of data collectors and contributors to the corpus comprised family members and friends of the researcher, all of whom are students at Rhodes University. This group was enlarged by referrals and contacts recruited through the Hindu Students' Society (HSS), one of several extra-curricular student clubs at Rhodes University. The HSS organizes social functions for its members around Indian religious and cultural themes and the aim of the society, loosely stated is: "[To provide] an insight into Indian culture as well as [to provide] events for… members" (http://hss.soc.ru.ac.za). However the HSS seems to fulfill a role in the lives of its members that extends beyond the provision of entertainment and the observation of cultural feast days. At Rhodes University where Indian students constitute a small minority group, it has been argued that the HSS "serves a vital function in providing the 'critical mass' necessary for a feeling of community" (Boshoff 2005: 109). Although the term "Hindu" is part of its name, this is misleading, as not all members of the HSS are adherents of Hinduism. It would be truer to say that membership of this society indicates alignment with the expression and experience of Indian culture in broad terms.

The HSS executive recommended the names of potential fieldworkers and contributors from their database. I screened these referrals by means of phone calls, followed up by individual meetings in order to confirm their eligibility as members of the target population. The field-workers thus selected were then thoroughly briefed, supplied with material detailing the recording procedure (Appendix B), the hardware for the recordings, and the relevant forms to accompany each recording (Appendices A and C).

Mindful of the fact that ISAE is not generally used in public discourse and that members of this speech community tend to adopt "more careful and formal styles in public interactions" (Mesthrie 2002*b*: 341), the fieldworkers were asked to use their access to existing social networks to identify other contributors to help build the corpus. This measure was guided by significant earlier research, notably Gumperz (1970), who recommended that a researcher who might in any way be perceived as an outsider should avoid interaction with the targeted social group, and also by Milroy (1987) who 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 purposes. In other research, Schmied (1996: 186) appropriately refers to the "famous sociolinguistic paradox" of the observer effect where the presence of a researcher who is not from the "in-group" causes the participants in an observed conversation to speak in ways that are not natural, in a bid to strive for more standard or prestigious forms. In a further attempt to secure more naturalistic data, the fieldworkers were specifically asked not to structure the social interactions as interviews. In interview situations the field workers are generally in control of the discourse since they determine the elicitation topics and techniques. It was considered essential that the field-workers should function as active participants with equal speaker rights in the discussions as a measure to preserve their "in-group" member status. It is believed that all these precautions assisted in modifying the observer effect and contributed to the procurement of representative linguistic data. As Milroy (1987: 37) observes with regard to using a member with access to the social group to research the vernacular, "…much more detailed information can be obtained from a fieldwork method which takes pre-existing groups as its unit of study", as it does not isolate informants from their social network and helps uncover "…the linguistic repertoire and the 'internal' grammar of a non-standard code. Information of this kind is not easily recoverable from texts recorded by means of single person interviews."

4.2.2 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 (see 3.4.1). Although many terms are shared by all groups, others are more specific to linguistic or cultural groups, especially those from culinary, kinship, clothing and religious domains. In this regard, Mesthrie (1992*a*) distinguishes between "restricted and unrestricted usage" in ISAE. The designation "restricted usage" refers to terms used by sub-groups with a common ancestral language. A selection of everyday terms with restricted usage in Table 3 below, illustrates this point (from *A Lexicon of SA Indian English*, Mesthrie 1992).

Table 3: Everyday ISAE lexical items with 'restricted usage'

Tamil	**Hindi**	**English meaning**
karo	thitta	strong or chilli hot
manja	haldhi	powdered turmeric
amma	ba	mother
mundhani	pallu	the decorated or tasseled end of the sari which is worn either draped over the shoulder or used to cover the head
nalengu	hardhi	pre-nuptial religious ceremony to invoke blessings for the bride
botu	tikka/bindhi	the (traditionally red) dot or decorative mark worn in the centre of the forehead by Hindu women
pake	pan	betel nut

"Unrestricted usage" applies to terms such as *agarbathi* (incense stick), *bhajia* (a spicy fritter made of chick-pea flour), and *thanni/thunee* (a popular card game), which are lexical items commonly used by all Indian South Africans regardless of language or cultural group affiliation

There are no recent national statistics for ancestral language affiliation or symbolic attachment to Indian language groups in South Africa. In the South African Indian community, where the majority of its members have shifted to English as the sole or predominant medium of communication, census questions which ask for "home language" do not yield historical information about the ancestral language or cultural affiliation of the respondent. Despite the absence of such nuanced cultural information, it was possible to establish very broad guidelines for substrate language affiliation amongst Indian South Africans. I chose the South African 1960 census records as a reference point, since 1960 appears to have been a linguistic watershed for the Indian South African community: there is a marked decline in the use of Indian languages as home languages after that date (Mesthrie 2002*a*: 165). Thus, as shown in Table 4 below, the ancestral language distribution in the corpus is closely aligned to the 1960 census figures for the group as a whole.

Table 4: 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	.5%	8%	4

Information about language loyalty or affiliation was gathered by asking each respondent to indicate which of the five main Indian language or cultural groups in South Africa they identified with most strongly (Question 8 Appendix A). The question also provided for anyone who did not align themselves with these groupings by providing a sixth option designated 'Other', and requesting specific details if this option was selected. Of the four respondents who chose 'Other', two specified 'Memon' [Meman] (a dialect of Sindhi), one specified 'Islam' and one specified 'English'.

The substrate language groups represented in the ISAE corpus are Hindi 37%, Tamil 27%, Gujarati 14%, Telugu 10%, Urdu 4% and other 8%, as Figure 1 below illustrates.

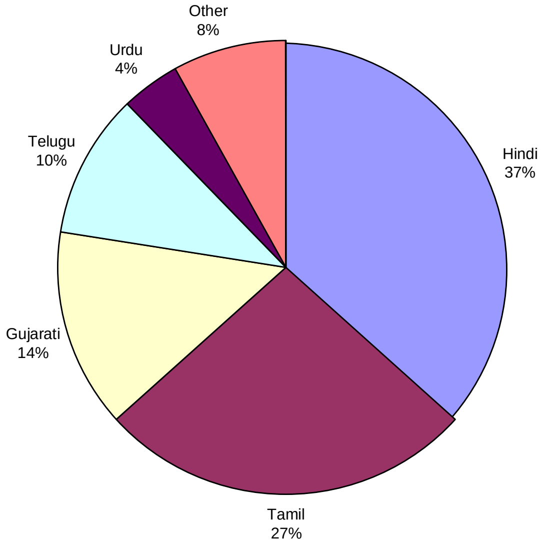

Figure 1: Substrate Language Groups in the ISAE Corpus

4.2.3 Actual Language Practice

In addition to gathering information about substrate languages, it was important to establish the current language practices of the corpus contributors, as it was felt that this would provide a more realistic and complete view of each respondent's linguistic profile. As stated earlier, since the Indian South African community has been involved in major language shift from the indigenous Indian languages to English (Chapter 3), the ancestral language is often a notional concept indicative of loyalty or symbolic identification with the language group. It was useful to have this trend confirmed by the data collected for this research. Of the 49 corpus contributors, 47 (96%) indicated that English was the first language they had learnt to speak at home (see Question 9 Appendix). Only two contributors indicated that they had learnt Indian languages first: one female respondent (aged 16-19) had learnt Tamil first and one male respondent (aged 20-24) had learnt Gujarati first. However despite learning Indian languages first, both these respondents said that currently English was the only language they could speak (see Question 10 Appendix A). Thus it is clear that neither were active speakers of the Indian languages learnt in their childhoods. In order to gain a broader picture of the other linguistic inputs that could have impacted on the contributors' performance, Questions 6 and 7 elicited information about the language practice in the home, by asking what languages the respondents' parents used most often in the home. The picture that emerged showed that all the mothers used English most often in the home and only one father used an Indian language (Hindi) at home. These responses indicate the almost exclusive use of English in the home by both parents. Only five contributors (10%) claimed to have any fluency in Indian languages: three indicated Hindi and two Gujarati. However 22 (45%) said that they could read and write various Indian languages in addition to English (Table 5 below).

Table 5: Literacy in Indian languages

Language	**Male**	**Female**	**Total**
Arabic	5	2	**7**
Gujarati	2	0	**2**
Hindi	3	4	**7**
Gujarati & Hindi	1	0	**1**
Tamil	0	3	**3**
Urdu	1	1	**2**
TOTAL	**12**	**10**	**22**

The fact that these 22 respondents claimed literacy, but not corresponding fluency, in Indian languages, suggests that the languages may have been learnt in formal settings, most likely in community-run vernacular schools. It is therefore reasonable to infer that they were not active speakers of the languages in question: at best they could be passive speakers capable of following conversations and making short contributions. The picture that emerges from these speaker statistics is consonant with findings by Mesthrie (2002*b*: 165) who observes that within the Indian South African community, families and individuals retain a strong "symbolic attachment" to particular Indian language or cultural groups, although *de facto* English has become the replacement language in the home. Probing the language practices of the corpus contributors in such detail provided solid background information against which to interpret the corpus findings. The statistics revealed that English is indisputably a first language for 100% of the speakers surveyed, and that for 90% of them it is the sole language in which they could confidently profess fluency.

Finally, it is necessary to comment on the two respondents who did not identify with the traditionally defined linguistic groups. One specified "English" (although he could also speak Gujarati and read and write Arabic) and the other indicated a religious affiliation, *viz*. "Muslim". The only common denominator that these two respondents shared is that they were both males, under 25 years of age. On the basis of these two isolated responses, it would be premature to surmise that this is indicative of a general trend away from language-based affiliations, as assumptions about identity constructs are beyond the scope of this study. But it would be interesting to survey ancestral language or cultural group affiliations in a similarly-constituted group in about ten years' time.

4.2.4 Gender

In order to avoid a gender bias in the corpus, right at the outset, when participants were recruited, attention was given to achieving a 50:50 gender distribution (see Table 6 in 4.2.5). In addition to maintaining an overall gender balance, care was taken to ensure that there were equal numbers of conversations and same-sex dyads as between mixed-sex pairs. Thus there were ten conversations between women only, ten between men only and ten in mixed gender groups. Such background considerations were important in planning and building the corpus as I felt that it would result in a corpus that would enable comparisons of language use in equally-weighted gender group configurations.

4.2.5 Age

The contributors to the corpus were all young adults ranging in age from 18 to under 29 (see Table 6 overleaf). They did not supply their actual ages on the "Personal Details and Consent Form", but ticked a box to classify themselves in one of several age group categories. Although there were 20 in the 16–19 age group, in reality no one was under 18 years of age, as all indicated that they had passed Grade 12 (see Question 4 Appendix A). The largest number of speakers (45 or 92%) fell into the 18–24 age group, as the data collectors, who were all similarly-aged university students, had used their existing social networks to identify corpus contributors. The age range (18–29) represented in the ISAE corpus makes it potentially useful for a variety of reasons. Firstly, it represents a small, highly focused collection of ISAE speech for this age and education sector. Secondly, it has potential value for comparisons with similarly profiled corpora, such as the locally-collected Xhosa-English Corpus (de Klerk 2002*a*, 2002*b*) and the Corpus of London Teenagers (COLT) abroad. (Stenström et al. 2002).

Table 6: Corpus contributors classified according to age and gender

AGE GRP	M	F	TOTAL
16-19	9	11	**20**
20-24	12	12	**24**
25-29	4	1	**5**
	25	**24**	**49**

4.2.6 Geographical distribution

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.

4.3 Data collection

4.3.1 Time frame for data collection

Since the research focus is *contemporary* South African Indian English, it was important to collect the data within a fairly narrow time frame, and to build a reliable synchronic corpus which would exclude, or at best minimize, variables related to language change. According to Meyer (2002: 46), a time-frame of five to ten years may be regarded as reasonable for synchronic corpora. Data for this project (toward a corpus of spoken ISAE) were collected from October 2004 to April 2006, a period of eighteen months in total.

4.3.2 Equipment

A decision had to be made about whether to use a professional recording studio or whether to make the recordings privately. The main advantages of professional studio recordings are the guarantee of superior sound quality (an essential criterion for research which intends doing any kind of acoustic analysis) and the ease with which recordings can be manipulated during playback, such as making just one speaker audible at a time, or eliminating background noise. As this research was not aimed at acoustic analysis, professional recordings were not considered to be essential. Furthermore, it was felt that the exercise of arranging a studio recording would put the speakers into an artificial setting and make it more difficult to capture the kind of spontaneous naturalistic data which was the focus of this research. Finally, on a practical level, the most economical method of obtaining the recordings had to be the guiding principle.

The simplest and most affordable solution therefore, was to supply the fieldworkers with small, battery-powered analogue tape recorders which had built-in flat microphones (Panasonic Model RQ-L10). The decision to make analogue recordings rather than digital ones was also a pragmatic one, influenced by financial constraints and technological availability. Admittedly, digital recordings do have definite advantages over analogue recordings: the sound quality is superior, digitized material is less likely to degrade, and lastly they can be easily transferred to computer or CD. However, Crowdy (1993: 261) reports that "under good recording conditions they [are] not significantly better than analogue..." I reasoned that, if required, at a later stage the analogue recordings could be digitized using specially designed software, such as Syntrillium's 'Cool Edit' program.[5] As the conversations did not involve more than two active participants in private settings, I hoped that the equipment chosen would be sufficiently sensitive to deliver fairly good quality recordings. The size and simplicity of the equipment was also a significant consideration. I aimed to ensure that the technology should be unobtrusive, unintimidating and manageable in informal, private settings.

Eventually, out of a total of 37 recorded conversations, only 30 were eligible for inclusion in the corpus. Six conversations were rejected because of the intrusion of background noise, or because the built-in microphone had not been optimally positioned during recording. I believe, in retrospect, that a compromise between a studio recording

[5] share-ware freely available from http://www.syntrillium.com/cooledit/index.html

and a private recording would have been to have arranged for each participant to be wired up with a small lapel microphone. This would have significantly enhanced the audio quality of the recordings, while using easily-managed equipment in a private informal setting. A similar practice has been successfully implemented in other research, notably COLT (Stenström et al. 2002). The seventh recording was rejected because the speakers were obviously "playing to the tape recorder" by engaging in a series of jocular disconnected exchanges, which parodied the Indian South African accent and speech. There was no evidence of a topic thread being followed or of the participants responding to each other in the manner that was displayed in the recordings selected to build the corpus.

4.3.3 The recordings

For ethical and legal reasons, all participants were informed at the outset that they were being recorded. In the interests of capturing naturalistic data, this practice has not always been upheld. Early corpora such as the LLC recorded people without their knowledge and only informed them after the recording had been completed. In the South African context Mesthrie (1992*b*: 40) has admitted to making and using covert recordings of close friends and relatives to draw inferences about style-shifting, and to make comparisons with the main body of data that he collected. However, in the case of this corpus, a firm decision was taken that recordings would not be surreptitious. To this end, signed permission was obtained from all participants prior to the recordings (Appendix A). In addition, a confidentiality clause assured contributors that their identities would be protected (Appendix B).

This did raise the scientific dilemma of the "observer's paradox", where the knowledge that data is being recorded makes participants self-conscious, thus jeopardizing the very spontaneity and naturalness that the research seeks to capture. Labov (1972: 181) states that "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation". In order to address this problem, a far longer piece of dialogue than the 2 000 words required for the text sample, was recorded. For this research, 30-

minute conversations were recorded, supplying on average about 4 500 words each. Each conversation therefore yielded ample data, making it possible to ignore the first 10 to 15 minutes of the recording but still extract a coherent text sample. I believed that by selecting data for transcription 10 to 15 minutes into the recording, the participants would have relaxed and adjusted to the presence of both the tape recorder and the field worker. A similar practice has been followed by other researchers, notably Holmes (1996), who found it successful in yielding reasonably naturalistic data for the New Zealand component of the ICE Project.

Since casual conversational speech data was the data collection target, participants were not interviewed or required to respond to a set of stimuli or triggers. In fact, the fieldworkers also functioned as conversation participants. Once the respondents had accepted that the data collection was for *bona fide* linguistic purposes and that their identities would be protected, field workers reported a high degree of co-operation from participants. They were amenable to being recorded in a variety of informal social situations such as relaxing with friends, playing computer games, while preparing meals and discussing academic problems.

4.4 Transcription and storage

As soon as a recorded tape was returned, each side was checked for quality and eligibility for inclusion in the corpus. Each recording was then assigned a file number from #01 to #37. A back-up copy was made, marked ‘Copy’ and stored separately from the original recording.

4.4.1 Speaker profiles

Before transcribing the data, relevant biographical information about each speaker was captured in the database. This included the details provided by each contributor with regard to gender, age group, level of education, schools attended, language used by mother in the home, language used by father in the home, participant’s language or cultural group affiliation, first language spoken, languages currently spoken at home and proficiency in reading and writing Indian languages (Appendix A).

As contributors to the corpus were anonymous, each speaker was firstly assigned a core identity number from 1–49, designated as follows: $01–$49. The fieldworkers assisted in identifying the voices on tape and linking the speaker numbers with biographical profiles in the database. Thereafter, other biographical details such as gender, age group and ancestral language group were encoded in 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 (Svartik 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. Names of public figures who are both a matter of public record and which may be crucial to making sense of the discussion, were not protected in this way. The basic principle which guided these decisions was to disguise details only if they would identify private individuals.

4.4.2 Classifying and storing the data collected

The thirty recordings used in the corpus were given file numbers which ranged from #01 to #37. Each file was prefaced by additional header information which encoded details about the material in the file, to facilitate identification and retrieval.

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. 2014)

Marking the texts in this way makes it easy to locate the actual tape used for the recording and match it to the transcription. This header information is stored separately, as it is not part of the speech text itself, but it does provide a useful "handle" for the extraction of material.

4.4.3 Transcribing and computerizing

"Transcription involves capturing who said what, in what manner (e.g. prosody, pause, voice quality), to whom, under what circumstances (e.g. setting, activity, participant characteristics and relationships to one another)" (Edwards 1995: 20). It involves the transfer of audible material to a visible medium. Since spoken language does not occur in clearly defined lexical units or in neat grammatical sentences, the activity of transcribing speech "is in essence a highly artificial process" (Meyer 2002: 71). Furthermore, any transcription system involves a selection of what is deemed to be important or significant. In order to capture every single aspect of a conversation in minute detail using writing conventions, one requires a very "narrow transcription" style, as opposed to the "broad transcription" style used to represent speech in a play (Edwards 1995: 20). As an example of a narrow transcription style, Edwards cites Pittenger et al. whose transcription of five minutes of interaction yielded 183 pages of transcript. It follows therefore, that every attempt to transfer speech into writing implies a set of choices around what aspects of the speech act to record, and what to leave out. These choices in turn depend on the use to which the transcription will be put. There is also much that occurs in speech contexts that is difficult to transfer objectively into writing. These features include stress, intonation, gesture, voice quality, phonetic/phonemic details of pronunciation, pauses, overlapping turns, and acoustic or non-verbal aspects of the discourse.

There is therefore no ideal transcription system. Johansson (1995: 97) observes, with regard to decisions taken by the Text Encoding Initiative (TEI), that as there is no "blueprint" for encoding spoken text, the practices that are followed are often subjective. There are essentially two issues at stake: one theoretical and the other practical. The theoretical issue hinges on the argument that "any transcription system is a theory of what is significant about language" (Chafe 1995: 55). This means that the transcriber will represent in visual form those aspects of the spoken language that are deemed to be important about the speech event: the sounds grouped into words, the words grouped into utterances or sentences, the sentences grouped into larger units such as paragraphs and so on. The transcriptions may be marked up with a conventional punctuation system that uses features such as capitalization, full stops, commas and inverted commas. However, it is the ultimate application of the data or the purpose of the transcription that will determine the type of transcription and the conventions selected.

In terms of time and energy this research sought the middle-ground and chose a transcription system that was inexpensive, simple and easily-recognizable. This system employed the established literacy conventions of identifying sounds grouped into word units (unless words had obviously been run together, e.g. *gonna* (going to) or *donno* (don't know)). Beyond the lexical level, non-speech information was added in the form of bracketed annotations such as *[coughs]* or *[laughs]*. Firstly, the recordings were transcribed manually and verbatim from the audio cassettes. This was the longest and most arduous part of the process because, as there were no funds to employ transcribers, the transcriptions and checking of data were all done by myself. However in retrospect, this prolonged initial contact with the recordings provided valuable insight into trends in the data and, in particular, contributed to high standards of quality control. The transcriptions used word-processing software (MSWord) and romanized standard orthography or "absolutely straightforward quick orthographic transcription" (Sinclair 1995: 99). Then all the files were checked against the recordings and converted to plain text to ensure interchangeability between computers with different operating systems and different programs.

Prosodic features were not marked up and phonetic transcription was not used. This is in line with other South African corpora e.g. the XE Corpus, the Corpus of South African English collected for the ICE Project and the proposed spoken language corpora for the nine official African languages of South Africa. The use of minimal annotation means that the corpus files are not 'bloated' and that the raw corpus is available in a simple form.

4.4.4 Speaker identity and speaker turns

The conversation was laid out like the script of a play with each speaker-turn on a new line, even if it consisted of just one word such as *ja* or *sorry*. In addition to the speaker identity (e.g. $10M1H), each speaker-turn was assigned a number starting from *001* for the first utterance in the conversation extract, going up in denominations of *5* for each new speaker-turn. This follows transcription practices of international corpora such as ICE and locally, the XE corpus. This is a useful system as it allows the researcher lee-way to make corrections or insertions during the checking phase without having to re-number the lines. All transcriptions are available on the CD which accompanies this thesis.

4.4.5 Punctuation

The transcriptions were done in lower case, with minimum punctuation. There were no capital letters, full stops, exclamation marks, commas to separate items in a list or clauses, colons or semi-colons. The use of punctuation indicates a level of interpretation by the transcriber, which I wished to avoid. The aim was to leave the spoken texts as open as possible for the users to interpret (Meyer 2002: 74–5). However, question marks were used to indicate obvious questions. For example, the surrounding context and a rising tone at the end of *What Mala and Usha and them all wore?* would have been used to discern that the utterance was a question rather than a statement.

Apostrophes were used for enclitic forms such as *isn't* and possessives, such as *my uncle's car*. Hyphens were used for hyphenated words such as *non-vegetarian*, if the hyphen was a part of the standard orthographic form of the lexical item. Double words such as *gonna* (going to) or *donno* for (don't know) were transcribed as they had been said, provided that they were listed as headwords in either the *South African Concise Oxford Dictionary (SACOD)* (2002) or the *Lexicon of South African Indian English* (Mesthrie 1992*a*). Incomplete words were indicated with an equal sign (e.g. *wed=* for 'wedding').

4.4.6 Pauses

A pause was indicated as: <,> and placed at the point where it had occurred within the conversation. Since a pause may influence the surrounding context, it was decided that the length of the pause should also be indicated. Thus <,> represents a one-second pause up to a maximum of <, , , ,> for a long pause of over three seconds. Vocalized pauses, hesitations and backchannels such as '*er*', '*um*', '*hmm*' were also indicated, if the sound appeared as a headword in *SACOD.*

4.4.7 Letter abbreviations and numbers

Initialisms or letter abbreviations e.g. *ANC* and *HIV* were transcribed *a n c* and *h i v* respectively, but acronyms pronounced as single words (e.g. *Aids* or *Rucus*) were transcribed as complete words (but in lower case) as *aids* and *rucus* respectively. Numbers such as *21* were spelled out as *twenty one,* and word and number combinations such as *History 101* were rendered as pronounced e.g. *history one o one.*

4.4.8 Non-verbal sounds

Human sounds such as coughs, laughter, crying etc. produced by one participant were indicated as third person singular verbs within the speaker's turn thus: *[coughs]*, *[laughs]*, *[cries]*, *[sneezes]*. If both participants produced those sounds simultaneously, they were indicated as nouns e.g. *[laughter]*, *[crying]*, *[coughing]* and placed on a separate line and not attached to any speaker turn. Non-human sounds such as doors banging, dogs barking, bells ringing, cars hooting etc. were also indicated within brackets with just sufficient detail to contextualize the sound e.g. *[dog barks]*, *[door bangs]*, *[telephone rings]* etc. It was important to insert these non-human sounds at the exact point where they had occurred in the conversation in order to preserve the context of the speech.

4.4.9 Non-fluencies

Non-fluencies refer to features of language production such as repetitions, self-corrections (or 'speech repairs') and hesitations which are peculiar to speech alone and which are not found in writing (Nelson 1996: 41). Because the aim was to achieve an utterly verbatim transcription style, non-fluencies were captured exactly as they occurred within each speaker turn and were transcribed orthographically. So repetitions such as *but but but* or the speech repair *he wish he wishes,* were not edited during transcription. However, words that had obviously been mispronounced were normalized (e.g. *bread* pronounced *breed*) to avoid throwing up distracting nonce-forms in the word-frequency lists.

4.4.10 Unfamiliar words and inaudible speech

Although every attempt was made to identify words from the context, occasionally there were unfamiliar or unrecognizable words. These were signalled with <*??*> to indicate that an approximate spelling had been used. Individual words or segments of speech that were unclear, either because of background noise, distance from the tape recorder or poor enunciation, were indicated as *[unclear]*.

4.4.11 Overlapping speech

Crowdy (1995: 230) observes that "Overlapping speech is very common in natural, informal conversation and any marking system has to be easy to employ (so that the transcription rate is not severely affected) and easy to interpret". Previous research has experimented with various methods to represent the turn-taking and overlaps iconically. Blachman, Meyer and Morris (1996: 54-64) devised a table-based visual formula, where each speaker was assigned a column and, within that scheme, individual rows for utterances. Using this scheme, overlapping or simultaneous speech was indicated along the same row within the relevant speaker's columns in a very readable format. Another visually representative method of showing overlapping speech advocated by Ehlich (1993), lays out the conversation like a musical score, with each speaker's contribution represented on a separate line, one below the other. Following this method, overlapping segments of conversation are then positioned directly beneath each other. However, the visual representation of overlaps is often lost when material is transferred between computers with different layout settings. For this reason, it was decided that it was not essential to strive for iconicity in the transcription of the data. It was more important to use a consistent and carefully constructed mark-up system that would indicate the different speakers, the speaker turns and the words spoken.

This study chose the ICE mark-up system for overlapping speech, in line with similar practices followed by existing South African English speech corpora e.g. the XE Corpus (de Klerk, 2006) and the South African component of the ICE Project. Using this system, <{><[> marks the onset of simultaneous speech during the current speaker turn; </[> marks the point where the current speaker stops talking and <[> marks the beginning of the new speaker's turn. The end of the overlapping speech section as a whole is indicated by </{>. In other words, each speaker's overlapped utterance is enclosed between <[> and </[> and the entire segment of simultaneous speech is enclosed within <{> and </{>. This system of mark-up has the added advantages of preserving each speaker turn as a unit, while using the mark-up to indicate the position and extent of the overlap. (Nelson 1996: 41). In the following punctuated extract[6] from the corpus of ISAE, the overlapping sections of the conversation would appear as follows:

> #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.</[></{>

4.5 Analysis of data

During the initial phase of familiarizing myself with the data or "examining the catch" (Barnbrook 1996: 43), I used Wordsmith Tools to produce alphabetical and frequency-based word lists of the corpus data. As the lists were not lemmatized, the alphabetically-arranged lists made it easy to track and analyze individual words and their inflected forms. The frequency-based lists gave useful information on the scope and range of vocabulary use within the ISAE corpus, while at the same time delivering data that would make objective comparisons with other spoken corpora possible.

[6] The original corpus data remains unpunctuated, but illustrative extracts used in this thesis have been punctuated for the reader's convenience.

Mesthrie has done extensive research into ISAE (1992*a*, 1992*b*, 2002 etc.) and its existence as a variety of SAE has already been established (Chapter 3). The function of this corpus was not to study ISAE as a new variety, but rather to discern and describe distinguishing features in the informal speech of a small group of young Indian South Africans who all have similar levels of education, in order to inform the design of a larger corpus of Indian South African English. Features selected for analysis were referenced against lexical data in the *Lexicon of South African Indian English* (Mesthrie: 1992*a*) and the *DSAE Hist*. (Silva et al.: 1996), and various analyses of the syntactic features of ISAE (Mesthrie 1992*b*, Crossley 1987, and Bughwan 1970). 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.

Handling the corpus data in this way, enabled some confirmation of previous research into ISAE (*ibidem*) which has remarked on the preservation of Indian language terminology such as *hardhi*, *bindi*, *roti* and *karo* to denote culturally-specific concepts, the vitality of grammatical structures such as *y'all* for the second person plural and the use of '*but*' in a sentence-final position. However, inasmuch as the respondents were also South African, their speech revealed typical SAE lexical items such as *ja*, *jol*, and *takkies* and these features also merit attention. Finally, it was observed that data such as the choice of swear words and discourse markers, had more in common with speech trends identified in COLT than with those in the BNC. These will be discussed fully in Chapter 5.

Chapter 5: FINDINGS

5.0 Overview

The most important product of this research is the Corpus of ISAE itself, which is offered as a resource for future researchers and which will be made available to peers in the field of linguistics via the Rhodes University website. However in addition, the results of actual analyses of the data within the corpus are important findings. In this chapter, tables together with examples of lexis and syntax, are used to describe and illustrate the general trends and specific features discerned in the corpus of ISAE. Corpus linguistics provided the methodology for data collection within an objective framework and for subsequent quantitative analysis. The quantitative analysis was done using the Wordsmith 5.0 software program which was used for generating frequency lists, Key Word in Context (KWIC) concordances, and for revealing collocational patterns. As the data was not grammatically parsed or prosodically marked up, features that were selected for further exploration had to be investigated in terms of their individual contexts.

The theoretical interpretation of the data drew on various sociolinguistic theoretical frameworks, specifically those relating to language variety (Andersson and Trudgill 1990, Mesthrie 1992*a*, 1992*b*, 1996, 2002*a*, 2002*b*) and discourse analysis (Schiffrin 1987, Romaine and Lange 1991, Fraser 1999, and Barbieri 2005).

This chapter discusses the findings in the corpus in terms of three selected areas. First the conversation trends in the corpus are described in order give a general overview of the kinds of topics that featured in the conversations and to provide some context for the extracts that are cited later (see 5.1); then a closer examination focuses on prominent ISAE features that were observed in the corpus (see 5.2); and finally other distinguishing features of this corpus are discussed in terms of whether they are manifestations of SAE or whether they link more broadly with patterns in general English (see 5.3). It is believed that this approach helps to situate the data in terms of the sub-variety (ISAE), the local variety of world English (SAE) and in terms of emerging international trends in spoken English.

5.1 Conversation trends

A total of 37 conversations were recorded, but ultimately only 30 were selected to build the corpus (see 4.3.2). Before trying to identify individual lexical or syntactic features in the corpus, it was important to gain a sense of the corpus as a whole. This was done by examining the corpus in terms of speaker word counts, by looking at the conversation trends or topics and then by looking more closely at features such as word frequency and syllable length. This approach provided useful indices for comparison with other corpora of spoken English (e.g. the local XE Corpus) and with relevant international corpora (*viz.* the BNC and COLT) which were collected in the UK.

5.1.1 Speakers and word counts

As there were 30 conversations, each involving two participants, the corpus could potentially have yielded contributions from 60 individual speakers. However nine contributors ($02M2H, $03F2T, $05F2H, $07F2G, $08M3T, $38F2H, $42FIH, $44F2T, $49M1E) also functioned as data collectors, making the total number of corpus contributors 49 (Table 7 overleaf).

Since the data collectors participated in an average of two conversations each, it will be immediately apparent that the word count for their contributions is slightly higher than the general average for each speaker (Table 7 overleaf). While strictly speaking this is not ideal, only speaker $02M2H, who participated in a total of four conversations, exceeds the average word count noticeably with 4317 words. In the initial stages of the research, speaker $02M2H served a valuable function of gaining access to the social network used for this study. His involvement was therefore crucial in helping to explore and expand contacts that might otherwise have been difficult to obtain.

Table 7: Individual speaker word counts

SPEAKER	WORD COUNT	SPEAKER	WORD COUNT
$01F2T	917	$26F2A	1313
$02M2H	4317	$27M2G	934
$03F2T	2321	$28F1T	695
$04M3H	929	$29M2T	1662
$05F2H	2146	$30F1T	937
$06M2T	886	$31M3A	1535
$07F2G	2514	$32F1T	1085
$08M3T	1886	$33M1A	498
$09F2T	723	$34F1H	1014
$10M1H	947	$35M1G	1345
$11M1G	985	$36F2A	366
$12F1U	782	$37M1H	779
$13M1G	590	$38F2H	1356
$14F2G	1143	$39M1H	1244
$15M2H	1279	$40F1H	1836
$16F2H	866	$41M2T	759
$17M2I	1212	$42FIH	1982
$18F1G	1156	$43M2T	863
$19M2H	1053	$44F2T	2091
$20F1M	855	$45M2H	1085
$21F1H	799	$46F1H	876
$22F2H	839	$47M2U	912
$23M2A	1048	$48F3T	691
$24M1M	1135	$49M1E	2034
$25M2H	1363		

Milroy (1987), in a review of various data-gathering techniques for sociolinguistic purposes, recommends the value of employing field workers who would be perceived as “insiders” or members of the in-group because, firstly, this approach takes cognizance of pre-existing social groups and secondly, it has been found that using members of this group to elicit the data, is successful in procuring data that is less constrained by the desire to adhere to external linguistic norms. In this regard Mesthrie (2002) has commented that ISAE is a *covert* badge of identity (my emphasis) and that speakers in this community select more careful styles in public or in the presence of anyone who is not a member of the speech community. In a tacit acknowledgment of the integrity of the social network and following precautions advocated by Gumperz (1970), I was careful to avoid contact with contributors that the fieldworkers had identified, as I felt that my age and professional status made me to some extent an ‘outsider’ and that this might constrain the use of natural language within the group (see 4.2.1). For obvious reasons, contact with the 9 fieldworkers who also contributed to the corpus was unavoidable.

With a closely balanced gender distribution in the corpus of 51% male speakers and 49% female speakers, it was feasible to compare the overall word count from each of these two groups of speakers. Word counts of 31280 (male speakers) and 29579 (female speakers) respectively, indicate a fairly (if not miraculously) even distribution of word contributions. However, a closer investigation of the word counts in the ten mixed gender conversations for females and males was not so closely matched. In the mixed dyads, the female participants spoke slightly less than their male counterparts, contributing 8748 words (42,7%), compared with the 11730 words (57,3%) of the male speakers. Since this thesis does not focus specifically on gender-based comparisons, this discrepancy was not regarded as serious.

5.1.2 Conversation topics

The data in the ISAE corpus were extracted from informal conversations rather than from formal interviews or elicitation exercises. The conversations therefore encompassed a range of topics and participants were free to switch between these with ease in a manner that is typical of informal, relaxed conversations in intimate settings.

Figure 2 below gives a gives a "broad brush stroke picture" of the range of topics covered, and the frequency figures give a general indication of how often they cropped up.

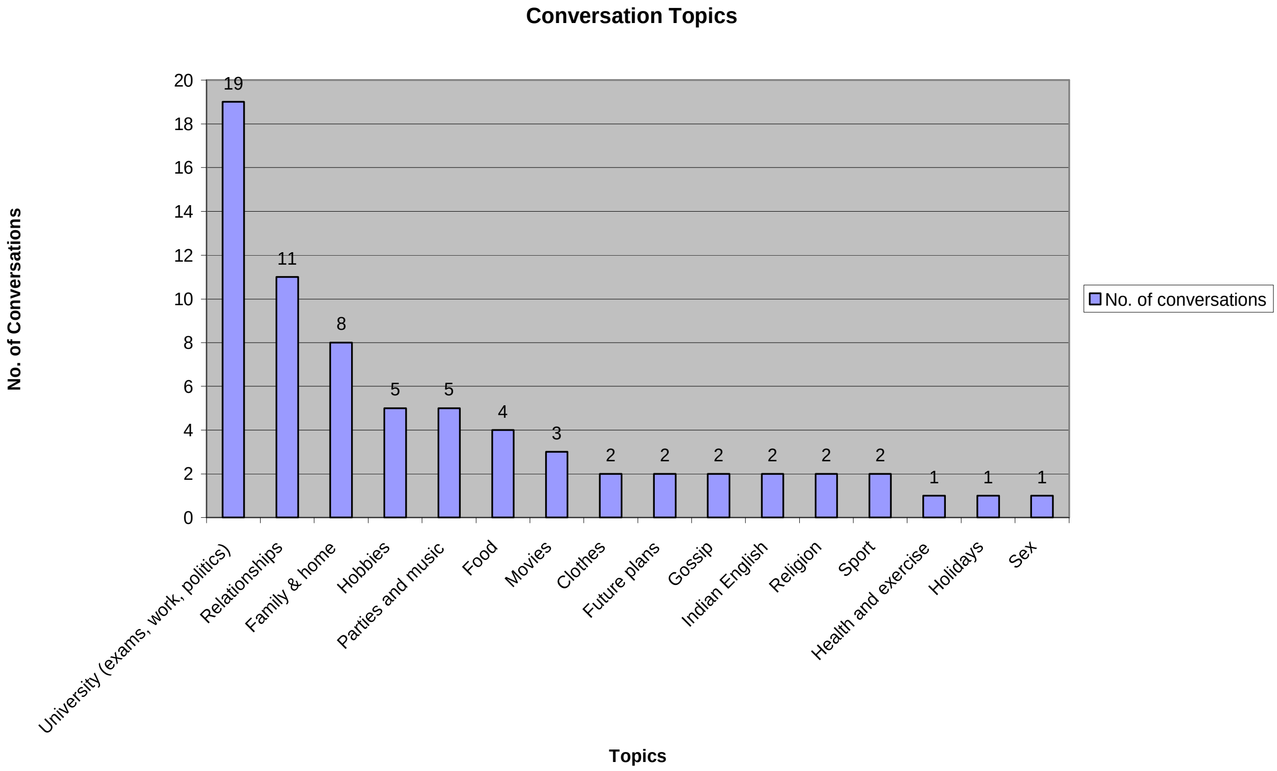

Figure 2: ISAE conversation topics

Table 8 (pg. 79-81 following) reveals more detail by listing the key lexical items associated with the topics. The frequency figures obviously do not take account of the number of times pronouns were substituted for nouns, but they do give an idea of some of the specific lexis associated with the topic areas.

Table 8 : Summary of popular conversation topics and associated lexical items

Topic	**Frequency**
University and studying	
Exams/Lectures/Assignments/Pracs	125
Names of Courses	53
Pass/Fail/Borderline/Accomplish/Achieve	42
Campus (Labs, Library)	38
Bursaries/Books	27
Lecturers/Tutors (names of lecturers/tutors)	26
Academic/ally	9
Applications	8
Sub-total	**328**
Partying and drinking	
Drinks/Drinking/Drunken	65
Party/Jol/Jolling/Birthday/Dancing	41
Cider/Cocktails/Wine/Beer/Rum/Vodka/Cane/Champagne	25
Clubs (names of various nightclubs in Grahamstown)	24
Bottles	12
Social	8
Caps (referring to drugs)	5
Cigarettes/Smoking	4
Drugs	2
Sub-total	**186**
Sociopolitical Issues	
Indians	47
Durban/Durbs	44
Whites/ Blacks/ Coloureds	36
South Africa, Zimbabwe, African	16
India	11
Afrikaans	6
Apartheid/Prejudice/Constitution	5
Sub-total	**165**

Topic	**Frequency**
Hobbies	
Movies/Cinema/Bollywood	57
Soccer/ Cricket/ Swimming/ Sport/ Exercise/ Gym	40
TV	10
Clothes	9
Reading, Books, Magazines	8
Sub-total	**124**
Family	
Brothers / Sisters	42
Parents	23
Mother/ Mummy/ Mum/ Mom	22
Dad/ Daddy/ Father/ ‘Ballie’	17
Cousins	11
Aunty/Uncle	10
Wife/Husband	8
Grandparents/ Granny/ Grandfather	5
Sub-total	**138**
Meals and Catering	
Breakfast/Lunch/Supper/Meals	22
Mushrooms/Bread/Meat/Eggs	21
Food	16
Bunny chow/Biryani/Roti/Soji	15
Curry/Curried	12
Vegetarian/Vegan/Vegetables	10
Sub-total	**96**

Topic	**Frequency**
Religion, Language and Cultural practices	
Islam/Muslim/Hindu/Christian/Church/Mosque	20
Incense, Pray, Prayer	12
Sari	10
Wedding, Hardhi, Bindi	8
Diwali	7
Tamil/Gujarati/Hindi	7
Dawah/ Ramadan/ Fasting/ Sehri	6
Christmas	2
Sub-total	**72**
Romance and Relationships	
Boyfriend/Girlfriend/Dating/Love	38
Marry(ied)	13
Condoms/Glove	11
Babe, Baby, Baby Doll, Hon, Munchkin	8
Sub-total	**70**

As can be expected of university students in an academic setting, their studies and university matters in general were recurrent themes, discussed with varying levels of passion and intensity, as extract (1) below shows:

(1)

<#05:$04M3H:245> Because every time I been doing experiments, experiments, experiments [pause]. I been fuckin' totally leaving this out and when I go home, I'm fast asleep. The other ous go home and do literature research. I've not been doing that. Now look at this. I'm a bit fucked.

#05:$02M2H:250> Ja literature research you can write very, very fast. It's a matter of writing up in your case. My case, we can't write up very fast. We haven't got=

<#05:$04M3H:255> No, no the thing is Abi, writing up for me is very hard. Experiments all, I can do it.

<#05:$02M2H:260> Ja no time to write.

<#05:$04M3H:265> But writing up I'm not very good at. I'm I'm like [pause] I'm not very good at these things [pause]. I was never good at it.

While many of the discussions centered on academic work and matters relating to life at university, this was balanced by other discussions which reflected the contributors' social activities. These activities involved drinking, dancing, partying (or *jolling*), dating, watching soccer (amongst the males), going to the gym (amongst the females) and watching movies. The Bollywood movies which were popular with the female participants, were occasionally the objects of gentle derision or self-deprecatory humour by the males, as text (2) demonstrates:

(2)
<#36:$46F1H:005> "Main Hoon Na"[7].
<#36:$49M1E:010> Hmm "Matrix" copycat. They tried, they tried.
<#36:$46F1H:015> [laughs]
<#36:$49M1E:020> Poor people tried. Doesn't work like that. You should tell them: I should become a director.
<#36:$46F1H:025> [laughs] Ohh no!

The exchanges on sociopolitical issues embraced current affairs in South Africa, as well as reflections on the country's past, the legacy of apartheid, personal family histories and the history of the Indian community in the country. Unlike the XE Corpus, in which Eastern Cape and Grahamstown news items (e.g. the National Arts Festival, unemployment and the lack of water-borne sewerage in the townships) featured prominently (de Klerk 2006: 56-57), the ISAE corpus did not contain references to such local issues. The differences in topic choices between these two corpora could be related to the fact that the speakers in the XE corpus were drawn mostly from the Eastern Cape and from Grahamstown in particular (*ibid*: 44), while the speakers in the ISAE corpus were all from KwaZulu-Natal. Although the latter group were students of Rhodes University, they were essentially 'in transit' in Grahamstown and the Eastern Cape. The most popular topical issues in the ISAE corpus were therefore geographically linked to

[7] A popular Bollywood movie.

events in Durban and KwaZulu-Natal. On a social level, issues involving family ties, family obligations and family politics featured regularly in the conversations. Extract (3) gives an indication of the candour and vigour with which participants discussed such topics:

> (3) <#23:$29M2T:170> It's our own cousins and they making her life so terrible. Ja now if Kessie does anything they have a problem. If they see him somewhere with a girl they have a problem. When Lesley came for the house prayer they had a problem with it. And I was like jeez we don't mind if at all
>
> <#23:$36F2A:175> So are you like supposed to like limit your friends and choose them select them on the basis of what your cousins think?
>
> <#23:$29M2T:180> Exactly [pause] and the thing is it's not like we mind what they do.

Predictably, conversations containing references to cultural practices around religion, festivals and food contributed the greatest number of lexical items of Indian origin such as *Ramadan*, *Diwali*, *hardhi* and *biryani*. Apart from this culturally orientated lexis, there were features of ISAE in every conversation, with the second person plural *y'all*, the use of '*of*' in partitive genitive constructions and the placement of '*but*' in a sentence final position, being the most pervasive (see 5.2). In terms of style, the conversations ranged from casual chats between friends, to intimate exchanges between partners. There was also a noticeably higher incidence of slang lexis (e.g. *ous*, *lakker*) and swearing (e.g. *fuckin'*, *bladdy*) in the exchanges between exclusively male participants (see 5.3.2)

5.1.3 Word frequency

An unlemmatized frequency list of all items in the corpus revealed that there were 4301 words (tokens) ranging in frequency from 1983 occurrences for the word 'I', to one occurrence for 'zone'. A comparison of the top 100 words on this list with a similar list from COLT was useful in delineating the distinctive features of each list. COLT was selected for comparison as it represents informal conversational data, 70% of which was collected from urbanized teenagers and young adults ranging in age from 14-19 years of age (Stenström et al. 2002: 19-20). The investigation revealed that although the actual rank order of the words did not match exactly, there were noteworthy points of similarity in the lists (see Table 9 below).

Table 9: Top 100 words in COLT and ISAE corpus

COLT			**ISAE**		
1-33	**34-66**	**67-100**	**1-33**	**34-66**	**67-100**
you	not	*you're*	you	then	him
I	me	he's	I	there	gonna
unclear	this	*going*	the	not	come
the	my	gonna	and	have	or
nv	well	think	*ja*	on	why
and	one	laughing	to	me	are
it	I'm	*goes*	it	do	two
a	go	him	like	ok	thing
to	up	*I've*	that	unclear	*were*
yeah	*erm*	er	a	now	she’s
that	get	ee	was	*hey*	about
what	your	why	no	with	from
no	all	come	so	her	well
in	we	say	in	your	see
know	are	said	he	all	had
he	be	now	know	that’s	*time*
of	with	them	is	at	must
laugh	or	how	but	go	cos
it's	that's	or	she	up	as
oh	can	mean	my	when	*hmm*
is	then	can't	they	get	did
like	there	as	it's	be	didn’t
on	right	off	what	out	his
do	about	here	this	got	*went*
was	if	she's	of	how	too
don't	cos	look	just	if	people
got	out	good	don't	can	can’t
have	really	two	oh	think	said
just	name	okay	for	*year*	mean
so	did	who	we	right	really
but	at	*down*	er	here	*huh*
she	her	want	one	he’s	them
they	when	had	I’m	because	tell
		some			want

The most significant finding arising from the comparison was the 74% overlap in the top 100 words on the two lists. A common core of 74 words indicates that despite cultural, geographic and time differences between the two corpora, the most popular words used in the casual conversation of these two groups of young people is very similar. In addition to common lexical items, the two lists revealed that laughter, hesitation cues (*er* or *erm*) and unclear dialogue typified the spoken language in both corpora. After identifying the common lexical core in the two lists, the words which were unique to each list were italicized for further investigation (see Table 9 above) as it was felt that this would contain valuable information about the distinguishing features of each list and the population it represented. They are discussed below.

On the ISAE list the unique words *ja*, *hey*, *year* and *time* need some explanation. In the South African speech context, *ja* and *hey* are common features of colloquial speech in all sub-varieties of SAE. *Ja* (directly from Afrikaans, but fully assimilated into SAE) is used to mean *yes*, *I understand*, or *I see* (Silva et al. 1996). Its equivalent in the COLT list is *yeah*, which does not occur in the ISAE list at all. In the ISAE corpus, at the beginning of an utterance, *hey* is often used as an exclamation or interjection to secure the interlocutor’s attention (see example (4) below). In an utterance-final position, *hey* is used to underscore the preceding utterance and if accompanied by a rising inflection, it turns a statement into a rhetorical question as in example (5) below.

(4) <#05:$04M3H:1005> *Hey* this Pharamacol doesn't interest me, man.

(5) <#36:$46F1H:005> *Ja* but it tastes really nice, *hey*?

The word *year* occurred most frequently in the collocation ‘last year’, ‘this year’, ‘next year’, ‘first year’, ‘second year’ and ‘third year’. The elliptical form of the pseudo-title ‘first year’, ‘second year’ was commonly used to refer to students by their year of study as in, “Ja, she was a *first year*…” and in extract (6) below. Almost every conversation featured some reference to student life as a phase and revealed a pre-occupation with what they were undertaking that year, what they had done the previous year, and what they proposed to do the following year, as the following texts illustrate:

(6)

<#08:$16F2H:810> Ja, she was a *first year* in Wakefield *last year*, on my corridor, Daksha, and then she moved *this year* to the other campus.

<#31:$41M2T:825> We don't write tests in *third year*, [we] only write two tests throughout the whole *year*.

<#31:$39M1H:1330> Ja I'm trying to think of words to tell him [pause] but er… [coughs] You got two more *years* here? Or one more *year* [pause] *this year* and *next year*?

References to *time* figured in the set expressions *at the time*, *the first time*, and *in a year's /two years' time*. However, in addition to these commonly-used collocations of *time*, it was also featured as a quasi-postposition (Mesthrie 1992*b*: 105, 179), as examples (7) to (9) from the corpus illustrate. A quasi-postposition is the insertion of a word (*time* in this case) after the preceding noun. Mesthrie explains that this is a fossilized basilectal usage, indirectly transferred from the substrate Indic and Dravidian languages. In examples (7) to (9) the word *time* is unstressed, and performs the functions of either a preposition or adverb in Standard English. Although there were only 10 examples in the corpus, the usage was distributed amongst a range of speakers, indicating that adherence to this form is still prevalent amongst some speakers.

(7) <#33:$45M2H:000> No waarheid bra [pause]. Don't act like you didn't try it one *time*? (=No waarheid bra [pause]. Don't act like you didn't try it *once*?).

(8) <#09:$05F2H:1431> Roxy's is closed December *time*. (=Roxy's is closed *in* December).

(9) <#03:$09F2T:1245> You must wear it summer *time*. (=You must/should wear it *in* summer.)

The contracted forms *I've* (I have) and *you're* (you are), which featured amongst the top 100 words on the COLT list, were noticeably absent in the ISAE top 100 words. A search for them in the ISAE corpus revealed only 29 occurrences of *I've* (ranked lower down at 257), and only 23 occurrences of *you're*, (even further down at 329 in the total rank order). The lower incidence of *I've* as a reduced form could be partially explained by the substitution of other grammatical forms in the ISAE corpus, notably *I got*, and the omission of the reduced *–'ve* (*have*) in the present participle forms of the verb as examples (10) to (12) illustrate:

> (10) <#09:$08M3T:770> Please man [pause]. *I got* work to do. (=Please man [pause]. I *have* work to do).
>
> (11) <#29:$42F1H:005> And I *been* living eighteen years in Tongaat. (=And *I've* been living in Tongaat [for] eighteen years.)
>
> (12) <#36:$49M1E:000> Hey, actually *I seen* worse, trust me. (=Hey actually *I've* seen worse, trust me.)

There is also a possibility that audibility on the tapes compromised the accuracy of the transcription, although every effort was made to remain true to the data. In the case of *you're*, it has been observed that ISAE is a non-rhotic dialect and that it is typical to omit the reduced auxiliary *–'re* (*are*). Therefore the omission of *–'re* is more than just the economy of production typical of speech. It is particularly noticeable if the initial sound of the word following the auxiliary is a consonant, as shown in examples (13) and (14) below. According to Mesthrie, occasional auxiliary deletion of this sort is one of the features that ISAE shares with other New Englishes, such as African American Vernacular English (AAVE) (1992*b*: 49).

> (13)<#10:$10M1H:925> *You* too close (=*You're* too close).
>
> (14) <#12:$18F1G:240> Ja one day *you* bound to get caught (=Ja one day *you're* bound to get caught).

5.1.4 Syllable length

The top 63 most frequently occurring words in the ISAE corpus were all one-syllable words, a finding consistent with Zipf's law of the "principle of least effort" in natural language use (Zipf 1949). This means that people prefer to use shorter words more frequently, rather than longer words. In the whole of the BNC (spoken and written components) it was found that, with the exception of the most common word *the*, frequencies for top ranking words in each syllable grouping from two to five were roughly twice that of the top ranking word in the next syllable group (Leech et al. 2001: 121). In the much smaller ISAE corpus which consists of only spoken conversational data, the decline in frequency between the syllable groups was much more dramatic. The frequency of the top-ranking word in each syllable group was three to four times greater than the top-ranking word in the next syllable group, as illustrated in Table 10 below. The conversations in the ISAE corpus were limited to examples of informal, spoken language collected from a very narrow age band (18-29 year olds), while the full BNC was composed of formal speech (e.g. sermons, lectures), informal speech and writing, collected from broad age band (15 to over 60). When set against the discourse context and demographic confines of the ISAE corpus, it would be reasonable to assume that the higher incidence of polysyllabic words in the BNC is attributable to the differences between the two corpora as described above.

Table 10 : Comparative frequencies of top-ranking words in successive syllable groups in the ISAE corpus

Syllable group	**Most frequent word**	**Frequency**
2	because	177
3	remember	66
4	everybody	13
5	quantification[8]	4

5.2 ISAE features observed in the corpus

Mesthrie's extensive research into ISAE (1992*a*, 1992*b*, 1996, 2002*a*, 2002*b*) has already established it as a legitimate rule-governed sub-variety of SAE. The function of this corpus was to see the collection as the first stage towards a larger endeavour which will hopefully ultimately result in the establishment of a fuller corpus of ISAE comprising

[8] *quantification* is an unusual word in casual conversation. Closer investigation revealed that the word occurred in the context of a course-work related discussion between two science students.

data extracted from a greater demographic and contextual range. Analyzing this data enables some corroboration of earlier descriptions and allows for the emergence of previously unremarked-on features. The data from this modest attempt could therefore constitute the first piece in the construction of the larger mosaic of ISAE. Its focus on a young cohort of speakers also offers a valuable window on current trends. As has already been illustrated in the examples relating to the frequency data, the corpus revealed several features of ISAE identified by Mesthrie.

Mesthrie (1996: 88-89) has identified three robust grammatical features that are observable across all ISAE lects from the basilect to the acrolect (see 3.5). All three were observed in varying frequencies in the corpus and will be discussed below. These features are:

- the use of '*y'all*' as a second person plural pronoun,
- the copula attraction to '*wh-*' in indirect questions and
- the use of '*of*' in partitive genitive constructions.

5.2.1. The use of 'y'all' as a plural pronoun

The corpus word list revealed 38 examples of *y'all* (the pronunciation rhymes with "ball") used as a plural pronoun and one example of the genitive equivalent *y'all's* (spread over 12 of the 30 files) in the corpus. The use of *y'all* was evenly divided between statements and questions. Mesthrie remarks that it is one of the "syntactic features that the acrolect shares with the basilect and mesolect…which is below the level of social consciousness for most ISAE speakers" (1992*b*: 61). While this exact form is not used in other varieties of SAE, some Afrikaans speakers employ *youse* to indicate the second person plural pronoun. Non-standard forms such as *y'all* and *youse* represent creative solutions to specify the second person plural, since standard modern English does not have separate lexical forms to distinguish between the singular and plural forms of *you*. In modern English, the one lexical form *you* does double duty for the archaic pronouns *thou* (singular) and *ye* (plural). In the face of this lexical limitation in standard modern English, context has to be harnessed to disambiguate the singular and plural meanings of the second person pronoun. In this regard Crystal (2004: 450) comments that dialects which allow *y'all* and *youse* are therefore "richer in their possibilities of expression than Standard English".

The following are some examples of the use of *y'all* found in both questions and statements in the corpus. In 11 of the 15 <u>questions</u>, *y'all* co-occurred with other ISAE syntactic features as well: deletion of the auxiliary verb (a–g) and the use of the simple past tense instead of the perfective form (l–o). In the statements 16o. reveals what looks like an interesting attempt at self-correction where the speaker uses *y'all* and immediately substitutes *you*, while obviously still intending the plural form of *you*.

(15) <u>Questions</u>

a. And er when [are] *y'all* writing tests?

b. [are] You sure *y'all* don't wanna call Biker Mehmood?

c. When [are] *y'all* starting that?

d. So [are] *y'all* just jolling now?

e. How [do] *y'all* know each other?

f. Where [are] *y'all* leaving me?

g. [do] *Y'all* stay in the same res?

h. Like do they help *y'all* at all ?

i. How many people are in *y'all's* maths class?

j. . And did *y'all* finish it?

k. But when do *y'all* wanna sleep over?

l. What time *y'all* <u>finished</u> off here? [did…finish]

m. *Y'all* <u>me</u>t him before? [have…met]

n. *Y'all* <u>did</u> it as group? [did…do]

o. Where *y'all* <u>met</u> him? [did…meet]

(16) <u>Statements</u>

a. Oh *y'all* not=

b. I was telling Pam *y'all* must come over.

c. Y*'all* all got together.

d. I think *y'all* did one more section.

e. It’s so cool *y’all* got two windows.

f. Sit and watch *y’all* from the top. [laugh]

g. No but still, *y’all* can just come.

h. *Y’all* picked me up.

i. Ok – cool so *y’all* are very closely related.

j. As long as *y’all* [are] still talk[ing].

k. But then the tut on what *y’all* [are] doing no=

l. Poor people, the noise *y’all* create!

m. Oh *y’all* you stayed in Parlock.

n. Ja, *y’all* [are] quite good.

o. Ok now, *y’all* you decide [Possibly a case of self-correction?]

Apart from its grammatical function to indicate the plural form of *you*, Crystal (2004: 451) elaborates that the use of *y’all* occurs in various informal American speech contexts and his analysis of those occurrences indicates that its pragmatic function is to express warmth and to signal “familiarity, friendliness, informality, and rapport, at least among young people”. The speakers represented in the ISAE corpus were all well known to each other (members of an existing social network) and their degree of intimacy ranged from those who were close friends to those who were partners, as was evident from the nature of the topics that were candidly discussed (see Figure 2 above) as well as the terms of familiarity (*bru*, *bra*, *ou*, *dude*, *ekse*) and endearment (*love*, *babe*) that were employed during the conversations. Therefore the use of *y’all* in the speech extracts quoted above could also be a pragmatic expression of kinship amongst members of the same student and speech community. The modern informal equivalent originally from spoken American English, but now used globally, is *you guys*. It refers to “members of a group regardless of sex” (Merriam Webster online http://www.m-w.com/cgi-bin/dictionary). The corpus revealed a very low frequency of *you guys* (four instances) used by only two speakers to designate the second person plural pronoun. The speakers who used *you guys* did not use *y’all*, which suggests that in the case of these two speakers *you guys* may have ousted *y'all*. The choice of *you guys* is in keeping with global speech trends for this age group.

5.2.2 Copula attraction to wh- in indirect questions

Crossley (1987) and Mesthrie (1992*b*) have both identified this phenomenon as one of the distinguishing features of ISAE. This form occurred with a very low frequency (6 occurrences used by 6 speakers) in the corpus. All the instances are given in (17) below:

(17)
a. <#01:$03F2T:655> I was there and then I was trying to explain to Lal like *where's* the digs.
b. <#32:$38F2H:120> But she's er ok. Ja I donno *what's* her personality like.
c. <#06:$06M2T:1185> So you can see, *where's* it – where it entered.
d. <#08:$02M2H:1145> I wonder *what's* she like now, huh?
e. <#21:$27M2G:095> Ja, South Africa is really cool when it comes to things like that. Like I was reading that even in Egypt people can't like do er dawah. You know *what's* dawah like?
f. <#33:$45M2H:000> Hey I donno *what's* her name huh.

Notably, all occurrences in the corpus revealed the copula in its contracted form. Crossley (1987) has questioned whether the copular attraction to *wh*- forms occurs in past tense forms as well as the present tense. Grammatically, example 19a would certainly require the past tense *were* for the second verb (<#01:$03F2T:655> I was there and then I was trying to explain to Lal like *where* the digs *were*). The small size of this corpus limits broader generalizations about its continued use in this speech community.

5.2.3. The use of '*of*' in partitive genitive constructions

A total of 13 occurrences of this feature were observed in the speech of 12 speakers. Although this feature is not found in other sub-varieties of SAE, it occurs frequently in the Indian English of the subcontinent (Nihalani et al. 1989). In this connection, they observe that *of* with words like "much" (too much *of* noise), "enough" (enough *of* money), “less” (less *of* drunkenness) and “little” (little *of* overt criticism), was a common syntactic structure in early nineteenth century British English and that it “has been maintained in I[ndian] V[ariant] E[nglish]” (Nihalani et al. 1989: 129). Mesthrie, on the other hand, argues that a similar grammatical construction occurs in French, and postulates that its occurrence in ISAE may be fossilized evidence of English language models acquired from contact between French missionaries and the Indian immigrants in South Africa (Mesthrie 1992*b*: 21; Brain 1983). Given the models of English language that were available for Indian learners of English, on the sub-continent and in South (see 3.4.3), both theories are indeed plausible. The corpus examples did not reveal the partitive genitive construction in any form other than *much of*. However *much of* co-occurred with ‘money’, ‘work’, ‘effort’, ‘shit’ and ‘weight’, as illustrated in examples (18) to (21) below:

(18) <#02:$05F2H:755> Did you ever think you'll make so much *of* money?

(19) <#31:$41M2T:005> Too much *of* fuckin' shit to do.

(20) <#03:$07F2G:900> You know I put so much *of* effort into it.

(21) <#31:$39M1H:000> I told her I got too much *of* work to do.

5.2.4 Topicalisation

Comparative studies by Mesthrie on topicalisation have revealed that this feature appears to be higher in ISAE than in other SAE speech communities, leading him to comment that ISAE speakers have a “predilection” for this type of construction (see 3.6.2) and to conclude that the sub-variety as a whole seems to have “the properties of topic-prominent languages” (Mesthrie 1992*b*:123). In the absence of grammatically parsed and marked-up data, electronic searches for this grammatical construction were not possible, so these were noted as they manifested themselves during searches for other features. Thus although actual statistical values are not available, the feature was observed in the speech of several contributors. Examples (22) to (25) are characteristic instances of the construction:

(22) <#10:$08M3T:595> Baby, *the geyse*r you put it off?

(23) <#05:$02M2H:250> Ja [pause] *literature research* you can write very, very fast.

(24) <#05:$04M3H:255> No, no the thing is Abi, writing up for me is very hard. *Experiments* all I can do it.

(25) <#32:$38F2H:000> No, *my tumblers* I need, not my plate.

5.2.5 '*But*' and '*like*' as reinforcing tags in clause-final positions to mark contrary pre-supposition

There were 9 occurrences of *but* fulfilling this function, roughly evenly divided between statements and questions. The only two occurrences of *like* in a clause-final position are included with this list, because *like* appears to be fulfilling the same function as *but*, in performing a type of "retroactive focusing" (Miller and Weinert 1995) and to counter possible opposition. Miller and Weinert's study, which analyzed the use of *like* in two bodies of Scottish speech data, is useful in recalling the speculative comment by Mesthrie (1992*b*: 21) that some forms in present day ISAE might derive from the dialect forms of early missionaries (Chapter 3). In this regard, Mesthrie specifically cites the use of *like* and *but* in Scottish and some northern English dialects (Mesthrie 1992*b*:21; 108). In the ISAE corpus, the use of *like* and *but* as reinforcing tags in clause-final positions was observed in the speech of ten contributors, typically as exemplified in (28) below:

(28)

a. <#02:$03F2T:975> I like to dance. I like to dance *but*.

b. <#08:$16F2H:530> All the Muslims shouldn't have been there *but*.

c. <#32:$43M2T:005> I don't drink *but*.

d. <#35:$19M2H:005> No like I had my boxers on *but* [pause] I thought it was inappropriate, so I put my jeans back on.

e. <#37:$25M2H:000> I donno how long it'll last *but*.

f. <#04:$12F1U:125> Don't you love spinach *but*?

g. <#36:$46F1H:255> Ja you you vegetarian? Don't you find= Do you eat eggs *but*?

h. <#32:$38F2H:160> That was the fun part *but* huh?

i. <#33:$45M2H:000> That ou can dance pantsula lakker *but* huh?

j. <#12:$18F1G:1015> But will you will you be able to have guys *like*?

k. <#08:$16F2H:260> Hmm abbreviations *like.*

5.2.6 Associative plural: *'and them'*

There were 10 examples of the associative plural *and them* which co-occurred with human nouns in the corpus, typically as follows:

(29)

a. <#20:$23M2A:1530> Ha your Daddy and Mummy *and them* were here.

b. <#03:$07F2G:930> Kamilla *and them* ordered a pizza for me.

The associative plural has been traditionally associated with African American and creole varieties of English, but more recently Tagliamonte (2000) has commented on the use of this structure across six varieties of English in North America and Britain. She observed the form in North America in Gullah (in the Sea Islands, off the coast of Georgia), in North Preston and in the Guysborough enclave (in Nova Scotia); and in Britain in York, Wheatley Hill (a village in northern England) and Buckie (in Scotland). Tagliamonte (*ibid*: 409) concludes by postulating "a plausible British source for…the associative plural that until now has been associated with creoles and/or creolization." Mesthrie (1992*b*) discusses the use of the associative plural in ISAE as a distinct feature of mesolectal (noun + *and them*), and basilectal (noun + *them*) speech, but concedes that it also occurs in other colloquial varieties of SAE. In the light of Tagliamonte's cross-variety findings, it is possible that a combination of sociolinguistic and historical linguistic investigative frameworks might usefully shed more light on the origin and prevalence of the associative plural in ISAE.

5.2.7 The use of *'never'*

There were 73 occurrences of *never* in the ISAE corpus (0.12% of 60 000 words) as opposed to 357 in COLT (.07% of 500 000 words). This meant that *never* featured almost twice as often in the ISAE corpus as in COLT. An investigation of the functions of this item in the ISAE corpus revealed that in addition to the meaning "not ever", *never* was employed as a standard past tense negator for a number of verbs, as the following examples (30) – (32) illustrate:

(30) <#04:$12F1U:695> I *never* went for that thing. (=I *didn't* go to that thing/function)

(31) <#19:$30F1T:675> Ok after that we went for Afrikaans tuition because we *never* understood Afrikaans. (=Ok after that we went for Afrikaans tuition because we *couldn't* understand Afrikaans)

(32) <#35:$49M1E:000> And you still *never* returned it to me till today. (And you still *haven't* returned it to me until today).

Although frequently regarded as non-standard (Labov c1972) and proscribed by upholders of formal style in all contexts, the use of *never* as past tense negator is in fact typical of hyper-colloquial English in general.

5.2.8. Use of '*it*' with plural noun phrases

Crossley (1987: 160) has commented on the use of the indefinite singular pronoun *it*, to replace plural nouns or plural noun phrases in data that she collected in a semi-formal test situation. This corpus revealed 15 occurrences of this phenomenon in the speech of 11 speakers, illustrated in the following selected examples:

(33) <#16:$26F2A:530> Her vertebrae [3-second pause] three four five right here just here and *it* was like disintegrating.

(34) <#17:$17M2I:000> …It's different from what we do in the pracs. Normally we have one week preparation where you can go over, find the ingredients whatever and use *it* .

(35) <#02:$05F2H:565> … These are the funny bottles. *It* doesn't have the part at the bottom

(36) <#23:$29M2T:580> …She took all the things from the car and left *it* back in our house.

(37) <#03:$09F2T:905> Black scarves?
<#03:$07F2G:910> No *it's* cream.
<#03:$09F2T:915> Oh both are cream?

5.2.9 Near misses

The term “near misses” is a catch-all term under which Mesthrie (2002: 352–354) groups a number of forms used by mesolectal speakers in targeting the standard form but which fall just short of it because of the preposition, adverb, adjective or noun selected, or because of the unusual way the idiom has been rendered. Mesthrie has concluded on the basis of these observations that the same speakers do not use these forms in “monitored speech”, which indicates that they are aware of the standard forms, but that in unguarded moments or in stressful situations, the standard idiomatic form is subtly changed. He suggests that these indicate the speaker’s position on the basilectal to acrolectal continuum as being not quite acrolectal. The fact that speakers slip into these forms unconsciously also indicates the dynamic nature of the speaker’s position on the polylectal continuum. The following are some examples of “near misses” that were observed in the corpus:

Preposition substitution

In the selected examples below (38) – (40) the set expression resembles standard English in all respects except for the choice of preposition. These are not instances of idiosyncratic usage as they are employed by speakers across the spectrum of lects, even by those who are capable of applying the standard form in formal situations or in the company of those who are not members of the ISAE community. In such cases the deliberate choice of the non-standard preposition may be associated with a desire to confirm membership of the speech community by not sounding too acrolectal and distant.

> (38) <#17:$22F2H:005> … but *in* school I used to play. (=but *at* school)
>
> (39) <#32:$43M2T:265> …so he went *by* the fountain and he leaned there. (=so he went *to* the fountain)
>
> (40) <#28:$40F1H:090> … about two hundred two hundred max will come *for* the wedding. (=*to* the wedding)

Non standard use of adverbials, adjectives, quantifiers

The same reasoning could be applied to the choice of non-standard adverbials, adjectives and quantifiers in (41) below.

(41)

<#02:$05F2H:660> So it means if you did *bad* in June you got sixty? (=did badly)

<#36:$46F1H:355> *For true*? (=truly/really?)

<#03:$07F2G:600> I think maybe January will be better – it's *more calmer* (calmer)

Lexis and idiomatic adaptations

Some of the "near misses" involve what appears to be the creation of novel lexis through the telescoping of two lexical items, which are usually closely associated in terms of sense, phonology or collocation. Although the actual neologisms vary between the varieties, lexical conflation as a phenomenon appears to be a common feature of new Englishes generally (Mesthrie 2002*a*: 353). Mesthrie (*ibid*) cites *sincing* from "since" + "seeing" and *long-cut* by analogy with "short-cut". The ISAE corpus data revealed *starting* as a conflation of "beginning" + "start" (see 42a below) and *swearing him* by analogy with "scolding him" (see (42b.) below). A more prevalent form of the "near miss" phenomenon in the ISAE corpus involved the rendering of set phrases or idioms in ways that differed slightly from the regular form, as examples (42) c–f. illustrate.

(42) a. <#06:$06M2T:920> I did it at the *starting* of this year. (=beginning: a blend of *beginning* and *start*)

b. <#29:$42F1H:005> Kay was *swearing him* there. (=scolding/swearing *at* him)

c. <#32:$43M2T:210> You *saying* the story *wrong*. (=*telling* the story *incorrectly*)

d. <#09:$05F2H:455> We *took out* this photo. (=*had* this photo *taken*)

e. <#20:$34F1H:1455> Hey oh, I *missed to get knocked* this morning. (=was *almost run over*/knocked down)

f. <#32:$43M2T:265> …and he slipped [laughs] and fell on his back and when he *woke up* then he fell down again. (*got up*)

5.3 Features of general English

The corpus also contained features of spoken English that have been identified in other corpora such as the BNC (Leech et al. 2001) and COLT (Stenström et al. 2002). The latter is particularly significant since it represents the speech of British teenagers, closer in age to the young adult contributors to the ISAE corpus.

5.3.1 The uses of *'like'*

As the lexical item *like* featured appeared in the top ten on the frequency list (924 occurrences) it necessitated further investigation. In addition to its appearance in the corpus data as a preposition ('It was *like* a wake-up call'), an adverb ('It's not *like* she's gonna stay') and verb ('I can't say I *like* her'), *like* also appeared to be fulfilling a number of other functions in the conversational discourse. Recent studies on the class of connective devices called variously discourse markers, particles, operators and connectives, provided useful insights. The debate has extended beyond nomenclature to embrace an understanding of the function of these discourse devices as well. In studies of social dialogue, Schiffrin (1987) distinguished a complex situation where several planes of discourse operate simultaneously: on the semantic, cognitive, and interpersonal levels. According to this theory, discourse markers act not only as connectives between utterances by providing textual co-ordinates, but they also relate the different planes of discourse to each other. The general term "discourse marker" is sometimes used to refer to these other functions of a lexical item. Schourup (1999:242) defines discourse markers as "syntactically optional, non-truth-conditional connective[s]" thus expressing the view that removing a discourse marker from a clause would not make any syntactic or semantic difference.

While it may be true that syntactical and semantic properties remain unaffected by their removal from a text, discourse markers are not functionally void. As the name indicates, they "mark" or signpost the discourse in various ways (Fuller 2003): by setting off or foregrounding unusual notions (43 a.), indicating obvious approximations (43 b.) or signalling a tentative comment or judgment, so that the listener is warned not to interpret the comment literally (43 c.)

(43)

a. <#23:$29M2T:200> They're *like* so childish and they got absolutely no respect for my mother which is *like* really terrible.

b. <#11:$013M1G:135> How old is he? *Like* thirteen?

c. <#01:$03F2T:450> But you you found it *like* manageable?

However, the above classification system fails to account for a host of corpus examples that contained the pattern "be + *like*", as illustrated in (44) below:

(44)

<#01:$03F2T:940> And then my mum was *like* what did Lalitha say?

<#35:$19M2H:000> And then I'm *like* what are you studying? Pharmacy? Third year?

The above examples illustrate the so-called "quotative" use of *like*, first attested by Butters (1982) who observed that it served to demarcate an "unuttered thought" or quotation from the rest of the utterance. The development of the quotative use of be + *like*, has since been extensively studied, as it has become established as a grammaticalized construction (Romaine and Lange 1991, Ferrara and Bell 1995 and Barbieri (2005). Studies of quotative *like* have included spoken language in America, as well as comparative studies of age-graded use in the UK and USA (Buchstaller 2006). When used as a quotative, *like* is an obligatory insertion, functioning as a substitute for *to say* or *to think* (45 below). It therefore introduces direct speech or frames unarticulated thoughts. It is typically associated with casual conversation where it is positioned as a preface to a direct quote, to render it more dramatic or immediate. There were 92 instances of the quotative "be + *like*" observed in the corpus. A selection of typical occurrences is illustrated below:

(45)

<#30:$44F2T:005> First the editor calls me in and *she's like* I really have a story about Bollywood and I need someone to do it, so would you like to do it? *I'm like* oh my goodness, you know out of all the people of course she's going to choose me, cos I'm Indian.

A comparative analysis of male and female usage of the quotative *like* in the corpus revealed a strikingly gendered pattern of usage, with a five times higher incidence in the speech of female speakers (see Table 11 below). Even a cursory glance at the occurrence of quotative *like* within individual speaker turns revealed that, when relating an event, the females tended to dramatize the story much more, employing *like* as a quotative device to introduce the actual words of different speakers if there were several in the story. Compare the use of *like* by the female and male speakers in the two examples in (46) below.

(46)

1. <#19:$32F1T:515> I know. And then he comes here like a great guy and he's *like*, this meal's on me and then Molly stands up and she's *like*, no, I'll pay myself and I'm *like,* yeah, me too.
2. <#14:$15M2H:315> I was *like*, listen, if you gonna ask my parents, please by all means, you ask them.

Table 11: Comparison of use quotative *like* by male and female speakers in the ISAE corpus

Male		**Female**	
Actual occurrences	%	Actual occurrences	%
15	16,3%	77	83,7%

A search for the use of quotative *like* in COLT did not uncover significant evidence and Stenström et al. (2002: 117) have suggested that when COLT was recorded in 1993, the quotative function “had not been grammaticalized [in the UK] to the same extent that it had been in American English” at that time.

5.3.2 Taboo and swear words

The ISAE corpus was compiled from the speech of 18–29 year olds, so it also features many characteristics of the language of young people, which are not necessarily generalizable to ISAE as a whole. It became apparent during the transcription and handling of the data that the use of taboo and swear words was strongly evident and that it should be the subject of further investigation. Swearing or 'cursing' as it was formerly referred to, is as old as language itself and is an established part of everyday language use (McEnery and Xiao 2004). It is particularly evident in spoken informal language, with the incidence of occurrence generally increasing with linguistic informality. Although swear words are hard to define, native speakers of a language can recognize swearing when it occurs. Swear words or taboo words, commonly termed "bad language" are signposted with the label *offensive* in dictionaries. Hughes (1991: 3-5) explains that the reason for the taboo surrounding certain words may be linked to religion, sex, or bodily functions (usually urination and defecation), and the act of swearing violates these "sacral notions". Taboos however, are linked to broad societal values and these shift over time: for example Victorian sensibilities required the substitution of the words 'stomach' for 'belly' and 'chemise' for 'shift' (McEnery 2006: 114), while 21st century taboos tend to be centred around the need to sound 'politically correct' in matters relating to, race, disabilities or sexual orientation. Since taboos and the words referenced by them shift over time and vary between cultures, it is useful to try and distill the essential features of swear words. Andersson and Trudgill (1990) have isolated three criteria in their definition of swear words. The first of these is in line with Hughes' definition, namely that they are expressions which are usually "taboo and/or stigmatized in the culture" (1990: 53). The second feature highlights the function of swear words in the discourse in expressing "strong emotion or attitudes". Thirdly, on the semantic level, they explain that swear words do not carry a literal denotative value and "should not be interpreted literally" (*ibid*: 53). The definition is useful, because although the choice of lexical items used as swear words may differ between languages or sub-varieties of a language, it is possible to compare swear words on the basis of the taboo concepts being invoked or violated, the emotion or attitude expressed and to use an assessment of literal meaning as the litmus test of their eligibility as a swear word.

A comparison between the ISAE corpus and the conversational component of the BNC did not reveal much commonality in terms of either swear words or their frequencies, as Table 12 below shows. For purposes of comparison, the raw frequencies were converted to percentages and in the case of the ISAE corpus, *bladdy* is treated as a full equivalent of *bloody*. *Bladdy*, favoured by English-speaking South Africans (Silva et al. 1996: 69), is the acknowledged SAE pronunciation-spelling of *bloody* when used as an expletive. The variant forms *bleddy* and *blerry* which typically represent the pronunciation of Afrikaans-speaking South Africans (Silva et al. 1996: 71), did not feature in the ISAE corpus at all. The most common swear word in the ISAE corpus was *fuckin(g)*, while in the BNC it was *bloody*. It is possible that these differences might be age-related: the ISAE corpus represents the language use of a relatively narrow age band of young adults between 18 and 29, while the BNC corpus represents the language use in a broader age range from under 15 to those over 60.

Table 12: A comparison of the most frequently-occurring swearwords in the ISAE corpus and the BNC

ISAE Corpus	**Actual frequency**	**%**		**BNC**	**Frequency per million**	**%**
fuckin(g)	80	.13		bloody	771	.08
shit	62	.10		fucking	504	.05
fuck	33	.05		shit	162	.02
bladdy	20	.03		fuck	100	.01

A comparison with the COLT data collected from London teenagers (69% of whom were aged between 14–19 years), yielded strikingly different results. While neither the BNC nor the COLT age variables match the ISAE age range exactly, the COLT age group offered a more reasonable basis for comparison with the ISAE corpus. The results of the comparison are captured in Table 13 below, where it is clear that there was an exact match of in the rank order of the top four swear words in the ISAE corpus and COLT (Stenström et al. 2002: 80), together with a fairly close correlation of the percentage of frequency.

Table 13: A comparison of the most frequently-occurring swearwords in the ISAE corpus and COLT

ISAE Corpus	**Actual Frequency**	**%**		**COLT**	**Actual Frequency**	**%**
fuckin(g)	80	.13		fuckin(g)	362	.07
shit	62	.10		shit	324	.06
fuck	33	.05		fuck	256	.05
bladdy	20	.03		bloody	219	.04

The similarity in the choice of swear words and their respective frequencies is interesting, because although the two corpora represent different regional varieties of English, age is the significant common variable shared by the two sets of data. One could therefore speculate that the findings are indicative of linguistic trends in a globalized youth culture. Furthermore, although more than ten years separate the two corpora, the data suggest that the choice of swear words and the taboos they reference have not shifted significantly in that time.

5.3.3 Innovative swearing

In addition to the well-known and well-worn swear words discussed above, the corpus revealed some striking instances of semantic innovativeness, especially when the social interactions became emotionally-charged. This is confirmed by other research findings on the relationship between emotion and semantic innovativeness. It appears that as terms become more "highly-charged, so they acquire greater grammatical flexibility" (Hughes

1991: 30). For example, the word *poes* (cunt) defined as a noun (Branford 1987) was observed in the corpus data functioning as both a noun (47) and a modifier (48).

(47) <#33:$45M2H:560> Sami, that *poes* that's fucked up now
<#33:$45M2H:580> And as for the other big *poes*, what's his name? Ishan?

(48) <#33:$47M2U:015> Last night no one was out my man no one! Plus too 'Uncle Jackie' was playin' fuckin' er his *poes* beats. (Speaker criticizing the disc jockey's choice of music).
<#33:$47M2U:085> The ref was calling fuckin' *poes* decisions, ekse. (Speaker outraged at the referee's decision during an international football match)
<#33:$45M2H:310> Ja ekse, but the TV we were watching it was fuckin' 'nother one *poes* TV! (Speaker complaining about TV set).

The examples in (47) and (48) indicate an extension of the usage of *poes* from a noun to a modifier or intensifier used to express strong disapproval and anger.

5.4 Slang

If one regards swearing as occupying a position on one end of the non-standard language continuum, then slang, on the positive extreme, was also strongly evident in all the conversations. The conversations in the corpus were collected in informal settings by fieldworkers who were all young students, who could conceivably have been regarded as members of the "in-group". The style of speech was therefore relaxed and informal and the speakers showed little inhibition in their choice of conversation topics, their lexis or their style of speech. According to Andersson and Trudgill (1990) everyone immediately recognizes slang as a linguistic phenomenon when it occurs, but objective definitions of slang are hard to formulate. They define slang as "language use that is below the level of stylistically neutral language usage" (*ibid* 69). Although the definition is vague, it indicates that slang is a relative concept. However they have identified the following typical features of slang: it occurs in informal spoken situations, it is not a dialect, it is not swearing, it is innovative, it is short-lived and it is group-related. More controversially, they maintain that slang involves lexical items rather than grammar.

However the instances of slang observed in the ISAE corpus confirmed Mesthrie's observations that ISAE slang lexis co-occurs with basilectal grammar and that slang is not purely lexical (Mesthrie 1992*b*: 148). Grammatical constructs that typify basilectal ISAE are topicalization of the complement (49), preposition deletion (50), lack of *do*-support (51), lack of the perfective *have* (52), and conjunction deletion (53) (*ibid*.). The following selection of slang lexical items (*ou*, *bra*, *check*, *lakker*, *ekse* and *vaai*) in a matrix of basilectal syntax is illustrated in examples (45)–(49) below:

(49) Hey terror that *ou* was bra! (=That ou **was a terror**, bra!)

(50) They bought these couple *ous* a shot. (=They bought this couple **of** ous a shot.)

(51) You *checked* that *ou's* feet bra? (=**Did** you check that ou's feet bra?)

(52) Ja but he got a *lakker* ma *ekse*. (=Ja but he **has** a lakker ma ekse.)

(53) Apparently one two bras *vaaied*. (=Apparently one **or** two bras vaaied)

The five most common slang lexical items in the ISAE corpus are given in Table 14 below. Amongst these, *bra*, *ou* and *vaai* are also part of the slang lexicon of general SAE (Silva et al. 1996). Discussion will therefore focus on the two uniquely ISAE items *ekse* and *lakker*.

Table 14: Five most common slang words in the ISAE corpus

Word	**Actual Frequency**
bra	70
ou/-s	58
ekse	57
vaai/-s/-ing/-ed	38
lakker	22

The corpus revealed 56 occurrences of the word *ekse* in seven files. *Ekse* is distinct from the SAE *ek sê* in pronunciation, meaning and usage although both probably derive from the Afrikaans *ek sê vir jou* (‘I’m telling you’ or ‘I say’) (Silva et al. 1996: 212). The pronunciation and stress placement of the ISAE form *ekse* (ˈæksɛ) (Mesthrie 1992*a*: 110) is noticeably different to the SAE *ek sê* (ɛk ˈsɛː). All the corpus examples of *ekse* were confined to conversations between males and there were no occurrences of *ekse* in conversations between males and females. This is consonant with Mesthrie’s definition of *ekse* as a mode of address typically used between young males (Mesthrie *ibid*). Unlike the SAE *ek sê* which can be used to preface entry into a conversation, the ISAE corpus did not reveal any instances of *ekse* in an initial position in utterances. All the observed instances of *ekse* were in utterance-final positions where they appeared to have been inserted for emphasis. The following examples from the corpus illustrate how, in each case, *ekse* provides the lexical underscoring for the preceding remark:

(54)
<#11:$013M1G:005> Hey, even I wasn't that bad *ekse*!
<#33:$45M2H:330> Hey and as for the graaf[9] bra: that thing just keeps fuckin' piling up *ekse*.
<#33:$45M2H:410> Ja mm she had the body of a goddess *ekse*!
<#33:$47M2U:715> Hey fuckin' dished him *ekse* when she found out.

The corpus revealed 22 occurrences of ‘lakker’ in 7 files. ‘Lakker’ which is synonymous with the SAE ‘lekker’ is pronounced /'lʌkə/ and spelt *lakker*. It is used to signal general approval and embraces a broad range of meanings which include ‘good’, ‘great’, ‘lovely’, ‘wonderful’, ‘pleasant’, ‘delicious’, and ‘smart’. It includes the adjectival (55) and adverbial functions (56) as indicated overleaf. This corpus did not reveal any examples of this lexical feature in the speech of females, suggesting that as a slang term it is more commonly associated with the verbal repertoire of young males.

[9] graaf *n. slang* work (Mesthrie 1990)

(55) <#11:$011M1G:005> Ja but he got a *lakker* ma ekse. His ma is overboard nice, you know what I'm saying? His ma is *lakke*r; even his sisi is nice. She's also *lakker*.

(56) <#32:$43M2T:005> And then er, so when w=we were going down, this ou dressed up *lakker*, had a shower and everything.

Two slang lexical items which appeared in the corpus data and for which I have not found any evidence in the literature relating to ISAE are *for waar* (really, truly) (4 instances, and *waarheid* (the truth, truly) (17 instances). Although they only featured in one conversation, they were produced by both participants as follows:

(57)
<#33:$47M2U:005> *For waar*?
<#33:$45M2H:840> No *waarheid*, that's gonna be my new motto. I'm gonna put it on m s n.

Both *for waar* and *waarheid* derive from the Afrikaans *waar* meaning true. Their significance for ISAE lies in the fact that they have been appropriated from a "foreign" language into English, and as such they conform to the adoption pattern followed for other slang words observed in this corpus e.g. *ekse* (Afrikaans), *lakker* (Afrikaans), *vaai* (Afrikaans). According to Mesthrie this is indicative of a subconscious desire by the users of slang to challenge "traditional kinship, class and sub-ethnic links" (Mesthrie 1992: 147) and create a new sub-culture defined by its own characteristic linguistic code.

5.5 South African English lexis

As ISAE is a sub-variety of SAE, it shares much of the lexis of the latter. Reference has already been made to the use of *ja* and *hey* (see 5.1.3) which typify colloquial usage of SAE in general. A selection of SAE lexis in the corpus data with frequencies alongside is given in Table 15 below.

Table 15: A selection of SAE lexis from the corpus data

Word	**Frequency**
ja	1078
bra	70
check	63
ou/-s	58
hectic/-ally	29
eish	21
joll/-s/-ing	16
chune/-ed/-ing	14
pozi	11
bru	9
sjoe	4
ouens	3

The selection is illustrative of informal usage on the gradient from colloquial (*ja, jol, hectic, eish, bra, sjoe*) to slang (*chune, pozi, check, ou*). Amongst these *ous*/ *ouens*, *hectic(ally)* and *eish* deserve attention. In SAE, the noun *ou* is a slang general term of address meaning 'chap', 'guy' or 'fellow', and it is almost invariably applied to men (Silva et al. 1996). A typical example of usage that occurred in the ISAE corpus was:

(58)

> <#11:$013M1G:365> Ja they say it's lakker [unclear]. I wanna vaai like on a cruise ship. The *ous* reckon you get like five thousand pounds a month.

In Afrikaans from which it derives, the plural form of *ou* is *ouens*. The ISAE corpus data revealed a much higher incidence of the Englished plural *ous* (24 instances) as opposed to only 3 examples of *ouens* appearing in the data.

The neologism *hectic(ally)* is a recent SAE general-purpose intensifier meaning anything from ‘over the top’ to ‘good’ to ‘really bad or difficult'. Typical of all varieties of colloquial SAE youthspeak, its appearance in the ISAE corpus is noted as an indication of ISAE observing the speech trends of the wider population of young people in the country. A slightly different case is revealed by ISAE speakers adopting the typically BSAE or neologism *eish* (sometimes represented as *eesh/aish*), an exclamation which conveys a range of emotions from surprise (wow!) to pain (ouch!). de Klerk (2006: 66) observed that the use of the exclamation *eesh/aish* imparted a particularly Xhosa flavour to the discourse data in her corpus. Its appearance in the ISAE corpus suggests a movement away from the linguistic ghettoisation of the past and the desire to reach for a new South African identity – one that identifies with the majority of the country’s population.

5.6 Concluding remarks

The ISAE corpus, although a modest 60 000 words in size and collected from a small demographic band within a limited context, nevertheless reveals complex and striking patterns about language usage within this group. In the ultra-informal speech situations which constituted the contexts for data collection, all the corpus contributors employed the distinctive ISAE syntax or lexis in varying degrees. This is all the more remarkable in the light of the fact that the contributors were all university students who would be required to employ formal standard English for successful functioning in academic contexts. Layered over the ISAE there was evidence of SAE lexis combined with examples of “global youthspeak”. This layered linguistic repertoire suggests that the speakers not only have access to multiple local and global varieties of English, but that they are able to use them with great facility in response to contextual demands.

Chapter 6: Conclusion

6.0 Overview

Whilst corpora offer vast bodies of data for linguistic analysis, the actual corpora themselves do not offer any explanations for the nature of linguistic phenomena. They merely provide the empirical evidence for further study. It is clear therefore that corpus studies cannot stand alone: they must be interpreted through the lens of interpretive frameworks such as intuition-based ones, in order to yield useful results. The data in the corpus collected as part of this thesis was accordingly investigated with sociolinguistic tools used to describe language varieties and discourse analysis.

6.1 Review and summary of the findings

As mentioned at the outset, the essential catalyst for this thesis was the lack of a corpus of SAE for linguistic and lexicographic investigation. The urgency of the need was highlighted by the increasing availability of corpus resources for most major national languages, for several minority languages and in particular, for most varieties of World English. A study of prior research into SAE revealed the multi-faceted nature of this variety and indicated that any efforts to compile a corpus of SAE would need to reflect the diversity inherent in this variety of World English. The evolution of ethnically- and language-based varieties of SAE is an historical fact and this in turn influenced the identification and naming of its sub-varieties. Guided by the existing taxonomy used to describe sub-varieties of SAE, it was felt that an initiative involving the collection of the sub-variety known as ISAE would be a useful small step towards the much bigger goal of building a corpus of SAE. To this end the construction of a 60 000-word spoken corpus of ISAE was undertaken and successfully completed. It is hoped that the corpus that is offered here in its untagged form will be of use not only to corpus linguists, but also to fellow-scholars from linguistics and linguistic-related disciplines. The corpus data represent a wealth of lexical, discourse, grammatical and semantic information that is available and ready for mining and evaluating. This thesis was but the first step “towards” a fuller corpus of ISAE. A full ISAE corpus could then serve as a standard against which future linguistic explorations into this sub-variety could be referenced. In addition to its value as a reference standard for ISAE, a full ISAE corpus, together with parallel corpora for other sub-varieties of SAE, could assist in providing a more nuanced understanding of SAE. The linguistic data in these parallel corpora could be a rich source of information that could facilitate cross-comparisons between the sub-varieties in order to enhance our understanding and definition of SAE. The demonstrable value of corpora of sub-varieties is evidenced by de Klerk's 500 000-word corpus of Xhosa English, which has furnished data and analyses on a variety of lexical and syntactic features (de Klerk 2006). The challenge to other researchers is to compare and contrast de Klerk's findings relating to Xhosa English with other sub-varieties of BSAE and with other sub-varieties of SAE.

Using the ISAE corpus that was compiled as part of this thesis, it was possible to identify and evaluate the robust lexical and syntactic features of ISAE against the background of prior research into this sub-variety of SAE. The corpus confirmed the existence of three noticeable features across the data *viz*. the use of '*y'all*' as a second person plural pronoun (see 5.2.1), the copula attraction to '*wh-*' in indirect questions (see 5.2.2) and the use of '*of*' in partitive genitive constructions (see 5.2.3). The fact that all the contributors to the corpus had received 12 years of formal English-medium schooling and that for 98% of them English was the preferred language of communication, is evidence of the resilience of these constructions. Other significant ISAE features prevalent in the corpus were the use of topicalisation (see 5.2.4), the use of 'but' and 'like' in sentence-final positions (see 5.2.5), and the use of the singular pronoun 'it' to replace plural noun phrases (see 5.2.9).

Although the corpus did not include phonetic mark-up, the use of *lakker* /'lʌkə/ as a pronunciation variant of *lekker*, and *ekse* /'æksɛ/ as a variant of *ek sê* /ɛk 'sɛː/ were noted as persistent pronunciation forms in the speech of the male contributors. Their pragmatic function was also discussed. These, together with other lexical examples of slang usage (*vaai* and *ou*), (all appropriations from Afrikaans and notably the preserve of the male speakers in the corpus), were highlighted as evidence of "outshifting" or the desire to challenge traditional ethnic or cultural boundaries. In this regard two other neologisms from Afrikaans *for waar* and *waarheid* were also identified in the speech of males.

In addition to the above features specific to ISAE, the corpus also revealed a variety of well-established SAE lexical items which included verbs, nouns, modifiers and exclamations. Amongst these were the colloquially pervasive *ja* (see 5.1.3), *bladdy* (meaning *bloody*) as rendered by most English-speaking South Africans (see 5.3.2), *jol(ling)* for partying or enjoying oneself, *hectic(ally)* as an intensifier for anything 'over the top', and the exclamations *sjoe* and *eish.* Amongst these, the appearance of *eish* and *bra*, originally common amongst speakers of BSAE, is a note-worthy indication of the cross-fertilization from one sub-variety to another and an indication of the blurring of ethnically-related linguistic boundaries. The topic choices in the ISAE corpus were contrasted with the corpus of Xhosa-English (the only other corpus of a sub-variety of SAE) and some differences were noted and explained (see 5.1.2).

Data for this building block of the corpus of ISAE were also usefully compared with data from the BNC and COLT. Although much smaller and collected from a narrower demographic band than the BNC, the data in the ISAE corpus confirmed Zipf's principle of least effort, as the top 63 most frequent words were one-syllable words. Thereafter, the syllabic complexity of the words exhibited a ratio that was consonant with spontaneous, informal speech. A comparison of the top 100 words in COLT and the ISAE corpus revealed interesting differences that were not all related to the cultural or geographic differences between these two varieties of World English. The most significant of these was the much higher frequency of *like* in the ISAE corpus (see 5.4.1). The use of *like* in this corpus was found to be in keeping with global trends for the term, which has since been acknowledged as a lexicalized discourse marker. In this respect the speakers in the ISAE corpus could be said to be keeping pace with their international cohorts in the English-speaking world.

6.1.1 The corpus design

In selecting corpus contributors great care was taken to ensure proportional representation from the substrate Indian language groups, and the 1960 census was used as a point of reference (see Table 4 in section 4.2.2). However, just over forty years on from that date, there has been a discernible language shift to English amongst Indian South Africans, and it is perhaps debatable whether such scrupulously-observed proportional representation was necessary. Nonetheless, guided by earlier research into ISAE (Mesthrie 1992*a*, 1992*b*) the careful selection of appropriate numbers from the ancestral language groups was undertaken as a precautionary measure in order to avoid making *a priori* judgments, as it could have offered a basis for comparison, despite the fact that the research sample was relatively small. It was only after studying the data gathered on the language practices of the ISAE corpus contributors, that I could legitimately confirm the strong thrust towards English: 96% of the participants indicating that English was the first language they had learnt to speak at home and 98% indicating that English was the preferred language of communication in the home. Language shift on this scale, combined with common schooling, shared socialization experiences and intermarriage (between substrate language groups) has tended to diminish all but the most robust of the formerly observed syntactic and lexically-marked differences between the language-affiliated groupings in this survey. In fact, on the basis of the information supplied by the respondents, it would be feasible to argue that the ancestral language affiliations now appear to exist in notional or symbolic form only. The value of meticulously observing that all substrate language groups are proportionately represented in a corpus might be an essential consideration only where there is still active use of those languages. The linguistic dynamics of this speech community suggest that frequent monitoring of language use together with attitudes to ancestral languages and cultural practices is necessary for future research.

6.1.2 Reviewing the method of data collection

Economic and pragmatic considerations led to the choice of small hand-held analogue tape recorders for recording the speech of the contributors to the corpus. However (as noted in 4.3.2) this did not always yield high-quality recordings and it must be conceded with hindsight that digital tape-recorders would have been preferable. They would also have simplified the transcription process, since sound files which are loaded directly onto the computer are more easily managed and manipulated.

A more serious factor relating to the method of data gathering was the obvious presence of the tape recorder. Although every effort was made to use technology that was unobtrusive in order to secure naturalistic data, it would be unrealistic to claim that the participants totally forgot about the tape recorder. The following extracts reveal the speakers’ awareness of the tape recorder:

<#12:$20F1M:185> You think it's taping properly?

<#35:$49M1E:160> You got that on tape isn't it?

However, on the whole, the conversations were usually uninhibited and the use of field workers who were perceived and accepted as members of the ‘in-group’ was crucial in securing relatively spontaneous data. A technological solution to the ever-present and visible tape recorders would be to use lapel microphones and small machines which the participants can slip into their pockets. The lapel microphones would deliver better quality sound recordings and the pocket-sized machines would make the tape recorders less obvious.

6.2 Limitations of this study

6.2.1 Social variables

The data for this corpus derived from a population of young adults (between 18 and 29 years of age) who had a minimum educational level of Grade 12. As this is a very narrow age band, the contributors to this corpus cannot be said to represent the demographics of the full targeted speech community (i.e. all speakers of ISAE), in terms of either age or educational level. In a fully representative corpus, the population sample should be broadened to include a greater age range and educational background. In addition, other social variables such as occupational background and religion could be encompassed as well. In a fully comprehensive study, the geographical distribution (which was restricted to Kwa-Zulu Natal) could also be extended to include other areas such as Gauteng and the Western Cape, which have smaller but noticeable populations of Indian South Africans. Extending the geographical range in this manner could furnish useful comparative data about the influences of geographical dispersion and contact with other speech communities on this variety.

6.2.2 Conversational variables

Using spontaneous informal dialogues as "the lowest common denominator" in the hierarchy of speech genres meant that the data collected for the ISAE corpus comprised only informal direct private conversations between two participants. An extension of the conversational data should ideally include multi-party conversations as well. Such conversations are more challenging to transcribe, but they would furnish data from the sort of conversation configuration that occurs every day in normal life. In this regard, multi-party conversations should also embrace family settings, as these would yield linguistic data from private contexts, but it should also include a mixture of age-groups and educational backgrounds. Such complexity was not possible in a limited study such as mine.

6.2.3 Telephone conversations

Data for this corpus derived from private face-to-face conversations. An extension of the genre of private dialogues should ideally also include telephone conversations, as the use of the telephone, especially the cell phone, is a common feature of communication amongst all communities in South Africa, particularly young people. Telephonic conversations have the potential to provide examples of private communication without the obvious intrusion of the tape recorder and/or the researcher. In addition, telephonic conversations add the dimension of *distance* as a variable in private communication. The two largest problems associated with recording telephonic conversation are legislation prohibiting the violation of privacy and the logistics of tracking down participants in such conversations. The first issue, *viz.* the violation of privacy, is an ethical problem surrounding the surreptitious monitoring of any private conversation, not just telephone conversations. Since the tapping of telephones is not recommended, an alternative method of securing the required speech data would be to use pre-recorded messages to alert callers that calls are being recorded. This practice is successfully used by financial institutions which typically warn callers that calls are being recorded 'for quality-control purposes'. A modified message could indicate that the call is being recorded for "The [Name] Corpus Linguistic Research Project at the [Name] Research Institute".

The second problem is related to the need to capture the essential demographic information (such as location, age, gender, level of education and linguistic background) of each speaker. Once again, the practice followed in the business sector is useful, where callers could provide the relevant information by keying in a digit on the telephone pad. Data collected in this way could be screened for eligibility and then sorted according to the required data categories.

I have described rather painstakingly and in some detail a possible method to collect telephonic data because I think that this is a potentially rich source of conversation data that is all too easily dismissed as "problematic" because it does require prior structuring to secure as well as considerable financial and technological resources to execute. However there are examples of corpora of telephonic data, the most recent being the Fisher Levantine Arabic Telephone Conversational Speech which forms a subset of other conversational data from this speech community.
(See http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0703&L=cllt&P=217)

6.2.4 Collecting data from formal contexts

While it has been observed that ISAE is used most markedly for "in-group casual communication", the contextual range of spoken genres could be extended to include some formal exchanges such as speeches, community meetings, classroom interactions and business transactions. Increasing the scope in this way could provide useful material for comparison with material gathered in strictly private settings where the discussants are known to each other.

6.2.5 Length of extracts

The length of each speech sample in the ISAE corpus was approximately 2 000 words, in line with practices followed ICE, Brown and LOB corpora. This entailed excising sections of dialogue, using the principle of selecting material after ten minutes into the conversations. While this practice ensured consistency across samples, it did omit the beginnings and endings of dialogues. A method for procuring more naturalistic samples might be to allow the type of discourse to determine the length of the speech sample along lines suggested by Holmes (1996: 164). Admittedly this will result in speech samples of varying length, but this would have to be weighed against the value of making realistic accommodations for discourse types such as telephone conversations, business transactions and community meetings.

6.2.6 Transcription and mark-up

This study sought the middle ground with regard to transcription and used simple orthographic representation of speech. While this proved to be adequate for lexical and grammatical analysis, the exclusion of phonetic and prosodic information limits its application and value. In converting speech to writing, one ultimately runs the risk of presenting spoken language as “a set of written texts which happen to have originated in spoken form” (Knowles et al. 1996:2), and it must be acknowledged that therefore the ideal transcription should include prosodic and phonetic segmentation, in order to facilitate a range of linguistic explorations.

6.3 Suggestions for further research

Despite significant technological advances in speech and language recognition, written corpora continue to outnumber spoken ones. Even large mixed corpora which contain both spoken and written data (e.g. the BNC and ICE) tend to be weighted in favour of written data. The fundamental reason for this, as discussed and illustrated in this thesis, is that the spoken language, and in particular spontaneous speech, presents a daunting set of challenges to corpus compilers and transcribers. Speech is typically ephemeral, unless it is preserved in a recording. Even then, it needs further processing from an audible medium into a visible form (writing or sound waves) for study and analysis.

6.3.1 Transcription conventions

In terms of transcription there is no absolute agreed standard for the semantic representation of spoken language, especially of spontaneous speech. There are different methods of indicating features such as latched utterances, pauses and non-verbal sounds, as well as vast differences in the levels of transcription in each corpus. Some corpora include semantic breaks, prosodic mark-up and phonetic mark-up in acknowledgement of their roles in the constitution of a speech act. While it is true that the ultimate purpose of the corpus determines the level of transcription detail, and the absence of rigid standard conventions allows for flexibility, it is equally true that the absence of standardized transcription conventions for spontaneous speech constitutes a challenge for cross-corpora comparisons within the genre. In addition, consideration needs to be given to the development of a system for tracking and detecting topic threads in spoken data. This will assist in giving a fuller representation of the semantic content of spontaneous speech.

6.3.2 Email correspondence

Any corpus compiled for linguistic investigation cannot afford to ignore email correspondence, as it represents a type of discourse that is located somewhere on the continuum between writing and speech. In terms of level of formality, email could be said to be closer to informal speech than it is to writing. One immediate advantage of email correspondence from the corpus compilers' point of view is that the data are already in electronic format. In terms of ISAE in particular, a corpus of email correspondence could be investigated for evidence of movement between the lects (acrolect, mesolect and basilect). Principled policy decisions would need to be made about the inclusion of forwards and replies, since these could significantly affect word counts. Other considerations might extend to which attachments would qualify as evidence of the genre under investigation. A well-known example of an email corpus is the Enron Corpus, made available to the public in 2003, following the scandal and subsequent collapse of energy and utilities company Enron in the USA. Although not compiled to service the aims of linguistic research, it nevertheless represents a vast data set, contributed by over 150 correspondents and has subsequently been referenced for linguistic and computer-based research (Bekkerman et al. 2004). Other linguistic research into corpora of email has involved the study of 'junk emails' (Orasan and Krishnamurthy 2002) which has yielded information on the frequency of specific lemmata in this genre.

6.3.3 Written data

Although it has been argued that ISAE exists chiefly in the spoken domain, a complete corpus of this sub-variety should include written material as well. On the basis of data gathered from the written mode it should be possible to evaluate to what extent features of ISAE are used in writing. This could usefully inform the design of educational programmes aimed at assisting users of ISAE discern the nuances between ISAE and other varieties of SAE and with World English in general.

6.3.4 Speakers’ attitudes towards their own usage

A corollary of research which highlights the distance between a linguistic sub-variety and the standard variety, is that the research ought to include data gathered from user-perspectives of the sub-variety as well. In other words it should reflect the speakers’ attitudes towards their own usage. In the case of ISAE, there is certainly a need for a survey of speakers’ attitudes towards ISAE. In the broader South African context attitudinal surveys should be undertaken for all sub-varieties of SAE and the results could be collected and compared. If data from such attitudinal surveys were interpreted together with shifts in the loci of political and economic power in the country, it is possible that there might be a need for a re-appraisal of the existing standardized norms of pronunciation, grammar and idiom of SAE.

6.3.5 Communities of practice

Linguistic data from corpora such as the ISAE need not stand alone and be interpreted independently of other social variables. Wisdom from related disciplines, notably sociology, suggests that the ‘speech community’ takes a rather narrow view of a set of phenomena that occur in a broader social matrix. An approach based in social theory has given rise to the term “communities of practice” (Eckert and McConnell-Ginet 1999, Holmes and Meyerhoff 1999) to describe the interplay of linguistic and other factors that collaborate to constitute the identity of the community and of the individual within the community. Traditionally communities of practice have been conceived of as sub-sets of speech communities, but their abiding value is the manner in which they link linguistic variables to the whole range of social practices. In the case of Indian South Africans, although there has been a discernible shift away from the substrate Indian languages towards English, the sub-variety of ISAE survives in a robust form that at times almost belies the lack of fluency in any Indian languages. It is possible that an ethnographically-based approach which uncovers identity constructs in this community of practice might offer deeper insights and ultimately a more holistic way of understanding the linguistic data in the ISAE corpus.

6.3.6 Examining "the core" and "periphery" of SAE

In the case of SAE, what is needed is a system that will acknowledge the permeability of any boundaries constructed 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 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 what constitutes exactly 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.

6.3.7 PanSALB or business sector involvement

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. 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. Since the act of transcribing in orthographic form what originates in speech discards much essential information, it is vital that future spoken language corpora ought to be available in at least two forms: audio and written. 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 labeled. Developing such a corpus is a formidable task and Williams (1996:19) concludes with the sober reflection that "This kind of corpus would be a major undertaking". Reference has already been made to the challenges and limitations of a sole researcher who undertakes to collect, transcribe and mark up a corpus. 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 publishing house HarperCollins. In the South African context, this 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.

6.4 Final remarks

This study required the collection of spoken data, the transcription of the data and the establishment of a corpus of 60 000 words. This completion of this step laid the foundation for a fuller corpus of ISAE that can be developed by researchers with an interest in corpus linguistics in general and in the sub-variety in particular. The subsequent identification of general trends and specific features of ISAE, SAE and general English in this corpus provided a sampling of the vitality of ISAE. For this reason there can be no last word on a corpus of a living language or language variety. It is therefore hoped that the corpus that accompanies this thesis will provide fertile ground for further explorations into themes that link ISAE with SAE and with world Englishes globally.

References

Adendorff, R. (2002) "Fanakalo: a pidgin in South Africa." In Mesthrie, R. ed. Language in South Africa. Cambridge: Cambridge University Press, 179-198.

Aijmer, K. and Altenberg, B. eds. (1991). English Corpus Linguistics : studies in honour of Jan Svartvik. London: Longman

Aijmer, K. (1996). Conversational Routines in English. London: Longman

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

Andersson, L. and Trudgill, P. (1990). Bad Language. England: Penguin

Barbieri, F. (2005). "Quotative Use in American English". Journal of English Linguistics, 33(3), 222-256.

Barnbrook, G. (1996). Language and Computers: a practical introduction to the computer analysis of language. Edinburgh: Edinburgh University Press

Baroni, M. and Bernadini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. Proceedings of LREC 2004, Lisbon: ELDA, 1313-1316.

Bekkerman, R., McCallum, A., and Huang, G. (2004). Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora. http://www.cs.umass.edu/~ronb/papers/email.pdf [Accessed 27.06.07].

Bhana, S. and Pachai, B. (1984). A documentary history of Indian South Africans, 1862-1982. Stanford, California: Hoover Institution Press

Bhatt, R.M. (2001). "World Englishes." Annual Review of Anthropology 30, 527-50.

Biber, D. and Finegan, E. (1991). "On the Exploitation of Computerized Corpora in Variation Studies". In Aijmer, K. and Altenberg, B. eds. English Corpus Linguistics : studies in honour of Jan Svartvik. London: Longman, 203-220.

Biber, D., Conrad, S. and Reppen, R. (1998). Corpus Linguistics : investigating language structure and use. New York: Cambridge University Press

Blachman, E., Meyer, C., and Morris, R. (1996). "The UMB and ICE Markup Assistant". In Greenbaum, S. ed. Comparing English Worldwide. Oxford: Clarendon Press, 54-64.

Bolton, K. and Kachru, B. eds. (2006). World Englishes: critical concepts in linguistics. Vol. 2. London: Routledge

Boshoff, P. (2005). "Diasporic Consciousness and Bollywood: South African Indian youth and the meanings they make of Indian film". Unpubl. M.A. Thesis. Rhodes University

Bourdieu, P. (1978). Outline of a theory of practice. Cambridge & New York: Cambridge University Press

Brain, J. B. (1983). Christian Indians in Natal: An historical and statistical study. Cape Town: Oxford University Press

Branford, J. (1987). A Dictionary of South African English. Cape Town: Oxford University Press

Bruthiaux, P. (2003). "Squaring the Circles: issues in modelling English worldwide". International Journal of Applied Lingistics, 13(2), 159-78.

Buchstaller, I. (2006). "Diagnostics of age-graded linguistic behaviour: The case of the quotative system". Journal of Sociolinguistics, 10(1), 3-30.

Bughwan, D. (1970). "An investigation into the use of English by the Indians in South Africa with special reference to Natal". Unpubl. Ph.D. thesis. Univ. of South Africa

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

Butters, R.R. (1982). "Editor's note [on be like 'think']". American Speech, 57, 149.

Census 1996 http://www.statssa.gov.za/census01/Census96/HTML/CIB/Population/210.htm [Accessed 11.12.07]

Census 2001 http://www.statssa.gov.za/census2001/digiAtlas/index.html [Accessed 2.09.06]

Chafe, W. Du Bois, J. and Thompson, S. (1991). "Towards a new corpus of spoken American English". In Aijmer, K. and Altenberg, B. eds. English Corpus Linguistics : studies in honour of Jan Svartvik. London: Longman, 64-82.

Chafe, W. (1995). "Adequacy, user-friendliness, and practicality in transcribing". In Leech, G., Myers G. and Thomas J. eds. Spoken English on Computer New York: Longman, 54-61.

Cheng W., and Warren, M. (1999). “Facilitating a description of intercultural conversations: the Hong Kong Corpus of Conversational English”. ICAME Journal 23, 5-20.

Chomsky (1957). Syntactic Structures. The Hague: Mouton

Chomsky (1962) Paper given at Third Texas Conference on Problems of Linguistic Analysis in English, 1958. Austin: University of Texas

Coetzee-Van Rooy, S. and Van Rooy, A. (2005). “South African English: labels, comprehensibility and status”. World Englishes: Journal of English as an International and Intranational Language 24 (1), 1-19.

COLT Word Frequency List: http://torvald.aksis.uib.no/colt/COLT1000.TXT [Accessed 29.10.2006]

Cook, G. (1995). “Theoretical Issues: transcribing the untranscribable”. In Leech, G., Myers G. and Thomas J. Spoken English on Computer New York: Longman, 35-53.

Crossley, S. (1987). “The syntactic features of South African Indian English among students in Natal, with regard to use and attitudes towards usage.” Unpubl. MA thesis. University of Durban-Westville

Crowdy, S. (1993). “Spoken Corpus Design”. Literary and Linguistic Computing, 8 (4), 259-265.

Crowdy, S. (1995). “The BNC Spoken Corpus”. In Leech, G., Myers G. and Thomas J. Spoken English on Computer. New York: Longman 224-234.

Crystal, D. (1995). The Cambridge Encyclopedia of the English Language. UK: Cambridge University Press

Crystal, D. (2004). The Stories of English. London: Allen Lane

Cruden, A., Irwin, C.H., Adams, A.D. and Waters, S.A. (1946) Cruden's Complete Concordance to the Old and New Testaments: with notes and biblical proper names under one alphabetical arrangement. London: Lutterworth Press

de Klerk, V. (2002*a*). “Starting with Xhosa English...towards a spoken corpus”. International Journal of Corpus Linguistics, 7(1), 21-42.

de Klerk, V. (2002*b*). “Towards a corpus of Black South African English”. Southern African Lingustics 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, 22 (4), 463-481.

de Klerk, V. (2006). Corpus Linguistics and World Englishes: an analysis of Xhosa English. S.l. : Continuum

De Schryver. G-M. (2002). “Web for/as Corpus: A Perspective for the African Languages”. Nordic Journal of African Studies, 11(2), 266-282.

Eckert, P. & McConnell-Ginet, S. (1999). “New generalizations and explanations in language and gender research”. Language in Society, 28, 185-201.

Edwards, J. (1995). “Principles and alternative systems in the transcription, coding and mark-up of spoken discourse”. In Leech, G., Myers, G. and Thomas, J. Spoken English on Computer: transcription, mark-up and application. Harlow, Essex: Longman, 19-34.

Ehlich, K. (1993). “HIAT: A Transcription System for Discourse Data”. In J. A. Edwards and M. D. Lampert eds. Talking Data: Transcription and Coding in Discourse Research. Hillsdale, NJ: Erlbaum, 123-148.

Ferrara, K. and Bell, B. (1995). “Sociolinguistic Variation and Discourse Function of Constructed Dialogue Introducers: The Case of Be + like”. In American Speech 70(3), 265-290.

Fraser, B. (1999). “What are discourse markers?” Journal of Pragmatics, 31, 931-952.

Fuller J. (2003). “Use of the discourse marker *like* in interviews”. Journal of Sociolinguistics, 7(3), 365-377.

Gast, V. (2006). “Introduction”. In Zeitschrift aus Anglistik und Amerikanistik (ZAA) - Special issue on The Scope and Limits of Corpus Linguistics – Empiricism in the Description and Analysis of English, 54(2), 113-120.

Gough, D. (1996). “Black English in South Africa”. In de Klerk, V. ed. Focus on South Africa. Amsterdam: John Benjamins, 53-77.

Green, J. (1997) Chasing the Sun: dictionary-makers and the dictionaries they made. London: Pimlico

Greenbaum, S. (1985). ed. The English Language Today. UK: Pergamon Press

Greenbaum, S. (1996). ed. Comparing English worldwide : the International Corpus of English. Oxford: Clarendon Press

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

Gumperz, J.J. (1970). “Sociolinguistics and Communication in Small Groups”. In Pride, J.B. and Holmes, J. eds. (1972). Sociolinguistics. Great Britain: Penguin, 203-224.

Holmes, J. (1996). “The New Zealand Spoken Component of ICE: Some Methodological Challenges”. In Greenbaum, S. ed. Comparing English Worldwide. Oxford: Clarendon Press, 163-181.

Holmes, J. and Meyerhoff, M. (1999). “The Community of Practice: Theories and methodologies in language and gender research”. Language in Society, 28, 173-183.

Hughes, A. and Trudgill, P. (1987) English accents and dialects : an introduction to social and regional varieties of British English. London: Edward Arnold

Hughes, G. (1991). Swearing: A Social History of Foul Language, Oaths and Profanity in English. Oxford: Blackwell

Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press

Johansson, S. (1995).”The approach of the Text Encoding Initiative to the encoding of spoken discourse”. In Leech, G., Myers G. and Thomas J. Spoken English on Computer New York: Longman, 82-98.

Kachru, B. (1985). “Standards, codification and sociolinguistic realism: the English language in the outer circle”. In Quirk, R. and Widdowson, H. G. eds. English in the World: Teaching of Learning of Language and Literature. Cambridge: Cambridge University Press, 11-16.

Kachru, B. (1994). “English in South Asia”. In Bolton, K. and Kachru, B . (2006) World Englishes: critical concepts in linguistics. UK: Routledge, Vol. 2, 255-310.

Kachru, B. (1997). “English as an Asian Language”. In Bolton, K. and Kachru, B. (2006) World Englishes: critical concepts in linguistics. UK: Routledge, Vol. 2, 324-342.

Kading, J. (1897). Häufigkeitswörterbuch der deutschen Sprache. Steglitz: privately published.

Kennedy, G. (1998). An Introduction to Corpus Linguistics. London: Longman

Kilgariff, A. and Grefenstette, G. (2003). “The Web as Corpus”. Computational Linguistics, V(N), 1-15.

Kilgariff, A., Richly, P., Smrz, P. and Tugwell, D. (2004). The Sketch Engine. Proceedings of EURALEX Conference, Lorient, France, 105-116.

Knowles, G., Wichmann, A. and Alderson, P. eds. (1996). Working with Speech: perspectives on research into the Lancaster/IBM Spoken English Corpus. London and New York: Longman.

Kolhapur Corpus: http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM#ABI [Accessed 10.12.06]

Krishnamurthy, R. (2002). "The Corpus Revolution in EFL Dictionaries." In Kernerman Dictionary News, 10 July. http://kdictionaries.com/newsletter/kdn10-9.html [Accessed 03.11.07].

Labov, W. (1972). Sociolinguistic Patterns. Philadelphia: University of Pennsylvania Press

Labov, W. (c1972) Language in the inner city : studies in the Black English vernacular Philadelphia : University of Pennsylvania Press

Landau, S. (2001). Dictionaries: the art and craft of lexicography. Cambridge: Cambridge University Press

Lass, R. (2002). "South African English". In Mesthrie, R. ed. Language in South Africa. Cambridge: Cambridge University Press, 104-126.

Lave, J. and Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge & New York: Cambridge University Press

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. (1991). "The State of the Art in Corpus Linguistics." In Aijmer, K. and Altenberg, B. English Corpus Linguistics: Studies in Honour Jan Svartik. London: Longman, 8-29.

Leech, G., Myers, G. and Thomas, J. (1995). Spoken English on Computer: transcription, mark-up and application. Harlow, Essex: Longman

Leech, G., Rayson, P., and Wilson, A. (2001). Word Frequencies in Written and Spoken English: based on the British National Corpus. Great Britain: Pearson

Malan, K. (1996). "Cape Flats English". In V. de Klerk. ed. Focus on South Africa. Amsterdam/Philadelphia: John Benjamins, 125-148.

McEnery, T. and Wilson, A. (1996). Corpus linguistics. Edinburgh: Edinburgh University Press

McEnery T. and Xiao Z. (2004). “Swearing in modern British English: the case of fuck in the BNC”. Language and Literature. 13 (3), 235-268.

McEnery, T. (2006). Swearing in English : bad language, purity and power from 1586 to the present. London: Routledge

Mesthrie, R. (1991) Language in indenture : a sociolinguistic history of Bhojpuri-Hindi in South Africa. Johannesburg : Witwatersrand University Press

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

Mesthrie, R. (1992*b*). English in Language Shift. South Africa: Witwatersrand University Press

Mesthrie, R. (1996). “Language Contact, Transmission, Shift: South African Indian English”. In de Klerk, V. ed. Focus on South Africa. Amsterdam: John Benjamins, 79-98.

Mesthrie, R. (2002*a*). “From Second Language to First Language: Indian South African English”. In Mesthrie, R. ed. Language in South Africa. UK: Cambridge University Press, 339-355.

Mesthrie, R. (2002*b*). “Language Change, Survival and Decline: Indian languages in South Africa”. In Mesthrie, R. ed. Language in South Africa. UK: Cambridge University Press, 161-176.

Mesthrie, R. (2003). “Children in language shift – the syntax of fifth generation, pre-school Indian South African English speakers”. Southern African Linguistics and Applied Language Studies, 21(3), 119-126.

Meyer, C.F. (2002). English Corpus Linguistics : an introduction. New York: Cambridge University Press

Miller, J. and Weinart, R. (1995). “The function of LIKE in dialogue”. Journal of Pragmatics, 23, 365-393.

Milroy, L. (1987). Observing and Analyzing Natural Language: a critical account of sociolinguistic method. Oxford: Blackwell

Moon, R. (1987). “The Analysis of Meaning”. In Sinclair, J. ed. Looking Up: an account of the COBUILD Project of lexical computing. London: Harper Collins, 86-103.

Nelson, G. (1996). “The Design of the Corpus”. In Greenbaum, S. ed. Comparing English Worldwide. Oxford: Clarendon Press, 27-53.

Nelson, G. (2006). “The core and periphery of world Englishes: a corpus-based exploration”. World Englishes, 25(1), 115-129.

Nihalani, P., Tongue, R.K. and Hosali, P. (1989) Indian and British English: a handbook of usage and pronunciation. Delhi: Oxford University Press

Orasan, C. and Krishnamurthy, R. (2002). A corpus-based investigation of junk emails. https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2002/LREC/pdf/113.pdf [Accessed 13.07.07].

Overstreet, M. (1999). Whales, candlelight and stuff like that: general extenders in English discourse. Oxford and New York: Oxford University Press

Oxford English Dictionary http://dictionary.oed.com/

Partington, A. (1998). Patterns and meanings: using corpora for English language research and teaching. Amsterdam: John Benjamins

Pickering, B., Williams, B. and Knowles, G. (1996). “Analysis of transcriber differences in the SEC.” In Knowles, G., Wichmann, A. and Alderson, P. eds. Working with Speech: Perspectives on research into the Lancaster/IBM Spoken English Corpus. London: Longman, 61-86.

Platt, C.L. (2004). “A corpus-based investigation of Xhosa English in the classroom setting”. Unpubl. M.A. thesis Rhodes University

Platt, J., Weber, H. and Ho, M.L. (1984). The New Englishes. England: Routledge & Kegan Paul

Reppen, R., Fitzmaurice, S.M. and Biber, D. (2002). Using Corpora to Explore Linguistic Variation. Amsterdam; Philadelphia: John Benjamins

Romaine, S. (1982). “What is a speech community?”. In Romaine, S. ed. Sociolinguistic variation in speech communities. London: Arnold, 13-24.

Romaine, S. and Lange, D. (1991). “The Use of Like as a Marker of Reported Speech and Thought: a Case of Grammaticalization in Progress”. American Speech, 66(3), 227-279.

Rundell, M. and Stock, P. (1992). “The Corpus Revolution” English Today 8(3) 21-32.

Sampson, G. and McCarthy, D. eds. (2004). Corpus linguistics: readings in a widening discipline. London and New York: Continuum

Schiffrin, D. (1987). Discourse Markers. Cambridge: Cambridge University Press

Schmied, J. (1996). “Second-Language Corpora”. In Greenbaum, S. ed. Comparing English Worldwide. Oxford: Clarendon Press, 182-196.

Schneider, E. (2003). “The Dynamics of New Englishes: from identity construction to dialect birth”. Language, 79(2), 233-273.

Schourup, L. (1999). “Discourse Markers”. Lingua, 107, 227-265.

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

Sinclair, J. M. (1987). Looking Up: an account of the COBUILD Project in lexical computing. London: HarperCollins.

Sinclair, J. M. (1995). “From theory to practice”. In Leech, G., Myers G. and Thomas J. eds. Spoken English on Computer New York: Longman, 99-109.

Sinclair, J. M. (2004*a*). “Corpus and Text: Basic Principles”. In M. Wynne, ed. Developing Linguistic Corpora: a guide to good practice. Oxford: Oxbow Books: 1-16.
http://ahds.ac.uk/linguistic-corpora/ [Accessed 24.08.06].

Sinclair, J. M. (2004*b*). Trust the Text: language, corpus and discourse. London: Routledge

Stefanowitsch, A. (2005). “New York, Dayton (Ohio), and the Raw Frequency Fallacy”. Corpus Linguistics and Linguistic Theory, 1 (2), 295–301.

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

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

Svartik, J. and Quirk, R. eds. (1980). A Corpus of English Conversation. Lund: C.W.K. Gleerup

Swales, J. M. (1990). Genre Analysis: English in academic and research settings. Cambridge : Cambridge University Press

Tagliamonte, S. (2000). “English . . . and them! Form and Function in Comparative Perspective”. American Speech 75(4), 405-409.

Taylor, L. (1996). “The compilation of the Spoken English Corpus”. In Knowles, G., Wichmann, A. and Alderson, P. eds. Working with Speech: Perspectives on research into the Lancaster/IBM Spoken English Corpus. London and New York: Longman, 20-37.

Tent, J. and Mugler, F. (1996). “Why a Fiji Corpus?” In Sampson, G. and McCarthy, D. (2006) Corpus Linguistics: Readings in a Widening Discipline. London: Continuum, 276-284.

The Chambers Dictionary. (2003). Edinburgh: Chambers.

Underhill, R. (1988). “Like is like, Focus”. American Speech, 63(3), 234-246.

Van Rooy, B. and Terblanche, L. (2006). “A Corpus-based Analysis of Involved Aspects of Student Writing”. Language Matters, 37(2), 160-182.

Watermeyer, S. (1996). “Afrikaans English”. In V. de Klerk (ed.) Focus on South Africa. Amsterdam/Philadelphia: John Benjamins, 99-124.

Williams, B. (1996). “The status of corpora as linguistic data”. In Knowles, G., Wichmann, A. and Alderson, P. eds. Working with Speech: Perspectives on research into the Lancaster/IBM Spoken English Corpus. London and New York: Longman, 3-19.

Wilson, A., Rayson, P. and McEnery, T. (2003). A Rainbow of Corpora: Corpus linguistics and the languages of the world. Munich : Lincom

Wynne, M. ed. (2005). Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books.
http://ahds.ac.uk/linguistic-corpora/ [Accessed 30.08.06].

Yoneoka, J.S. (2001). “The English Umbrella: Model of a Multicultural Language System.” Asian English Monographs, 7-2.

Zipf, G. K. (1949). Human behavior and the principle of least effort: an introduction to human ecology. Cambridge, Mass.: Addison-Wesley

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 yrs?
□Yes □No
If yes, please state which country………………………

6. Which language does/did your **mother** use **most often** in your home?
□Tamil □Telugu □Hindi □Gujarati □Urdu □English

□Afrikaans □Other (please specify)…………………

7. Which language does/did your **father** use **most often** 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………………..

Appendix B

Guidelines for Fieldworkers

Dear

Thank you for offering to help me by recording samples of conversations for the corpus. I know that I could not get ‘real’ data without your help!

Some guidelines:

1. Please record 2 x 30 min. conversations: one on each side of the tape.

2. Aim for a relaxed conversation in English, preferably with someone you know. Be as natural as you would normally be in such circumstances. I don’t mind slang, swear words, the use of ‘Indianisms’, laughing, joking etc. – if it comes naturally, just do it! And you may talk about any topic you like.

3. Try to avoid unnecessary background noise (e.g. lawn mowers, motor bikes, radios, television) when recording.

4. Always tell the person in advance that they are being recorded.

Confidentiality
Your own identity and the identity of the person you talk to will be protected at all times. Although I do not require the names of the people you record, I do require their permission. Please ask them to fill in the form and sign it.

Payment

30 min. recording	R15
60 min. recording	R30

Return of equipment
Please return the tape recorder and tape within two weeks so that I can arrange for you to be paid promptly for your efforts.

Thanks once again!

Leela Pienaar

Appendix C

Background to Recordings

Date recorded……………………….

Place recorded……………………....

Tape no.:…………… Side❑A ❑B

Number of people present……………

Private or public………………………

Audience or not (state approx. no.)………..

Domain (home, business, university, public place, school)………………………………

Topic(s)…………………………………………………………………………………

Any other relevant information……………………………………………………….

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………