

Corpus Linguistics and African Englishes

EDITED BY

Alexandra U. Esimaje

Ulrike Gut

Bassey E. Antia

Studies in Corpus Linguistics
88

JOHN BENJAMINS PUBLISHING COMPANY



Corpus Linguistics and African Englishes

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
<http://benjamins.com/catalog/scl>

General Editor

Ute Römer
Georgia State University

Founding Editor

Elena Tognini-Bonelli
The Tuscan Word Centre/University of Siena

Advisory Board

Laurence Anthony
Waseda University

Antti Arppe
University of Alberta

Michael Barlow
University of Auckland

Monika Bednarek
University of Sydney

Tony Berber Sardinha
Catholic University of São Paulo

Douglas Biber
Northern Arizona University

Marina Bondi
University of Modena and Reggio Emilia

Jonathan Culpeper
Lancaster University

Sylviane Granger
University of Louvain

Stefan Th. Gries
University of California, Santa Barbara

Susan Hunston
University of Birmingham

Michaela Mahlberg
University of Birmingham

Anna Mauranen
University of Helsinki

Andrea Sand
University of Trier

Benedikt Szmrecsanyi
Catholic University of Leuven

Elena Tognini-Bonelli
The Tuscan Word Centre/University of Siena

Yukio Tono
Tokyo University of Foreign Studies

Martin Warren
The Hong Kong Polytechnic University

Stefanie Wulff
University of Florida

Volume 88

Corpus Linguistics and African Englishes

Edited by Alexandra U. Esimaje, Ulrike Gut and Bassey E. Antia

Corpus Linguistics and African Englishes

Edited by

Alexandra U. Esimaje

Benson Idahosa University

Ulrike Gut

University of Munster

Bassey E. Antia

University of the Western Cape

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.88

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2018047817 (PRINT) / 2018051235 (E-BOOK)

ISBN 978 90 272 0219 2 (HB)
ISBN 978 90 272 6293 6 (E-BOOK)

© 2019 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Company · <https://benjamins.com>

Table of contents

Preface	IX
<i>Doug Biber</i>	
Introduction: Corpus linguistics and African Englishes	1
<i>Alexandra U. Esimaje, Ulrike Gut and Bassey E. Antia</i>	
Part I. Corpus linguistics and corpus building	
CHAPTER 1.1	
What is corpus linguistics?	7
<i>Alexandra U. Esimaje and Susan Hunston</i>	
CHAPTER 1.2	
Corpus-based research on English in Africa: A practical introduction	37
<i>Robert Fuchs, Bertus van Rooy and Ulrike Gut</i>	
CHAPTER 1.3	
The purpose, design and use of the Corpus of Nigerian and Cameroonian English Learner Language (Conacell)	71
<i>Alexandra U. Esimaje</i>	
CHAPTER 1.4	
Introducing a corpus of English(es) spoken in post-independence Namibia: Insights into corpus design and quantitative analyses	97
<i>Helene Steigertahl</i>	
CHAPTER 1.5	
The historical corpus of English in Ghana (HiCE Ghana): Motivation, compilation, opportunities	119
<i>Thorsten Brato</i>	
CHAPTER 1.6	
Addressing a coverage gap in African Englishes: The tagged corpus of Cameroon Pidgin English	143
<i>Gabriel Ozón, Sarah FitzGerald and Melanie Green</i>	

CHAPTER 1.7

- Practical corpus linguistics: Designing and exploiting a written corpus
for research with special reference to Cameroon English 165
Daniel Nkemleke

Part II. Corpus-based analysis of African Englishes

CHAPTER 2.1

- Evaluating explanations for past-time reference with unmarked verb forms
in African Englishes 185
Bertus van Rooy

CHAPTER 2.2

- The use of stance markers in West African Englishes 205
Ulrike Gut and Foluke Unuabonah

CHAPTER 2.3

- Namibian English on the web: Lexical and morphosyntactic features
in a Corpus of Namibian Online Newspapers (CNamON) 231
Alexander Kautzsch

CHAPTER 2.4

- Lexical expansion in Ghanaian English from a diachronic perspective:
A structural and semantic analysis 259
Thorsten Brato

CHAPTER 2.5

- Capturing the lexicon of Ugandan English: ICE-Uganda, its limitations,
and effective complements 293
Bebwa Isingoma and Christiane Meierkord

Part III. Applications of corpora in English language teaching and learning

CHAPTER 3.1

- A corpus-based analysis of conjunctive cohesion in English essays
of Nigerian university learners 331
Adeyemi Iyabo

CHAPTER 3.2

African corpora for standards in African academic English: Case studies on prepositions	355
<i>Josef Schmied</i>	

CHAPTER 3.3

Semiotic signature of transformation in a diachronic corpus of a South African political party	373
<i>Bassey E. Antia and Tamsyn Hendricks</i>	

Index	401

Preface

Doug Biber

Northern Arizona University

One important application of corpus linguistics is for the documentation of World Englishes: the many varieties of English spoken around the globe. While these descriptions are obviously of interest to linguists, they are probably even more important as the basis for language teaching materials and approaches. Thus, there have been numerous books and articles published over the past few decades focusing on the distinctive linguistic characteristics of particular World English varieties.

Given the importance of this line of research, it is no surprise that there have been many studies published on the use of English in Africa. There have also been several corpus-building projects, dating back at least to the 1990s, focused on representing the use of English in particular African contexts. However, no book to date brings together descriptions of those corpora and linguistic studies in a single outlet. The present book aims to fill that gap.

The editors – Alexandra Esimaje, Ulrike Gut, and Bassey Antia – are to be congratulated for assembling such an impressive array of authors, who collectively cover a wide range of topics relating to the corpus-based study of African World English varieties. The book is divided into three major parts: The first introduces the field of corpus linguistics, surveys previous corpus linguistics research in Africa, and provides detailed descriptions of several major corpora compiled to represent particular African English varieties. Then, the second part of the book presents several new corpus-based studies on the distinctive linguistic characteristics of particular African English varieties. The third part shows how corpora can be applied in English language teaching and learning in Africa. Taken together, the chapters in this volume provide a wide and thorough introduction to this topic. It is highly recommended reading both for scholars in Africa who are interested in learning more about corpus linguistics, as well as scholars around the world who specialize in the study of World Englishes and want to learn more about the initiatives in this domain of research currently taking place in Africa.

Introduction

Corpus linguistics and African Englishes

Alexandra U. Esimaje, Ulrike Gut and Bassey E. Antia

Corpus linguistics has become one of the most widely used methodologies across different linguistic subdisciplines. The list includes syntax, phraseology, discourse analysis, pragmatics, sociolinguistics, phonology, terminology, lexicography, language change, language teaching, and language acquisition. Especially the study of world-wide varieties of English uses corpus-based investigations as one of its chief methodologies, relying on a wide range of available corpora such as the ICE (International Corpus of English) corpora¹ (Greenbaum 1991), GloWbE corpus² and many national initiatives such as the SCOTS (Scottish Corpus of Texts and Speech) corpus, COCA (Corpus of Contemporary American English), and MICASE (Michigan Corpus of Academic Spoken English). The many African Englishes, however, are still severely underrepresented in this corpus-based research branch. To date ICE East Africa (Schmied 1989) and ICE Nigeria (Wunder et al. 2010) are the only available large reference corpora. Yet, this situation is currently changing with many new corpus initiatives both within and outside Africa that aim to compile both reference and highly specialized corpora of African Englishes and to thus promote corpus-based research on African Englishes and the use of corpora in pedagogic contexts within African institutions.

This book brings together a number of these cutting-edge initiatives. It is divided into three parts. In Part I, general introductions to corpus linguistics and to corpus linguistics in Africa in particular are given and a number of recently completed and ongoing corpus compilation projects on African Englishes are described. In Chapter 1.1, Alexandra Esimaje and Susan Hunston offer a theoretical introduction to the concept of corpus linguistics, explain some of the key terms and concepts used in it and consider the different types of corpora as well as the scope and applications of corpus linguistics.

1. <<http://ice-corpora.net/ice/>>

2. <<http://corpus.byu.edu/glowbe/>>

Chapter 1.2 by Robert Fuchs, Bertus van Rooy, and Ulrike Gut gives a practical introduction to corpus linguistics in Africa: it introduces the field, its methods, tools and resources and demonstrates the applications of corpora with three case studies on lexical differences between varieties of English, the present perfect in African Englishes, and the deletion of the phoneme /h/ in some sub-varieties of Nigerian English.

Chapter 1.3 by Alexandra Esimaje discusses the purpose, design and use of an error-annotated corpus of learner English from Nigeria and Cameroon. The chapter shows how the corpus supports the study of lexical forms and tenses used by some communities of learners of English in Nigeria and Cameroon.

In Chapter 1.4, Helene Steigertahl describes the compilation of a corpus of Englishes spoken in post-Independence Namibia and illustrates the kinds of analyses that the corpus allows by offering a quantitative analysis of the morphosyntactic structures of spoken Namibian Englishes.

In Chapter 1.5, Thorsten Brato reports on the compilation of the Historical Corpus of English in Ghana (HiCE Ghana), covering the period between 1966 and 1975, and discusses how methodological and theoretical concerns and challenges have affected the final corpus design. Moreover, he demonstrates that some lexical and grammatical properties of Ghanaian English have been in use for several decades.

Chapter 1.6 by Gabriel Ozón, Sara FitzGerald, and Melanie Green illustrates the compilation and uses of a tagged pilot corpus of spoken Cameroon Pidgin English (CPE). Quite apart from being a resource for the investigation of CPE, the tagged corpus is shown to be particularly useful for automatic retrieval of language phenomena above the level of the lexicon.

Chapter 1.7 by Daniel Nkemleke describes the basic steps involved in the design and exploitation of a corpus of Cameroonian English and presents results of two studies on the frequency and use of modals in the corpus.

Whereas studies in the first part focus on general principles of corpus research and on accounts of the building of corpora with exemplifications of how such corpora can be used, the thrust of studies in the second part of the volume is more on providing in-depth studies of how corpora have been and can be used to investigate a variety of linguistic aspects of African Englishes. In Chapter 2.1, Bertus van Rooy reports on a corpus analysis of spoken conversation in Nigerian English and Black South African English, which indicates that the non-marking of the past tense occurs in about one in every five contexts where an event in the past is represented.

In Chapter 2.2, Ulrike Gut and Foluke Unuabonah present an analysis of the use of stance markers in two West African Englishes (Nigerian English and Ghanaian English) and compare this to their use in British English. They show that the two West African Englishes use stance markers less than speakers of British English but reveal the same stylistic variability in the use of stance markers across different text

types. Moreover, they found systematic differences in stance marker usage between the two West African English varieties.

In Chapter 2.3, Alexander Kautzsch introduces the Corpus of Namibian Online Newspapers (CNamON) and demonstrates the corpus's potential for analyses on the linguistic levels of lexis and (morpho-) syntax while providing qualitative stock-taking and exemplary quantitative analyses of structural characteristics of English in Namibia.

Chapter 2.4 by Thorsten Brato exemplifies semantic analyses of lexical expansion in the Nativization phase of Ghanaian English based on the Historical Corpus of English in Ghana and the Ghanaian component of ICE. Analysing 'Ghanaianisms', innovative lexical items in Ghanaian English, he shows that semantic shift is the most important lexical process used in both periods, while word-formation processes like derivation or compounding play only a subordinate role.

In Chapter 2.5, Bebwa Isingoma and Christiane Meierkord investigate the lexicon of Ugandan English and document how analysing the Uganda component of the International Corpus of English allows for a contextualisation of observations made from current uses of English and for an assessment of the textual genres in which such innovations occur.

Part III of the present volume contains three studies that demonstrate how corpora of African Englishes can be applied to address issues in English language teaching and learning in Africa. In Chapter 3.1, Iyabo Adeyemi examines the use of conjunctions in written texts by Nigerian university students. She compares the use of conjunctions in the Nigerian Learner English Corpus (NLEC) to native-speaker use in the Louvain Corpus of Native English Student Essays (LOCNESS) and finds both underuse and overuse of specific conjunctions, which may contribute to an overall reduced stylistic variation and reduced cohesion in the students' writings.

In Chapter 3.2, Josef Schmied proposes the replacement of the traditional native-speaker model at African universities by a sophisticated and stratified corpus model of nation-, university-, and department-specific usage. Based on a qualitative and quantitative analysis of prepositions and their variation in Cameroon English academic writing, he suggests that corpus analyses should provide the basis for all discussions of new functional standards for African Englishes.

The final Chapter 3.3 by Bassey Antia and Tamsyn Hendricks shifts the focus away from varieties of English to other applications of corpus research. Their chosen focus is political discourse analysis. The authors query a diachronic corpus associated with a South African political party in order to determine if there is semiotic evidence of the party having transformed over time. They draw on social semiotics to argue that meanings that are socially significant can be read not only from what is said in corpora but also from a range of other resources, such as names of persons and places as well as language choices made in texts.

The field of African Englishes has elicited criticism (e.g. Banda 1996) for its blunt descriptions and generalisations that neither consider the local identities which language is contextually appropriated to perform, nor acknowledge that individuals may possess a rich English repertoire, dimensions of which are deployed differently according to situations. This is of course apart from the fascination with codifying errors or other deviations from an exogenous reference norm (see e.g. Jowitt 1991). Arguably inherent in both corpus linguistics and African Englishes is a view of language as a closed and enumerable set of features that are consigned to specific spaces or geographical borders. There is a sense, then, in which African Englishes and corpus linguistics have traditionally worked with an ontology of language as a set of fixed, stable and enumerable features that are tied to broadly defined contexts, including nation states.

This volume addresses these questions by covering a wide range of African Englishes and animating these discussions in a number of ways: For instance, Chapters 2.1, 2.2 and 2.3 provide rich data for reflecting on the ontology of language or the non-essentialism of labels (e.g. Cameroonian English, Ugandan English, Nigerian English). Moreover, Chapters 2.5 and 3.2 seek to explain features from within the context of situation and promote endogenous models of use, blend qualitative analyses into corpus work and thus address certain perceptions of corpus and African Englishes research. Beyond an audience of researchers and teachers interested in the linguistic properties of English, this volume should also be of interest to academics and teachers working in other languages. This book is addressed to readers who are both new to, and experienced with, corpus work.

We would like to acknowledge the contributions of all the authors and reviewers of the chapters which make up this volume, as well as the series editors.

References

- Banda, Felix. 1996. The scope and categorization of African English: Some sociolinguistic considerations. *English World-Wide* 17(1): 63–75. <https://doi.org/10.1075/eww.17.1.04ban>
- Jowitt, David. 1991. *Nigerian English Usage. An Introduction*. Lagos: Longman.
- Wunder, Eva, Voormann, Holger & Gut, Ulrike. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.

PART I

Corpus linguistics and corpus building

What is corpus linguistics?

Alexandra U. Esimaje and Susan Hunston

Benson Idahosa University / University of Birmingham

Corpus Linguistics has made great strides in language research and teaching but it is only fairly known, and thus its potentials lost, to many African academics and linguistic communities. The aim of this chapter is to introduce corpus linguistics to those African researchers and others who are not yet familiar with, or have limited knowledge of, the field and who are interested in using this method for linguistic analysis. The chapter introduces the concept of corpus linguistics (Section 1), explains some of the key terms and concepts used in it (Section 2), and considers the types of corpora, as well as the scope and applications of corpus linguistics (Section 3).

1. The concept of corpus linguistics

In this section we introduce and define corpus linguistics, offering an insight into the history, significance and scope of the field.

Corpus: A definition

A corpus (plural: corpora) is a collection of written or spoken material, occurring naturally, stored on computer, and typically used to carry out some kind of linguistic analysis. A corpus has three main characteristics:

- It consists of naturally occurring texts; that is, the texts are not written specially to go into the corpus, but are collected from sources such as newspapers, novels, blogs or (transcribed) conversations.
- The texts are selected to represent a language or a variety of a language (Francis 1982; Nelson 2000). This might be a geographical variety such as ‘the English spoken in East Africa’, or a register such as ‘the English spoken in academic conferences’, or the language of a specific group of people such as ‘English used by students in year 3 in secondary schools in Kenya’. The texts are chosen to be

as representative as possible of the chosen variety. Corpus design is discussed further in Section 2 below.

- It is relatively large. This means that there are so many texts in the corpus it would be difficult to study them ‘by hand and eye alone’ (McEnery & Hardie 2012: 2). Instead, the corpus is stored on computer and accessed using specialised software.

Most corpora consist of written texts, such as newspaper articles, academic journal articles, novels, blogs, text messages or student essays, and transcribed oral texts, such as speeches, conversations, or interviews. Increasingly, however, corpora incorporate visual material, such as pictures, photographs, or videos. This allows the range of languages studied to include signed languages used by deaf communities (Johnston & Schembri 2006); it also allows the study of written language to take account of accompanying pictures, and the study of oral languages to include the study of gesture (Adolphs & Carter 2013).

Corpus linguistics: History and significance

Corpus linguistics is any kind of language enquiry which relies on corpus data or corpus evidence to answer questions about language, including those related to the description of a language variety, theories of language, how language is learnt, and how language is used in specific contexts. The history of corpus linguistics is very much the history of corpora themselves.

The history of corpus linguistics might be traced back to the compilation of the Brown corpus (Francis & Kucera 1964), which consists of 1 million words of American English. It contains texts published in the USA in various fields in 1961 and could therefore be said to be a ‘snapshot’ of standard American English at that time. A counterpart to the Brown corpus, known as ‘LOB’ (Lancaster, Oslo, Bergen – the three universities that collaborated to produce the corpus) was made available in 1978. It consists of 1 million words of texts published in 1961 in the UK, from the same fields as the Brown corpus. It is therefore possible to directly compare American and British English. This reflects an early concern of corpus linguistics: to compare geographical varieties of a language. This concern has continued with the development of the International Corpus of English, which includes 1 million words from each of 23 varieties of English, allowing a direct comparison between them.

More recently, corpora have been compiled using the same design as Brown and LOB, but consisting of texts published in the early 1990s. These are the Freiburg-Brown corpus of American English (Frown) and the Freiburg-LOB corpus of British English (F-LOB). Using these corpora, it is possible to compare

British and American English from the 1960s and the 1990s. Language change is another chief concern of corpus linguists.

Later corpus linguistic work, such as that by Biber (1988), stressed the importance of register variation in describing language. Biber built corpora divided into registers such as news reportage, conversation, academic prose and fiction (Biber et al. 1999). Using a variety of methods, but most famously multi-dimensional analysis, he compared the sub-corpora, or registers, and noted that they vary in complex ways. Comparison between types of text is another recurring theme in corpus linguistics.

A corpus that was very large for its time was the Bank of English, compiled as part of a project at the University of Birmingham that led to the publication of COBUILD learners dictionaries. In the 1980s the Bank of English consisted of 200 million words. It was used to investigate what words mean, but also how they are used, and in particular the kinds of phrases they are used in. The word *back*, for example, has a slightly different meaning in each of the phrases *move back*, *think back* and *shout back* and is also used in idioms such as *behind someone's back*, *break the back of a task* or *back to front*. Looking at concordance lines from a corpus (see below) focused attention on the patterning of words and led to a renewed interest in the importance of vocabulary rather than grammar in the way language works.

The history of corpus linguistics might be summarized as a movement towards:

- more languages;
- greater size;
- more specialization;
- more use of corpora by individuals;
- more accessible software;
- more sophisticated statistical packages.

We can deal with each of these, briefly, in turn.

Languages

As noted above, the earliest corpora were of English, but these were soon followed by other languages. At first, it was difficult to build corpora of languages using non-Roman script, such as Arabic and Chinese, but these difficulties were soon overcome. Today, SketchEngine (a publicly-accessible corpus search tool: see below) gives access to corpora in 81 languages, including Afrikaans, Hausa, Igbo, Oromo, Setswana, Somali, Swahili, Tigrinya, and Yoruba.

Size

Since the 1970s, corpora have grown in size from a maximum of a million words to billions of words today. As long ago as 1991, Sinclair commented: “thirty years ago ... it was considered impossible to process texts of several million words in length. ... Today it is very popular” (Sinclair 1991: 1). The current Oxford English Corpus contains almost 2.5 billion words.¹ This has been made possible partly by the ease of downloading texts from the internet, but also by increases in computer power that make searching and processing large amounts of text feasible. As shall be noted below, larger corpora tend to be less carefully designed than smaller ones.

Specialisation

Many very large corpora are ‘general’ corpora, that is, they are intended to represent a language (such as English) or more usually a variety (such as American English or Nigerian English), and include examples of all kinds of registers. Increasingly, though, researchers work with more specialized corpora. At a recent conference on corpus linguistics (Corpus Linguistics 2017, University of Birmingham), papers were delivered based on specialized corpora such as ‘Irish post-primary school textbooks’, ‘UK universities’ careers pages’, ‘Advertisements for antidepressant medication in China’, ‘Judges’ statements about women found guilty of murder’ and ‘On-line reviews of video games’. Such specialized corpora are valuable for linking language to culture and ideology. Studies which at one time would have been carried out on a few texts, involving manual analysis, can now involve hundreds or thousands of texts and include quantitative as well as qualitative information.

Individuals and software

One reason for this development is that doing corpus linguistics no longer necessitates the involvement of large teams of people and specialized computer programmers. Because textual material is available in electronic form, and because corpus investigation software (e.g. Wordsmith Tools, Antconc, SketchEngine, Skylight – see below for details) are available that can be used by a non-expert, individuals with very little training can build and investigate their own corpora, or can easily access and investigate existing corpora. Online resources, such as those produced by UCREL at Lancaster University, can be accessed to permit corpus annotation and statistical processing by individuals who are amateurs in these fields.² University researchers, but also language teachers and language learners, can therefore carry out simple corpus research with only the most basic computer training, though reliable internet access is a necessity.

1. <<https://en.oxforddictionaries.com/explore/oxford-english-corpus>>

2. <ucrel.lancs.ac.uk>

Statistics

At the same time, however, there has been a growth in the sophistication of the statistics used in corpus linguistics. Many early corpus linguists used simple frequency counts, or measures of significance that were built into available corpus software (Hunston 2002, though see also Biber 1988 for a much more complex approach). Present-day researchers are likely to use much more sophisticated measures of significance (e.g. Gries 2010), or complex means of clustering texts or features (Durrant 2014; Thompson et al. 2017). As Hardie (2017) says, however, corpus linguistics begins and ends with the simplest methodology: the study of concordance lines (see below). Many of the ground-breaking insights associated with corpus linguistics can be made through the simple observation of these lines, even if complex statistics are used to corroborate or adjust the concordance line findings.

It is no exaggeration to say that corpus linguistics has revolutionised the study of language. Instead of relying on intuitions about what is or is not said, or basing a theory on very small samples of language, corpora have made it possible to look at large quantities of language as it is actually used (Aijmer & Altenberg 1991: 2; McEnery & Wilson 1996; Hunston 2002). It is possible to see how language varies according to context and to apply quantitative methods alongside qualitative ones (Meyer 2002: 6). Corpus linguistics is now widely used to study the language of language learners (Granger 1998), the practice of translators (Kruger et al. 2011), how newspapers (mis)represent particular communities (Baker 2008; Partington et al. 2013), and how academic knowledge is communicated (Hyland 2012). It has become central, not only to linguistics itself but to the applications of linguistics.

Corpus linguistics: Historical debates

The kind of research described as corpus linguistics covers an expanding range of methodologies and applications. These methods have affected our understanding of what languages are like and how they are used. Such diversity has led to a number of debates over the years; these are still worth mentioning because much work in corpus linguistics finds a compromise between the two positions in each.

Research which uses a corpus is often referred to as ‘corpus-based’ (that is, based on a corpus rather than on intuition), but the term ‘corpus-based’ has also been used in contrast with ‘corpus-driven’ to refer to contrasting ways of conceptualizing corpus research. The term ‘corpus-driven’ was coined by Tognini-Bonelli (2001) to indicate that the researcher approaches the corpus without preconceptions about linguistic categories. She contrasts this with ‘corpus-based’ research, which uses existing linguistic categories as the basis for corpus searches. Historically, corpus-driven research has tended to prioritise lexical approaches to language. Such approaches to words, grammar and discourse (e.g. Sinclair 1991,

2004; Hunston & Francis 1999; Hoey 2005) derive descriptions of language that are possible only with corpus resources. Corpus-based research is more likely to be based on traditional linguistic concepts, for example, grammatical categories. Biber (2009) summarises the distinction thus:

corpus-based research assumes the validity of linguistic forms and structures derived from linguistic theory. The primary goal of research is to analyse the systematic patterns of variation and use for those pre-defined linguistic features [but] corpus-driven research is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus.

The ‘corpus-based’ / ‘corpus-driven’ distinction has been extensively discussed (see, for example, McEnery et al. 2006 for a critique), but has to a large extent disappeared as a point of contention. Increasingly, all corpus linguists use some pre-conceived ideas and also search for novel insights, so find themselves on a cline between the two extremes.

The second debate is concerned with the relative importance of theory and method in corpus linguistics. As noted above, corpus linguistics as a methodology can be applied to many aspects of Linguistics: the study of grammar, language change, variational sociolinguistics, literary stylistics and many others. McEnery and Hardie (2012) are among those that argue that it is a method that can be used in conjunction with any theory of Linguistics. Nowadays, even proponents of Universal Grammar have recourse to corpora (e.g. Leivada et al. 2017). Many corpus linguists, however, have studied the particular role that words and phrases play in the way languages work (e.g. Sinclair 1991, 2004). This emphasis on words has led to a view of language that sees lexis as particularly important to language structure (Sinclair 1991) and to language acquisition (Hoey 2005). Some have gone so far as to suggest that corpus linguistics represents a novel theory of language based on words and discourse (Teubert 2005). In other words, researchers differ on whether corpus linguistics is best seen as a new way to answer old questions, or as constituting a novel approach to language (see Taylor 2008 for a summary of the issue). The difficulty with seeing it as ‘only’ a method that is at the service of various approaches to linguistics (as Gries 2006; McEnery et al. 2006; Thompson & Hunston 2006; McEnery & Hardie 2012 do) is that this appears to be somewhat dismissive, with a corpus being just one research tool among many. The difficulty with seeing it as a theory (as Tognini-Bonelli 2001; Mahlberg 2005; Teubert 2005 do) is that this would limit the scope of corpus linguistics to a particular set of ideas, and to deny that researchers taking other theoretical standpoints are ‘doing corpus linguistics’.

A compromise between these positions, and one that is increasingly adopted, is to argue that corpus linguistics is a method that can be applied to any language

theory, but that it tends to lead towards particular theories of language, some of which have been developed specifically in response to corpus evidence. These include proposals that a distinction between lexis and syntax is artificial, and that the essence of the structure of language lies in the patterning of lexis (Sinclair 1991, 2004). Scott and Tribble (2006: 4f.) argue that although corpus linguistics is primarily a method, it is a method that “[has] a potential to shake the foundations of the field”; this is the case because the software enables the researcher to identify ‘latent’ patterns, that is, those which cannot be observed in other ways.

2. Key concepts in corpus linguistics

Some of the key concepts and terminology in the field of corpus linguistics are explained here under the following headings: corpus design, corpus output, and corpus annotation.

Corpus design

If a researcher wishes to compile a corpus for their own use, a number of issues have to be considered: what is going into the corpus (content), how large it should be (size), and to what extent the corpus can be balanced and representative. Each of these issues is explained here.

Content

What goes into a corpus will depend on the use that is to be made of the corpus (Hunston 2002: 25). This means that the content of the corpus will depend on the aim of the study. Once that aim is decided, the texts have to be found. For example, if the aim of a project is to study research articles written in the field of applied linguistics, journals devoted to applied linguistics will be the most obvious source of material. It is likely that a variety of journals will be selected, with expert informants consulted to recommend the most highly-regarded journals in the field from which texts can be chosen. A more difficult task would be to compile a corpus consisting of newspaper articles dealing with a given topic, such as educational reform. The difficulty here is to identify which articles have this topic. Typically, a search is made of digital editions of selected newspapers, using terms such as ‘education’ and ‘reform’. Frequently, though, the researcher has to scan each article visually to ensure that it really deals with the topic being studied. Yet another issue arises when very short texts, such as tweets or text messages, are being studied. If the aim is to

collect a corpus dealing with a given topic, the Twitter symbol # is often used to ensure that texts relevant to the topic are collected. In short, a corpus will be designed based on the research question being asked. The content is restricted to the target text types, but should also be as varied as possible within that specification (and see the discussion of representativeness below).

A more prosaic consideration is what texts are available for use. Some genres, such as police interviews, disciplinary hearings or medical consultations, take place in highly confidential conditions, and it is difficult to obtain access to them (Carter 2009; Atkins et al. 2016). Even if permission is given for a corpus to be constructed, it may well be restricted in terms of which researchers have access to it. Issues of copyright or ethical considerations are also important, and can restrict corpus construction. It is not ethical, for example, to build of corpus of student essays unless each student has given informed consent for their work to be used in this way.

Size

As noted above, corpora have increased in size dramatically since the early days of corpus linguistics and it is usually said that the larger the corpus the better. This is because many words and phrases occur relatively infrequently, so to study them adequately, a large amount of data is needed. For example, in the Times sub-corpus of the Bank of English, the word *very* occurs 678.3 times in every million words, but the word *exceptionally* occurs only 9 times in every million. Therefore, to study the patterning of *very*, one million words would be quite sufficient, but to repeat the exercise with *exceptionally*, many millions of words would be needed.

In many studies, however, the size of the corpus may be limited by the availability of data. If a corpus is being built of student essays, for example, the size will be limited to the number of students whose permission can be obtained to use their essays, and by the number and length of essays they write. The pressure of time may be another limiting factor. To take the same example, if the essays have to be typed from manuscripts, or if a decision has been taken to identify all errors in them (see below), then the size of the corpus will be limited by the availability of person-time available to carry out these tasks. Another consideration (see, for example, Nelson 2000, citing Murison-Bowie 1996), is the balance of the corpus. If the aim of the student essay corpus is to compare, say, students studying Politics and those studying Theology, with the same number of essays from each discipline, and if there are 10 times as many students taking Politics as Theology, then the size of the corpus will be limited by the number of students studying Theology. In general, studies in corpus linguistics tend to prioritise either size or balance of the corpus.

Balance

In those cases where a corpus will consist of a number of different parts (or sub-corpora), and where a large amount of data is available for each part, it may be wished to achieve balance between the sub-corpora. Biber (1993) gives details on how to achieve corpus balance through sampling. First the population to be studied is defined; for example, the population may consist of publications in English in a certain country in a certain year. This is the sampling frame. Then the hierarchical structure (or strata) of the population is determined; for example, the publications are divided into newspapers, novels, non-fiction books, web pages and so on. The next consideration is how many texts from each stratum should be included. This involves some difficult decisions. For example, there will probably be many more newspapers than novels published. A very large corpus consisting of thousands of newspaper articles but only two novels could be built, but that would not be balanced between newspapers and novels.

To take another example, suppose it is decided to build a corpus to trace changes in literary style over a period of 100 years. The sampling frame may be determined to be all novels published, in a particular country, from 1915 to 2015. A list might be drawn up of 5,000 novels, comprising the population. The strata might be described in terms of time period. For example, a logical decision might be to have five time periods: 1915 to 1934, 1935 to 1954 and so on. Another decision might be to distinguish between romantic novels, children's novels, science fiction and so on. The consequence might be a total of 15 different sub-corpora, each a combination of genre and time. The decision then has to be taken of how many novels to include in each sub-corpus. What is to be done if, for example, no science fiction novels were published in that country between 1915 and 1934? What is to be done if some novels are very long and others much shorter? Or if twice as many novels were published in 2015 than in 1915 – should the 2015 sub-corpora be twice as large as the 1915 ones? It might be decided that, in order to have the maximum number of novels represented in each sub-corpus, whole novels will not be included, but samples of about 2,000 words each. And so on. It can be seen that achieving balance involves a great many subjective decisions.

In practice, most corpora are a compromise between the demands of size and balance. For example, suppose someone wishes to build a corpus that is balanced between spoken and written language. It is more time-consuming to record and transcribe spoken language than to collect written language, which is often available electronically. If the quantity of each is to be exactly equal, then the corpus will be limited in size. A larger corpus will be unbalanced. Most corpus investigation software allows the researcher to compensate for the problems of an unbalanced corpus by calculating the 'normalised' frequency of words. An example was given above of

the word *exceptionally*. We can compare the frequency of this word in three newspaper sub-corpora of the Bank of English: *exceptionally* occurs 469 times in the Times sub-corpus, 287 times in the Guardian sub-corpus and 266 times in the Sun sub-corpus. We might think that *exceptionally* is used most in the Times and about equally in the Guardian and the Sun. However, the three corpora are of very unequal sizes, though all are large. The Times sub-corpus has nearly 52 million words, Sun has nearly 45 million, and the Guardian has over 32 million. To compensate for this difference, the frequency can be normalised so that the number of instances of *exceptionally* per million words is obtained. In the Times this is 9.0 instances, in the Guardian it is 8.9 instances and in the Sun it is 5.9 instances. This shows that *exceptionally* is used with equal frequency in the Times and the Guardian and with less frequency in the Sun. Thus, although the three sub-corpora are not balanced, accurate comparative information about frequency can be obtained.

Representativeness

As well as being balanced, a corpus is often designed to be representative of a given population. According to Biber (1993), “representativeness refers to the extent to which a sample includes the full range of variability in a population” (243). As an example, consider the hypothetical corpus of literature from 1915 to 2015 mentioned above. It was proposed in that example that many kinds of fiction would be included, the aim being to achieve a corpus that was truly representative of the variety of literature published in that century.

How difficult it is to achieve true representativeness depends on the aim of the corpus compilation. For example, it may be possible to achieve a ‘saturation’ corpus, that is, a corpus that contains all possible texts. A corpus containing all the plays written by Shakespeare, for example, is of necessity representative of Shakespeare’s plays. If it is wished to compile a corpus of newspaper texts reporting a natural disaster in a given country it may be possible to identify all the articles written, and to make a balanced selection from them, or include all of them. On the other hand, if the objective is to study tweets written by teenagers the task of achieving representativeness may be much more difficult, as it may not be known how many such texts exist or who is writing them. If the population or the strata are not known, then achieving either balance or representativeness can be a problem.

To some extent, size, balance and representativeness are conflicting requirements. To return to the literature example: to be truly representative, we may wish to include poetry in the corpus. Poems are short compared with novels, so a corpus that is representative of fiction and poetry is unlikely to be balanced between them. If the corpus is very large, the poetry may effectively disappear in it. If equal amounts of poetry and fiction are used, the corpus will be too small. Again, building

the corpus is a compromise between these demands. In many cases, what is to be represented is adjusted to allow for balance and corpus size. For example, suppose we are building a corpus of university student essays in order to study their use of academic English. The population is all students, at all stages of their university life, studying all subjects. To be representative we may wish to collect essays from all kinds of students, in each university year and in each subject. This would lead to a corpus that is composed of many very small sub-corpora, however, and that would make comparison difficult. Another would be to focus on some subjects only, collecting more from each, so that the size of each sub-corpus is as large as possible and comparison between the sub-corpora is more meaningful. However, the representativeness of the study is reduced.

Corpus output

Once a corpus has been built, it is investigated using specialised software. Available software and corpora are listed below. Most software packages will give information about word frequency, collocations and keywords. In addition, they allow the researcher to search for specific words and see these words in context: the ‘concordance line’. This section explains these terms. These kinds of output can be used without much formal training.

A word frequency list shows all the words in a corpus and how many times that word occurs. Frequency can be shown as ‘raw’ frequency (the number of times the word occurs in the corpus) or as ‘normalised’ frequency (the number of times, on average, the word occurs in each thousand or million words of the corpus). The normalised frequency is useful for comparing word frequency in corpora of different sizes, as demonstrated above. A word frequency list is often the starting point for corpus research, as researchers often wish to study the most frequently-used words, or they may wish to study words of a particular kind. In a corpus of political speeches, for example, the researcher may study the word frequency list to identify words relating to education, or finance, or community relations.

Concordance lines, as Hardie (2017) has said, are the bedrock of corpus linguistic research. This is because it is in concordance lines that the researcher actually sees the language that is in the corpus. To obtain a set of concordance lines the researcher searches for a word. The software shows every instance of that word in the centre of the screen, with the words immediately before and after it to the left and right. Concordance lines can be sorted alphabetically, based on the words occurring to the right or left of the centre (or node) word. Bringing all the instances of the word together and sorting the words right and left allow the researcher to see patterns that might otherwise be overlooked. The Appendix to this chapter shows 50 concordance lines centred on the word *obvious*, sorted so that the words

immediately to the right of the node word are in alphabetical order (*advantages*, *ain't*, and etc.). (The lines come from the Bank of English corpus and are randomly selected from the total of 26,997 instances of *obvious* in that corpus.) A researcher, or language learner, might notice the following:

- *obvious* is used both attributively (*obvious advantages*) and predicatively (...*was obvious*).
- It is often followed by *that*.
- The phrases *it is obvious that*, *it is obvious to ... that* and *made it obvious that* are also common.
- It is used in place of a noun, following *the* e.g. *to state the obvious*, *with some new examples of the obvious*.

One of the things that become clear from concordance lines is that certain words frequently go together. For example, *obvious* is often followed by *candidate* or *danger* or *reason*. Words that co-occur in this way are known as collocates. Collocation can be observed in concordance lines, but it can also be identified by software that calculates the significance of each collocate pair, taking into account more information than the individual researcher can handle. For example, *reasons*, *choice*, *answer* and *reason* are the most significance lexical collocates of *obvious*. (The lexical collocates exclude grammatical collocates such as *it* and *that*.) That information is based on all the 26,997 instances of *obvious*.

It is often useful to know which words are particularly frequent in a specialised corpus compared with a more general corpus. Such words are known as keywords. For example, Scott and Tribble (2006: 60) list the words that are significantly more frequent in Shakespeare's 'Romeo and Juliet' compared with all Shakespeare's plays. These include character names (*Capulet*, *Mercutio* etc.) and also words connected with the plot or themes of the play, such as *banished*, *death*, *love* and *poison*. Scott's Wordsmith Tools (1996) was the first software to calculate keywords in this way.

Corpus annotation

It is possible to move straight from compiling a corpus to investigating it, using concordance lines and collocates. However, in many cases a corpus is annotated before it is used, and this takes specialised skills. For a detailed treatment of corpus annotation, see Garside et al. (1997). Researchers such as Leech (1997) describe annotation as an enrichment of a corpus, and corpora often need to be annotated before particular kinds of study can be carried out. The annotation sometimes involves human judgement, and this restricts the size of the corpus that can be used. Increasingly, however, annotation is applied automatically or at least

semi-automatically (for example, it is carried out by a computer programme but checked by a human being). In this section, various kinds of annotation, including tagging and parsing, are described.

Tagging

Tagging refers to the assigning of a part of speech to each word in a corpus. It is a pre-requisite annotation to other complex annotations such as syntactic or semantic annotation and it serves the useful function of dividing text into words and sentences (Garside & Smith 1997: 102). As Baker (2008: 40) remarks, tagging enables more specific analyses and saves a lot of the researcher's time. For example, a researcher looking for the noun *light* in a corpus will find it easier and quicker if they can find just the nouns and not also the verb *light* and the adjective *light*. When a corpus is tagged, a code is added to each word; these codes are also called tags. In the BNC for example, a tag consists of three characters; the first two characters indicate the part of speech and the third shows the sub-category. For instance, the tag AJC represents comparative adjective: AJ stands for adjective and C stands for the sub-category, comparative (e.g. *better*, *older*). There are many different sets of tags; a corpus annotator is at liberty to use either an existing tagset or to devise their own. Annotation can be done manually or automatically using computer tagging software such as CLAWS (Garside & Smith 1997). However, it is to be noted that automatic taggers are not always accurate and it is frequently the case that the tags added by automatic taggers are corrected manually; however, manual correction is not feasible in a very large corpus. Hunston (2002) confirms this in the assertion that automatic tagging will be reasonably but not entirely accurate. The example below of a tagged text from the BNC comes from Leech (1997: 30). It consists of the sentence *Even the old women manage a slow Buenas, just as they're passing you*. The tags used are:

- AVO = adverb: *even, just*
- ATO = article: *the, a*
- AJO = adjective: *old, slow*
- CJS = subordinating conjunction: *as*
- NN2 = plural noun: *women*
- PNP = personal pronoun: *they, you*
- PUN = punctuation mark: ,
- UNC = unclassified (non-English) item: *Buenas*
- VBB = present tense forms of BE: *'re*
- VVB = verb, finite base form: *manage*
- VVG = verb, -ing form: *passing*

Figure 1 shows the tagged text.

```
<w AVO>Even <w ATO> the <w AJO> old <w NN2>women
<w VVB>manage <w ATO>a <w AJO>slow <w UNC>Buenas<c PUN>,
<w AVO>just <w CJS>as <w PNP>they<VBB>'re <w VVG>passing
<w PNP>you<c PUN>.</PUN>
```

Figure 1. A tagged text extract (adapted from Leech (1997: 30))

Tagged output like this is difficult to read, and normally the tagging remains hidden from the researcher. However, some corpus searches and processes will make use of these hidden tags, for example:

- searching for *light* as a noun only;
- searching for all combinations of ‘adverb + adjective’ in a corpus, to identify which adverbs and adjectives are likely to be used together, such as *utterly ridiculous*;
- searching for all instances of *criminals* preceded by a preposition, to identify phrases such as *rights of criminals*, *a haven for criminals*, *dealing with criminals*;
- calculating the relative numbers of nouns and verbs in essays written by different groups of students, to see which have the most nominalised writing style (Halliday 1985/1994).

Parsing

While tagging indicated the grammatical class of each word in a corpus, parsing identifies the constituent parts of sentences. Parsing, according to Hunston (2002: 19) “is the analysis of texts into constituents, such as clauses and groups, that is, doing a grammatical analysis”. She explains that a parser identifies the boundaries of sentences, clauses and phrases and assigns labels to the parts identified. Examples of identified elements are: noun clause, verb phrase, prepositional phrase. As with tagging, parsing is laborious and as such is often done automatically. As with tagging, accuracy is an issue, and manual editing is usually needed to enhance this. This is particularly true because parsers work with tagged corpora, so if the tagging is inaccurate the parsing will be too. Figure 2 is an example of a parsed sentence from a corpus as shown in Leech (1997: 37).

```
[S But_CCB ,_, [[N the_AT thing_NN1 N][V was_VBDZ V]] ,_, [N you_PPY N] often_RR
[V found_VVD [Fnthat_CST [Faalthough_CS [N you_PPY N][V had_VHD [N a_
ATreserved_ll seat_NN1 N]V]Fa ,_, that_CSTthere_EXjust_RR [V would_VMn't_XXbe_
VBO [N room_NN1 N][P on_ll [N the_AT train_NN1 N]P]V]Fn]V] _.. S]
```

Figure 2. A parsed sentence (adapted from Leech 1997: 37)

This parsing identifies the following as phrases:

- the thing
- a reserved seat
- wouldn't be
- room on the train
- on the train
- the train

It identifies the structure of the sentence as:

- But
 - the thing was
 - you often found
 - that
 - although you had a reserved seat
 - that there just wouldn't be room on the train

As with tagged corpora, the parsing annotation is often hidden from the researcher but is useful in finding all instances of a given structure, such as all verbs followed by that-clauses (e.g. *found* + *that there just wouldn't be room on the train*), or in calculating the relative frequency of such structures when comparing corpora.

Error tagging

Error tagging is a specific kind of annotation carried out on learner corpora with the dual aim of tracking the learning process and identifying the learners' areas of difficulty. It is one important way of investigating a corpus of learner language. As McEnery et al. (2006) say, corpora annotated for errors reveal the relative frequency of error types produced by learners of different L1 backgrounds and proficiency levels. A common way of proceeding is as follows. The researcher first decides on an error annotation scheme, that is a list of errors and their codes, and this is fed into the computer software. The linguist carefully reads through the corpus to identify instances of particular errors according to their judgement and uses the codes in the software to label them and to provide a corrected version. This is a time-consuming process that has not so far been automated. The researcher can then specify a type of error and the software will calculate the frequency of that error type. In setting the initial tagset, the researcher can be more or less specific as required. For example, a single tag (e.g. 'tense') can be used to capture all instances of such errors, or categories of tense can be specified and codes assigned to each one. For example, present and past can be distinguished and lexical verbs (e.g. *walk*,

sing) can be distinguished from *be*, *do* and *have*. An instance of annotation of syntax errors is shown below in which each error is shown by a code; in addition, a colour scheme is used for better visibility e.g. red for spelling, purple for tense, and blue for determiner. A few error-tagged corpora available are the Cambridge learner corpus, the Longman learners' corpus (McEnery et al. 2006) and the corpus of Nigeria and Cameroon English learner language (Conacell). Figure 3 is an example of error-tagged learner English from Conacell. In this example, a student's essay is error-tagged and shows the various errors identified, tagged and/or corrected. The tags are named following the error types, for example, spelling errors are tagged 'spelling' and are shown in red while tense errors are tagged 'tense' and shown in purple. In both cases, the software enables the annotator to both identify the error and provide the correct forms. This means that learners are able to 'notice' the errors as well as learn their correct forms.

An **embarrassment** is a situation in which you **fine** yourself which you less expected. This situation is alway a bad one which most at times **disgraces** you and **in** public. I have had many of such challenges in life but below, we shall see the most **embarrassing** one **was** which made me almost **lost** my character trade of contentment and satisfaction.

On this particular day, we were **suppose** to have a contest for the miss of my village. I knew I could make it because I had all that was required from the **delegates** candidates. This particular contest, I saw it as a **way** step into my career of modeling **but** the **embarrassment** I **recieved** made me **lost** that dream till today. I prepared so well that I knew nobody could meet up with me **but** just a small thing spoiled everything.

The night before the contest, I had invited all my friends and **talk** to them about it with **slot** of confidence. There was one thing which made me mad. I had my best dress prepared for this day but I had no shoes, not even a sister that I could get one from. I had to borrow one from my neighbour which was really nice and I practiced with it the whole night and everything was fine.

On **The** set day, I **met** with the other girls, we discussed and shared ideas and **every-body** was set to go **in** when time came **but** **Every-body** was to enter when your name **is** called. The judge called about five names and they performed so well, I could hear the crowd shouting from where I was standing. I was the next person he called. I **move** in but I felt some strange behaviour with the shoe I had **practice** with. I struggled **then** to **wait** walk with it but behold when I got to the center of the stage, my heels came out. I was **domfounded**, I could not take any further step but the judge told me to continue. **with** the **chilidings** I **recieved** from the crowd I was **confuse** and more than the word **embarrassed**. I **look** at the other contestants who had already talked **and** on my right, they were all laughing, deep in me I knew I had already lost it. I **strugled** to talk but all **what** I prepared **out** of the embarrassment had left my head. I could hardly say out anything clear so I was asked to go. **while** leaving the stage I was called to come and pick my broken heel along. **with** all the shame I took **it** out with me.

When I got home that day I could not even explain to anyone what happened. when I **open** my mouth tears **will** drop from my eyes and I will stop. The other girls came to our **housemy** home that day, I **though** they were coming to consol me but unfortunately for me **They** came to mock me and left. Even my own family members could not encourage me. I was all left alone.

Figure 3. Example of error-tagging in Conacell

Semantic tagging

A relatively recent innovation in corpus annotation has been semantic tagging (Rayson 2008), where words in a corpus are matched against a pre-prepared set of semantic labels (e.g. 'food and farming', 'education', 'science and technology', 'time', from the UCREL Semantic Analysis System (USAS)). This overcomes one problem with examining word frequency in a corpus: each word is a unique entity, and a word frequency list shows each one separately, but often a researcher will

wish to add together the frequency figures for words that mean similar things or that belong to the same semantic domain (such as ‘education’). It is difficult for the individual to do this with any degree of consistency, but the semantic tagger will perform the task without subjectivity. This allows very broad comparisons between corpora in terms of what they are about. For example, in two corpora that consist of texts concerned with global warming, one corpus may contain more scientific terms than the other, and one may be more concerned with agriculture than the other. Semantic tagging facilitates this kind of comparison.

It is apparent from this discussion that tagging and annotating a corpus involves both time and expertise. While many researchers will wish to annotate the corpus they are using, it is worth noting that corpus linguistics can be undertaken on a corpus which has minimal tagging (e.g. Part of Speech tags only), or has no tags at all. The questions that can be answered using a tagged or an untagged corpus are different, however.

3. Types and applications of corpora

There are different classifications of corpora in the literature of corpus linguistics. In this section we describe eight types as follows: general, specialised, comparable, parallel, learner, pedagogic, diachronic, and synchronic as described in Sinclair (1991), Kennedy (1998), Hunston (2002) and McEnery et al. (2006). This section will also address the question of what kinds of linguistic investigation are amenable to the corpus approach. First, we shall discuss the various types of corpora and the software available for their use.

Types of corpora

General corpus

A general corpus is a systematic collection of texts of different types and genres, which is normally very large in size and used for comparing languages (it is also called a reference corpus). Kennedy (1998) observes that such corpora are made simply with the aim of providing database for unspecified linguistic research, hence the name, and notes that since such corpora are designed to meet a wide range of research tasks, they normally aim at representativeness and balance; having in them texts representing different genres and different domains of language use, spoken and written alike. However, Hunston (2002) argues that such a corpus is unlikely to be representative of any particular ‘whole’ but will simply include a very wide spread of texts. Examples of general corpora are the British National Corpus of 100

million words and the Bank of English of 400 million words. The BNC, for instance, represents a wide range of spoken and written British English from the 1990s. It includes texts from newspapers, periodicals, academic books and journals, fiction, letters, memoranda as well as college and university essays. A general corpus is useful for answering questions about specific aspects of the given language, whether grammar, vocabulary or discourse and it can also be a reliable basis for contrastive analyses especially in the case of corpora of other languages with a comparable sampling frame (Kennedy 1998; McEnery et al. 2006).

Specialised corpus

Unlike a general corpus, a specialised corpus is compiled with a particular research task in mind. It therefore is representative of a given type of text and used to investigate a particular type of language or topic. For instance, a collection of newspapers articles, works of genres of fiction, works of a particular author, in a specific context or time, students' conversations, female writings are all illustrative of this kind of corpus. As Hunston (2002: 14) says, however, although there is no limit to the degree of specialisation involved, parameters are set to limit the kinds of text included. Specialised corpora are often also compiled by individual researchers to investigate language phenomena of interest to them. For example, Esimaje (2012) compiled a corpus of religious sermons of a well-known preacher in Nigeria in order investigate the lexical characteristics of his sermons. This corpus, called Chris Oyakhilome sermons (COS), contains 400 sermons and is made up of 65,000 words. Many such corpora are not published and consequently are not known or at best little known. Some of the examples of specialised corpora found in the literature of corpus linguistics are the Michigan corpus of academic English (MICASE) of 1.7 million words and the corpus of professional spoken American English (CPSA) which is made up of two sub-corpora of 1 million words each; one of these is a collection of committee meetings and the other is a compilation of transcripts of White House press conferences (McEnery et al. 2006). In some cases, a specialised corpus can be built from a general corpus. Su (2015), for example, built a corpus of biographical texts from a part of the BNC.

Comparable and parallel corpora

Comparable and parallel corpora are used in studying translation between languages. A comparable corpus consists of two or more sub-corpora, each comprising texts in one of the languages concerned. The sub-corpora are designed to contain the same number of the same kinds of texts, so that they can be directly compared. For example, the frequency of coordinating conjunctions (*and*, *but* etc.) in each language can be compared. Another use is to find translation equivalents. For example, in English the verb *do* is used with nouns such as *housework* or *homework*.

If the equivalents of *housework* and *homework* are found in a comparable corpus of another language, the verb(s) used with them can be observed. A parallel corpus also consists of two or more sub-corpora, but in this case the texts in one corpus are translations of the texts in the other corpus. Parallel corpora make it possible to study in detail how particular words and phrases have been translated and therefore what the candidate translation equivalents are. Parallel corpora can also be used to examine the grammatical choices that translators have made when rendering a text in another language.

Learner corpus

A learner corpus is a collection of texts produced by learners of a language. The purpose of building a learner corpus is often to study features of interlanguage, frequently in comparison with native language, and to analyse learner errors with the aim of identifying learner needs, as exemplified above. This second aim has been a topic of controversy because some believe that learner language is a genre in its own right and should not be expected to match native-like production (Granger 1998) nor be subjected to correction (Truscott 1996; Broughton et al. 2003: 135). It is certainly important to differentiate between individuals producing language in the context of learning it and individuals (possibly the same people) producing language in the context of using it (Mauranen 2012: 3f.). While tracing the history of computer text corpora, Aston et al. (2004) note that corpora interact with learners in three main ways: learners may be authors of corpora; learners may be beneficiaries of corpora, and learners may be users of corpora. Thus, there are corpora by learners, corpora for learners and corpora with learners. An example of learner corpus is the international corpus of learner English (ICLE) compiled by Sylvianne Granger (1998) and her team. This includes samples of writing from learners of English in 20 countries. There are others such as a corpus of Polish learners of English (PELCRA; Lenko-Szymanska 2004) or the corpus of Nigeria and Cameroon English learner language (Conacell; Esimaje 2016), which has proved very useful in various analyses of learning needs and contrastive studies of learning in the two contexts. Also of use to learners is a pedagogic corpus which is a collection of all the texts a learner has been exposed to, often collected by teachers or researchers, which may consist of all the books, reference materials and tapes a student has used in the course of their study.

Diachronic corpus

A diachronic corpus is a corpus of texts taken from the same language at different periods of time. Such corpora are useful for tracking changes in language evolution (McEnery et al. 2006: 65). The best known example of diachronic corpus is the Helsinki corpus which contains about 1.5 million words of text samples dating from

the 8th to 18th centuries. A more recent example is the Diachronic English Web Corpus which is made up of texts extracted from the internet between 2000 and 2010.³ The corpus is divided into months, with the texts for each month comprising about 1 million words. It is therefore possible to track the use of a word or phrase over that time period. Table 1, for example, shows the frequency of the word *crisis* in the corpus in each year of the first decade of the 21st century.

Table 1. Frequency of *crisis* 2000 to 2009

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
379	542	637	611	604	677	601	549	1111	1033

This shows a relatively steady increase in the use of the word over the decade, as evidenced by internet use.

A corpus that is not diachronic can be described as a synchronic corpus.

Available corpora and software

As noted above, it has now become possible for researchers to access corpora compiled by other people, and to use software to investigate corpora they have compiled themselves. This section lists some of the most commonly-used corpus investigation software and some of the corpora accessible to researchers and learners. The first three entries in this list are examples of software only, designed to be used with the researcher’s own corpus.

AntConc. This is a set of tools developed by Laurence Anthony. It is free of charge and downloadable. It can be used with any corpus compiled by the individual researcher.

CQPweb. This is a set of tools using Open Source software. The researcher can put any corpus into it.

WordSmith tools. This is a set of tools developed by Mike Scott. A one-off charge is made for downloading it. It can be used with any corpus compiled by the individual researcher.

The next set of entries comprises corpus search tools that have a range of corpora built into them.

Corpus.byu.edu. This suite of corpora has been developed by Mark Davies. It is accessible free of charge via the internet and it consists of a set of corpora accessed through its own web interface. The corpora include the Corpus of Contemporary American English, the Corpus of Historical American English, the British National Corpus, and the Corpus of American Soap Operas.

3. <wse1.webcorp.org.uk/home/dia.html>

SketchEngine. This is a set of tools developed by Adam Kilgariff and colleagues accessed via the internet on payment of an annual charge. Users have access to over 300 corpora, including over 50 in English. They can also add their own corpus and access it using the SketchEngine tools. SketchEngine has a number of unique features. As well as generating word lists and concordance lines, SketchEngine shows collocates arranged according to function. For example, it will show the most frequent subjects or objects of a given verb in the chosen corpus.

Skylight. This is a set of tools developed by Gill Francis and Andy Dickinson. It is accessible via the internet and is free of charge. It is designed to be easy to use for teachers and learners. The web interface gives access to 16 corpora in English, including the British National Corpus and the American National Corpus.

The next set of examples comprises specialised corpora.

The British Academic Written English Corpus (BAWE) is a corpus of essays written at UK universities. It can be accessed via SketchEngine or it can be downloaded and used with other corpus tools such as Antconc.

The International Corpus of English (ICE) contains 1 million words each of varieties of English from 20 countries or regions, including Cameroon, East Africa, Great Britain, Kenya, Nigeria, Sierra Leone, South Africa and USA. Each sub-corpus follows the same design parameters (for example, contains the same amount of newspaper text, etc.); thus, the sub-corpora are directly comparable. The ICE is syntactically parsed (see above) and is accessed through its own specialised software.

The International Corpus of Learner English (ICLE) is a corpus of written English from learners with 20 mother tongue backgrounds. It can be purchased on a CD-ROM which also has a built-in concordancer.

The Michigan Corpora. This collection consists of two corpora compiled from academic texts. They are: the Michigan Corpus of Academic Spoken English (MICASE); and the Michigan Corpus of Upper-Level Student Papers (MICUSP). The collection is accessed through its own web interface.

The scope and applications of corpus linguistics

As will be apparent from the discussion above, corpus linguistics is an approach to language investigation which can be applied to almost every aspect of language, including lexis, grammar, semantics and discourse. Many of these applications are illustrated in the chapters in this book.

A major application of corpus linguistics is to the study of the lexis and grammar of languages and geographic varieties of a language. This is amply exemplified in this book by a number of studies of African varieties of English: Isingoma and Meierkord's chapter on the lexis of Ugandan English (Chapter 2.5); Kautzsch's chapter on lexis and morpho syntax in Namibian English (Chapter 2.3); Ozón, FitzGerald,

and Green's chapter on verb use in Cameroon Pidgin English (Chapter 1.6); van Rooy's chapter on unmarked verb forms in various African varieties of English (Chapter 2.1).

The study of language is inextricably entwined with the study of its social, historical and political contexts. History is exemplified in this volume by two papers by Brato on diachronic studies of Ghanaian English, tracing developments in lexis and in the status of Ghanaian English as a variety. The application of corpus linguistics to the study of political discourse is addressed in Antia and Hendricks' paper on texts associated with a South African political party.

A more practical application of corpus linguistics is in the study of learner language. The papers by Esimaje and by Iyabo both demonstrate how useful corpus linguistics can be in this regard. The paper by Nkemleke discusses the direct use of corpus information in the teaching context.

The papers in this collection draw on the wider context of applied corpus studies. Early work on lexis, collocation and phraseology led to the publication of the COBUILD learners' dictionary (Sinclair 1991; Sinclair et al. 1987), and learners' dictionaries, among others, now routinely rely on corpus information to ensure accurate representation of word use. The emphasis on lexis also inspired a lexis-centred 'pattern' approach to grammar (Hunston & Francis 1999), which finds an echo in more recent work on constructions (Goldberg 2006).

With the various computer tools available to corpus analysts, the comparison of language with language, language variety with language variety, is now an increasingly manageable task. As such, descriptions of varieties of English, for specific purposes, in terms of lexis, grammar, syntax and semantics have become easily accessible. It is in view of this that Baker (2004) notes that lexical corpus linguistics exonerates lexical investigation from the traditional criticism that it offers no significant solution to linguistic problems since it seems to focus solely on frequency. Work on language variation in this volume and elsewhere locates differences in lexical frequency with phraseology, grammar and culture.

The study of register variation has benefited considerably from corpus studies since Biber's early ground-breaking work. In particular, the language associated with different academic registers and disciplines (Biber 2006; Hyland 2012; Gray 2015) has usefully been documented. Hyland (2012), for example, studies personal and professional identities in various academic genres ranging from traditional academic papers to the more recent phenomenon of academic websites. Gray (2015) uses the multi-dimensional analysis pioneered by Biber to propose that academic discipline alone is insufficient to account for register variation in academic research articles; differences in research paradigms, such as qualitative and quantitative, also play an important part.

As noted above, another fertile application for corpus linguists is the study of diachronic changes in language. The effects of temporal changes on language vocabulary and meaning are well known. However, investigating these variations to track specific changes, in order to inform the designs of subsequent language curricula, had been an age-long challenge until the advent of corpus and computational linguistics which enabled the systematic handling of substantial texts in quite a limited time frame. An example is the work of Koteyko (2007), who applied a corpus approach to the study of diachronic changes in the use of lexical items by particular discourse communities. Her immediate focus was on changes in the use of English loan words in a Russian pro-communist community. Another example is that of Westin and Geisler (2002) who used a corpora of twentieth-century British newspaper editorials (published between 1900 and 1993 in three newspapers: *The Guardian*, *The Daily Telegraph*, and *The Times*) to investigate diachronic changes in grammatical features in English based on Biber's multi-dimensional model. The study concludes that clear linguistic and stylistic differences were apparent in the different corpora examined and this shows that "by using different periodization strategies, we can stress different aspects of language change" (Westin & Geisler 2002).

The field of discourse analysis has benefited no less than others from the incursion of corpus linguistics. This is demonstrated, for example, by Gabrielatos and Baker (2008) who used a corpus approach to analyze discourse patterns and diachronic changes in the UK press, from 1996–2005, by examining the discursive constructions of refugees and asylum-seekers in UK. This study, which shows the extent to which a corpus-based methodological stance can enhance critical discourse analyses, finds that the descriptive terms associated with the discourse of refugees and asylum seekers in the UK press are confusing, ambiguous and that often the terms are conflated. They do, however, recommend further analysis, a multi-dimensional critical discourse analysis (CDA) to establish why certain linguistic patterns are found or not found.

Also, in literary stylistics, the value of corpus linguistics is very well catered for as the following study testifies. Starcke (2006: 87–103) examines Jane Austen's *Persuasion* and demonstrates that phraseological units in language are the carriers of meaning. The paper finds that the explicit and implicit markers of time in the novel are collocates of the novel's most frequent phrases and that they not only relate to the chronological sequence of events in the novel but also contribute to the novel's melancholy atmosphere and thematic focus. Mahlberg (2013) and Mahlberg et al. (2013, 2016) investigate various aspects of Dickens' style based on a freely-available corpus of all his novels. They note, for example, that many of Dickens' famously memorable characters are distinguished by phrases denoting physical activities such as having hands in pockets.

Corpus linguistics has a great deal to offer language teachers. One application is the study of the registers that learners are likely to use: the extensive work in English for Academic Purposes is an instance of this. Another is the study of learner corpora, used to identify areas where language instruction is most needed and most useful, as noted above. Increasingly, learners are encouraged to explore corpora themselves: Charles (2017) gives an example of international university students working with corpora specific to their subject areas in self-directed learning. Granger and Tribble's (1998) paper is a much earlier example of how concordance lines can be used with language learners to encourage 'noticing' of language features. Language learners often find it fascinating to explore a corpus (e.g. Davies 2004), and there is some research on how best to exploit corpora with learners (e.g. Sripicharn 2004). Difficulties remain, however; for example, it is not often possible to equip classrooms with computers, and many available corpora contain language unsuitable for younger learners. Teachers, too, may be unfamiliar with corpus resources, and corpus investigation may not fit neatly into exam-driven syllabuses. With these caveats in mind, it is nonetheless likely that the use of corpora by learners will gradually increase.

Conclusion

Corpus linguistics is a method of linguistic research which relies on real or naturally-occurring language or language-in-use as data. Its main object is the corpus which is itself a collection of actual utterances or writings of people and thus, it is in fact actual human behaviour. This highlights the first merit of corpus linguistics: having as data actual language in use. Beyond this, its second strength is the ability to 'see' far beyond the capability of ordinary human 'eyes' because of computer technology. This combination of software and corpora enables a spectrum of observation of linguistic phenomena that was previously impossible. The third strength is the quantification of data and the application of statistical measures which make possible the systematization of research handling, thereby increasing the reliability of findings. For these reasons the popularity of this approach continues to grow.

It is therefore important that knowledge about corpus linguistics and its approaches is shared or disseminated to those linguistic communities not yet acquainted with it to enable linguists make informed choices of approach for different research enterprises. This is the overall objective of this chapter – to draw the attention of African linguists, in particular, to the corpus approach and the force of its relevance in language studies today.

References

- Adolphs, Svenja & Carter, Ronald. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. London: Routledge. <https://doi.org/10.4324/9780203526149>
- Aijmer, Karin & Altenberg, Bengt (eds). 1991. *English Corpus Linguistics*. London: Longman.
- Aston, Guy, Bernardini, Silvia & Stewart, Dominic. 2004. Introduction: Ten years of TALC. In *Corpora and Language Learners* [Studies in Corpus Linguistics 17], Guy Aston, Silvia Bernardini & Dominic Stewart (eds), 1–18. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.17>
- Atkins, Sarah, Roberts, Celia, Hawthorne, Kamila & Greenhaigh, Trisha. 2016. Simulated consultations: A sociolinguistic perspective. *BMC Medical Education* 16(16): 1–9. <https://doi.org/10.1186/s12909-016-0535-2>
- Baker, Paul. 2004. Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics* 32(4): 346–359. <https://doi.org/10.1177/0075424204269894>
- Baker, Paul. 2008. *Using Corpora in Discourse Analysis*. London: Continuum.
- Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511621024>
- Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–57. <https://doi.org/10.1093/lilc/8.4.243>
- Biber, Douglas. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.23>
- Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275–311. <https://doi.org/10.1075/ijcl.14.3.08bib>
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Broughton, Geoffrey, Brumfit, Christopher, Flavell, Roger, Hill, Peter & Pincas, Anita. 2003. *Teaching English as a Foreign Language*, 2nd edn. London: Routledge.
- Carter, Elisabeth. 2009. Policing talk: An investigation into the interaction of the officer and the suspect in the police interview. *International Journal of Speech Language and the Law* 16(1): 165–168. <https://doi.org/10.1558/ijsl.v16i1.165>
- Charles, Maggie. 2017. Do-it-yourself corpora in the EAP classroom: Views of students and teachers. In *Faces of English Education: Students, Teachers, and Pedagogy*, Lillian Wong & Kenneth Hyland (eds), 107–123. London: Routledge.
- Davies, Mark. 2004. Student use of large, annotated corpora to analyze syntactic variation. In *Corpora and Language Learners* [Studies in Corpus Linguistics 17], Guy Aston, Silvia Bernardini & Dominic Stewart (eds), 259–269. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.17.19dav>
- Durrant, Philip. 2014. Discipline and level specificity in university students' written vocabulary. *Applied Linguistics* 35(3): 328–356. <https://doi.org/10.1093/applin/amto16>
- Esimaje, Alexandra. 2012. A Lexical Study of the Sermons of Pastor Chris Oyakhilome. PhD dissertation, University of Maiduguri.
- Esimaje, Alexandra. 2016. The Corpus of Nigeria and Cameroon English Learner Language (Conacell).

- Francis, W. Nelson. 1982. Problems of assembling and computerizing large corpora. In *Computer Corpora in English Language Research*, Stig Johansson (ed.). Bergen: Norwegian Computing Centre for the Humanities.
- Francis, W. Nelson & Kučera, Henry. 1964. *The Brown Corpus*. Providence RD: Brown University.
- Gabrielatos, Costas & Baker, Paul. 2008. Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005. *Journal of English Linguistics* 36(5): 5–38. <https://doi.org/10.1177/0075424207311247>
- Garside, Roger & Smith, Nicholas. 1997. A hybrid grammatical tagger: CLAWS4. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Roger Garside, Geoffrey Leech & Anthony McEnery (eds), 102–121. London: Longman.
- Garside, Roger, Leech, Geoffrey & McEnery, Anthony (eds). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.
- Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: OUP.
- Granger, Sylviane (ed.). 1998. *Learner English on Computer*. London: Longman.
- Granger, Sylviane & Tribble, Christopher. 1998. Learner corpus data in the classroom: Form-focused instruction and data-driven learning. In *Learner English on Computer*, Sylviane Granger (ed.), 199–209. London: Longman.
- Gray, Bethany. 2015. *Linguistic Variation in Research Articles: When Discipline Tells only Part of the Story* [Studies in Corpus Linguistics 71]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.71>
- Gries, Stefan T. 2006. Exploring variability within and between corpora: Some methodological considerations. *Corpora* 1(2): 109–151. <https://doi.org/10.3366/cor.2006.1.2.109>
- Gries, Stefan T. 2010. Useful statistics for corpus linguistics. In *A Mosaic of Corpus Linguistics: Selected Approaches*, Aquilino Sanchez & Moises Almela (eds), 269–291. Frankfurt: Peter Lang.
- Halliday, Michael. 1985/1994. *An Introduction to Functional Grammar*, 1st/2nd edn. London: Edward Arnold.
- Hardie, Andrew. 2017. Exploratory analysis of word frequencies across corpus texts: Towards a critical contrast of approaches. Paper given at Corpus Linguistics 17 conference, University of Birmingham, July 2017.
- Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139524773>
- Hunston, Susan & Francis, Gill. 1999. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.4>
- Hyland, Kenneth. 2012. *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: CUP.
- Johnston, Trevor & Schembri, Adam. 2006. Issues in the creation of a digital archive of a signed language. In *Sustainable Data from Digital Fieldwork*, Linda Barwick & Nicholas Thieberger (eds), 7–16. Sydney: University of Sydney Press.
- Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.
- Koteyko, Nelya. 2007. A diachronic approach to meaning: English loanwords in Russian opposition discourse. *Corpora* 2(1): 65–95. <https://doi.org/10.3366/cor.2007.2.1.65>

- Kruger, Alet, Wallmach, Kim & Munday, Jeremy (eds). 2011. *Corpus-based Translation Studies: Research and Applications*. London: Continuum.
- Leech, Geoffrey. 1997. Introducing corpus annotation. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Roger Garside, Geoffrey Leech & Anthony McEnery (eds), 1–19. London: Longman.
- Leivada, Evelina, Papadopoulou, Elena & Pavlou, Natalia. 2017. Functionally equivalent variants in a non-standard variety and their implications for Universal Grammar: A spontaneous speech corpus. *Frontiers of Psychology* 8: 1260. <https://doi.org/10.3389/fpsyg.2017.01260>
- Leńko-Szymańska, Agnieszka. 2004. Demonstratives as anaphora markers in advanced learners English. In *Corpora and Language Learners* [Studies in Corpus Linguistics 17], Guy Aston, Silvia Bernardini & Dominic Stewart (eds), 89–107. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.17.07len>
- Mahlberg, Michaela. 2005. *English General Nouns: A Corpus Theoretical Approach* [Studies in Corpus Linguistics 20]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.20>
- Mahlberg, Michaela. 2013. *Corpus Stylistics and Dickens's Fiction*. London: Routledge.
- Mahlberg, Michaela, Smith, Catherine & Preston, Simon. 2013. Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels. *International Journal of Corpus Linguistics* 18(1): 35–56. <https://doi.org/10.1075/ijcl.18.1.05mah>
- Mahlberg, Michaela, Stockwell, Peter, de Jooode, Johan, Smith, Catherine & O'Donnell, Matthew. 2016. CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* 11(3): 433–463. <https://doi.org/10.3366/cor.2016.0102>
- Mauranen, Anna. 2012. *Exploring ELF: Academic English Shaped by Non-native Speakers*. Cambridge: CUP.
- McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics*. Cambridge: CUP. <https://doi.org/10.1093/oxfordhb/9780199276349.013.0024>
- McEnery, Tony, Xiao, Richard & Tono, Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York NY: Routledge.
- McEnery, Tony & Wilson, Andrew. 1996. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Meyer, Charles. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511606311>
- Murison-Bowie, Simon. 1996. Linguistic corpora and language teaching. *Annual Review of Applied Linguistics* 16: 182–199. <https://doi.org/10.1017/S0267190500001501>
- Nelson, Michael. 2000. A Corpus Based Study of Business English and Business English Teaching Materials. PhD dissertation, University of Manchester.
- Partington, Alan, Duguid, Alison & Taylor, Charlotte. 2013. *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies (CADS)* [Studies in Corpus Linguistics 55]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.55>
- Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549. <https://doi.org/10.1075/ijcl.13.4.06ray>
- Scott, Michael. 1996. *Wordsmith Tools*. Oxford: OUP.
- Scott, Michael & Tribble, Christopher. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education* [Studies in Corpus Linguistics 22]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.22>
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

- Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
- Sinclair, John et al. 1987. *Collins Cobuild English Language Dictionary*. London: HarperCollins.
- Sripicharn, Passapong. 2004. Examining native speakers' and learners' investigation of the same concordance data and its implications for classroom concordancing with EFL learners. In *Corpora and Language Learners* [Studies in Corpus Linguistics 17], Guy Aston, Silvia Bernadini & Dominic Stewart (eds), 233–245. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.17.17sri>
- Starcke, Bettina. 2006. The phraseology of Jane Austen's *Persuasion*: Phraseological units as carriers of meaning. *ICAME Journal* 30: 87–102.
- Su, Hang. 2015. Judgement and Adjective Complementation Patterns in Biographical Discourse: A Corpus Study. PhD dissertation, University of Birmingham.
- Taylor, Charlotte. 2008. What is Corpus Linguistics? What the data says. *ICAME Journal* 32: 179–200.
- Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13. <https://doi.org/10.1075/ijcl.10.1.01teu>
- Thompson, Geoffrey & Hunston, Susan (eds). 2006. *System and Corpus: Exploring Connections*. London: Equinox.
- Thompson, Paul, Hunston, Susan, Murakami, Akira & Vajn, Dominik. 2017. Multi-dimensional analysis, text constellations, and interdisciplinary discourse. *International Journal of Corpus Linguistics* 22(2): 153–186. <https://doi.org/10.1075/ijcl.22.2.01tho>
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.6>
- Truscott, John. 1996. The case against grammar correction in L2 writing classes. *Language Learning* 46(2): 327–369. <https://doi.org/10.1111/j.1467-1770.1996.tb01238.x>
- Westin, Ingrid & Geisler, Christer. 2002. A multidimensional study of diachronic variation in British newspaper editorials. *ICAME Journal* 26: 133–152.

Appendix. Concordance lines for *obvious*

to the next. As well as the	obvious	advantages to both your health and
And while Boyzone, to state the	obvious,	ain't U2, Keating is moving from
e State Department acknowledge the	obvious,	and officially pronounce Iraq to
us with some new examples of the	obvious.	And as Flexner made clear almost
b pages. The methodical, sometimes	obvious	approach usefully opens up the
and people. <p>Yet, despite its	obvious	benefits, the EEA has been a long
reject the view, considered simply	obvious	by most scholars, that the
of herself must be female. The	obvious	candidate for antecedent, Rachel,
effrey or Horner – Brougham is the	obvious	candidate – would be guilty of a
long been considered as the next	obvious	candidates to float. Both say they
or their troubles. Witches were an	obvious	choice. Accused witches were most
re need of a basket. He seemed the	obvious	choice, for example, after the
ou can see that although it has no	obvious	colour scheme, in fact the shades
xpert was needed to appreciate the	obvious	danger of sailing with open bow
and religious leaders. There are	obvious	dangers of a right to die becoming

But all this indulgence has had an
than just a clothes show. The most
he deciding set for displaying his
said the paper revealed the
thin the buying centre. This is of
and past connections. The most
he springboard of this victory was
prosecutors and judges seems most
50 and 75 a week. <p>There is an
tonishing capacity for turning the
to us," he says. 'Smuggling is an
his impression more specific. The
are in direct contradiction to its
<p> I am warned against posing
tourism to that island for the
Late last autumn without any
to apply to life insurance for the
in the Champions League and, for
is certainly no threat to her most
<p> Nevertheless, there are some
in the filming, either. It was
nts us to meet, and he has made it
rnest with whispers in bed. It was
it had been stolen. It is far from
from itself, because it was
and challenging programs. It is
who claimed 12-107. That made it
the same view to make it readily
real and complex numbers was
III <f> a year earlier. It was
on Saturday and made his feelings
woman writer, but it becomes
few surprises. The first and most
they even fail to mention the most
for contemporary women, this isn't

effect. Quick to sense their effect is to make the play's frustration. The match was 'HIPC has failed to deliver interest to salespeople as they interpretation to those on the Mike Catt was making the latest Mr. Wintery, the prosecutor, need for rent assistance to on its head. Schumacher's 37-point opportunity open to the newly practical implication of primacy-provisions and to the views of questions. 'Someone asked him what reason: hard currency is in short reason there was a noticeable reason that the actual cash value reasons, that is the game everyone rivals, Missy Elliot and Mary J, strategies that have worked well that they were going to that he thinks we are already half-those things were not right.The that using a parking bay without a that its various peoples had no that talent and dedication of the that at some point Australia would that for all its rich variety, to the millions of potential to all that she found working with to all when he was substituted by to her that she cannot write and was the extent to which it was way of helping poor parents-We are not as likely as our

Corpus-based research on English in Africa

A practical introduction

Robert Fuchs, Bertus van Rooy and Ulrike Gut

University of Hamburg / North-West University / University of Münster

This chapter provides linguists and students not yet familiar with corpus-based research on varieties of English in Africa with a practical introduction to the field. After explaining the rationale and aims of corpus-based research on varieties of English (in Africa), we introduce methods, tools and resources commonly used and applied in the field in order to provide readers with a point of entry into the field. Most of the corpora and software that are introduced can be obtained free of charge. The software is introduced in a practical way to permit readers to use it in their own research. The application and value of corpus linguistics are exemplified with three case studies. These examples are based in part on previous research, retracing the methodological steps, but are also expanded with more data from across Africa. Case study 1 shows how corpora allow researchers to investigate lexical differences between African varieties of English, arguably an area that is amenable to scholarly inquiry with relatively limited methodological means. Case study 2 considers a grammatical phenomenon, the present perfect in African Englishes, and demonstrates how a corpus tagged for parts of speech permits syntactic analyses. Case study 3 illustrates the analysis of a phonological corpus with an investigation of the optional deletion of the phoneme /h/ in words such as *house* in Nigerian English. The chapter concludes with recommendations for further reading, allowing readers to explore selected topics in more depth according to their interests.

1. Introduction

Corpora are large, representative and digital collections of texts and/or speech that provide an empirical basis on which verifiable and replicable research on language can be based. A corpus is for the linguist what a survey of the population residing in a certain place is to the economist or political scientist. When scholars from these fields want to determine the income distribution of a population, or their political preferences, they might opt to conduct a survey. Instead of requesting information from all residents of a country, which would be wasteful for most purposes, they

will restrict themselves to a representative (e.g. in gender, ethnicity, linguistic and educational background) sample of the population. Trends that become visible from an analysis of this sample can, with a certain margin of error, be generalised to the whole population.

In the same way, a linguist interested in studying a language or variety cannot make a record of all the words spoken and written or heard and read in the language over any period of time. A representative sample of a language will rely on the inclusion of speakers with a representative range of personal characteristics such as gender and education. Moreover, to be representative in this holistic sense, a corpus also needs to cover the full range of situations in which a particular language is used, and therefore it should include a range of text types, or registers, such as private conversations, speeches, newspaper articles and legal writing. In practice, representativeness in corpus compilation is often more strictly enforced for registers than for the personal characteristics of speakers/writers. A corpus may also be intended to be representative of private conversations in a specific city or of a specific genre such as newspaper articles.

Corpora are usually made available publicly, at least to other researchers, provided that permission and copyright clearance are taken care of. This enables the replication of studies and thereby increases the reliability of research. It also allows for faster progress to be made in research on particular languages or varieties, as many different researchers can potentially contribute to knowledge by analysing different aspects of the same corpus. More generally, corpora also allow researchers to make evidence-based generalisations and judgements on a language or variety, compared to anecdotal and chance observations, which might be influenced by the (often unconscious) bias of the researcher in question and are difficult to falsify. In this sense, corpora that are publicly available support the scientific method. This makes it possible for studies to be replicable (another researcher can repeat the analysis in order to check its accuracy), falsifiable (hypotheses and research results can be tested on the data and possibly shown not to be supported by the data), and objective (all steps in the analysis can be retraced and repeated by other researchers; Leech 1992).

Given the spread of English around the globe, more resources have been spent on corpus linguistics research on this language than on any other. It is little wonder then that corpus linguistics research on English is not new to Africa. When the International Corpus of English (ICE) project (Greenbaum 1996) was launched with the aim of compiling similar corpora of all national varieties of English where this language plays an official role, the challenge of extending the ICE corpus design to non-native varieties of English in Africa was taken up by Schmied (1990). Since then, a number of other ICE corpora have been collected or are in the process of being collected in Africa (e.g. ICE Nigeria, see below, and ICE Uganda, see

Chapter 2.6, this volume). ICE corpora are particularly valuable for research into the features of a specific variety and the comparison of features across varieties of English, since the different ICE corpora have the same, or very similar, designs. Beyond ICE corpora, a number of other corpora of African Englishes have also been compiled for purposes of linguistic research, such as the spoken Xhosa-English corpus (de Klerk 2006), the Black and Coloured South African English in contact corpus (Meierkord et al. 2009) and all the corpora presented in this volume. The present chapter examines a number of ways in which corpus-linguistic research on African Englishes can be conducted.

A second line of inquiry that is of much value for applied linguistics is the investigation of student or academic writing by means of corpora. Nkemleke (2011) considers a range of features of academic writing in Cameroonian English, while a number of studies have been conducted with the Tswana Learner English Corpus, which forms part of the International Corpus of Learner English (Granger et al. 2009). The findings of such studies can serve to provide input to classroom practice, curriculum design and ultimately contribute to the improvement of academic writing, both by students and researchers, in English. These studies often require analyses of a more textual nature, looking at the realisation of cohesion, for instance, or techniques for identifying learner errors, which are not discussed in this chapter (but see Chapter 2.7, this volume). However, analyses of academic writing can sometimes benefit from the more linguistic approach that is profiled in this chapter, to understand the ways in which linguistic features are used differently by learners at different stages of development, as exemplified by Van Rooy and Kruger (2016) for instance, with reference to the progressive aspect use in corpora that represent learners at different stages of development.

In Section 2, tools and resources for corpus-based research on African Englishes will be described. In Section 3, the use of corpora for the study of diverse aspects of African Englishes is demonstrated with three case studies focusing on lexical features, the present perfect and the phonological process of /h/-deletion, respectively. The chapter concludes with suggestions for further reading in Section 4.

2. Tools and resources for corpus-based research on African Englishes

Corpus linguistics can be done only if there are corpora available for analysis, and if there are tools that can assist the researcher in the analysis. This section examines a selection of the available corpora for African Englishes, alongside the guidelines that the compilers of a new corpus should bear in mind. This is followed by a discussion of different kinds of tools available for a range of different kinds of analyses of corpus data.

2.1 Corpora

When English corpus linguistics received a major growth injection with the launch of the International Corpus of English, the East African corpus was one of the first to be completed. All ICE corpora (i.e. subcorpora or ‘components’ of the International Corpus of English), as set out below, have a relatively fixed structure, which has two important virtues: they represent a particular variety in a balanced way, and they make comparative work across different varieties of English possible. Although the original aim (Greenbaum 1996) was for fully identical corpora, Schmied (1990) points out that certain text types are not common or widely produced in particular societies, and therefore motivated deviation from the original corpus design template has to be tolerated sometimes. Nevertheless, corpus compilers generally tried to remain close to the original blueprint. For each variety of English, Greenbaum’s (1996) original design recommended that a corpus consisting of one million words (400,000 of written English and 600,000 words of spoken English) be collected, comprising different text categories such as private conversations, broadcast interviews, private letters, novels and academic writing (see Table 1).

Table 1. Text categories and number of words in the ICE corpora

Written part	Required total no. of words	Spoken part	Required total no. of words
Academic text: humanities	20,000	Broadcast discussions	40,000
Academic text: natural sciences	20,000	Broadcast interviews	20,000
Academic text: social sciences	20,000	Broadcast news	40,000
Academic text: technical	20,000	Broadcast talks	40,000
Administrative/instructive texts	20,000	Business transactions	20,000
Business letters	30,000	Class lessons	20,000
Editorials	20,000	Commentaries	40,000
Essays	20,000	Conversations	180,000
Exams	20,000	Business transactions	20,000
Novels	40,000	Demonstrations	20,000
Popular texts: humanities	20,000	Legal cross-examinations	20,000
Popular texts: natural sciences	20,000	Legal presentations	20,000
Popular texts: social sciences	20,000	Non-broadcast talks	20,000
Popular texts: technology	20,000	Parliamentary debates	20,000
Press reports	40,000	Phone calls	20,000
Skills, hobbies/instructive texts	20,000	Unscripted speeches	60,000
Social letters	30,000		
Total	400,000	Total	600,000

Two ICE corpora representing African Englishes have been completed so far: ICE East Africa (Hudson-Ettle & Schmied 1999), which contains Kenyan and Tanzanian English, and ICE Nigeria (Wunder et al. 2010). Two further corpora, ICE Ghana and ICE Uganda (see also Chapter 2.6) are under construction, alongside ICE South Africa (Jeffery 2003). The latter contains exclusively native-speaker data (mainly White, but some native speakers from other population groups as well), and is therefore more comparable to other colonial varieties such as those represented by ICE New Zealand and ICE Australia. In addition, ICE Great Britain may be used to compare colonial varieties to the present-day version of their input variety (although it is important to take into account that the actual input came from a socially and regionally diverse group of speakers and that British English has also evolved in the meantime). The speakers in these corpora nearly all have a university degree or are university students with the possible exception of some politicians and actors. Their professions range from university student to university lecturers, reverends, broadcasters, journalists, politicians, actors, medical doctors, barristers, and lawyers and therefore represent the educated usage of English in the respective country.

Most ICE corpora can be obtained for free for academic research via the internet and include CLAWS part-of-speech (POS) tagging and semantic annotation (at the time of writing, the website is being moved to the University Zurich, where an online concordancing interface will be made available).¹ ICE Nigeria is freely available from <http://pacx.sourceforge.net/projects/ice-nigeria/>. ICE Great Britain is available for a fee from University College London, and includes manually verified POS-tagging and syntactic annotation that can be accessed via specialised computer software that is shipped together with the corpus.²

In contrast to the other corpora, ICE Nigeria is a phonological corpus. It was collected with the help of the Pacx software, which is based on the Eclipse platform and extends it by the addition of several plug-ins: the XML editor Vex, the image viewer QuickImage and the client for the version control system Subversive.³ Furthermore, it builds on the software ELAN, which allows time-aligned transcriptions of video and audio data. Figure 1 illustrates the annotation of a video recording that was carried out with the software ELAN. The video is displayed in the top left-hand corner. Beneath the speech wave form, the annotation is represented on three different levels, referred to as 'tiers'. Text-to-tone alignment links the annotation to the raw data, which means that the boundaries of all annotated

1. <<http://www.es.uzh.ch/en/Subsites/Projects/dbank/iceonline.html>>

2. <<https://www.ucl.ac.uk/english-usage/projects/ice-gb/>>

3. <<http://www.eclipse.org>>

elements are defined by time stamps. Information about the exact beginning and end of each element is available in the corresponding XML file that is created by the program. The time-aligned annotation illustrated in Figure 1 provides direct access from each annotated element to the raw data, i.e. the original audio or video file. By clicking on any annotated element, the corresponding part of the recording will be played back by ELAN.

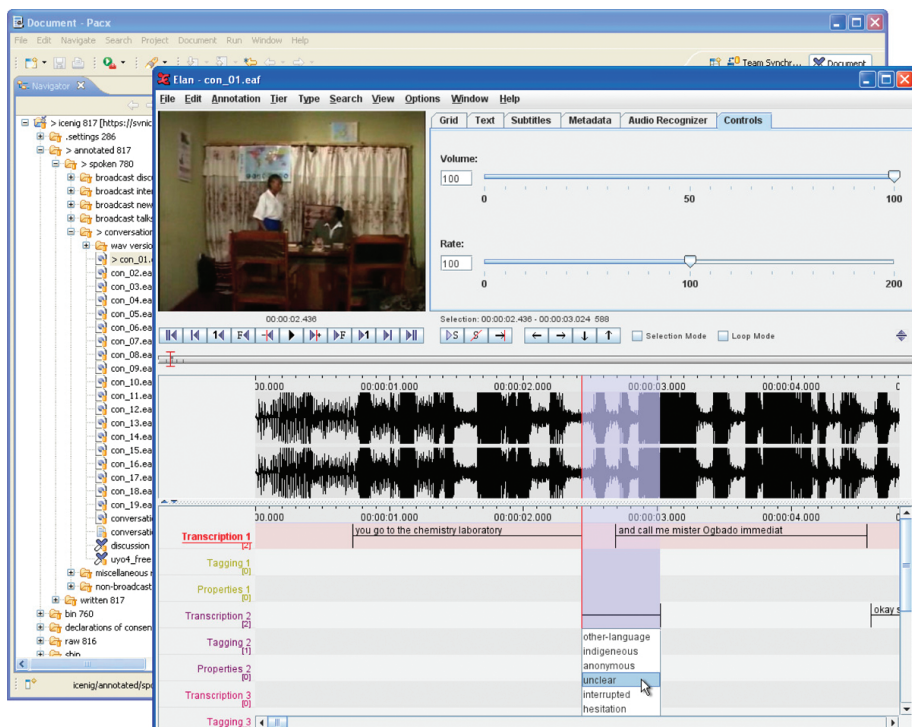


Figure 1. Annotation of spoken language data in Pacx (reprinted from Wunder et al. 2010)

Beyond the ICE corpora, a number of researchers opted for the collection of corpora that follow a different structure. Constraints of resources sometimes mean that a smaller corpus is collected, but these corpora are still very valuable as they provide data on particular varieties about which very little is still known. This is illustrated by the work of Marungudzi (2017), who collected a 400,000 word balanced corpus similar to ICE, representing Zimbabwean English, an under-researched variety of English in Africa (see also Chapter 1.6 this volume for a corpus of Cameroon Pidgin English that is partly modelled on the ICE format). While the ICE template is a very good standard reference point, particularly for studies aimed at the description of educated national varieties, corpus researchers can make valuable contributions

to scholarship with other corpus designs too. ICE is, by contemporary standards, relatively small, especially if one is interested in rare syntactic or lexical phenomena or a particular register of language. Thus, some researchers chose to collect corpora that focus on a smaller selection of registers, in order to gain more insight into such a subpart of the language. De Klerk (2006), for example, compiled a corpus of spoken English conversation of 500,000 words from native speakers of isiXhosa (an indigenous language spoken in South Africa), in order to enable a better look at the features of this register than would be afforded by the 180,000 words in an ICE corpus. Another example is Nkemleke's Corpus of Cameroon English, which focuses only on written registers, and comprises 820,000 words (Nkemleke 2012; Chapter 2.8 this volume).

The Global Web-Based English Corpus (GloWbE, pronounced 'globe'; Davies & Fuchs 2015) follows a different philosophy than the corpora discussed so far. Consisting of 1.9 billion words gathered from the internet from 20 English-speaking countries around the world, it is much larger but also less structured than, for example, the ICE corpora. GloWbE is not split into several finely distinguished registers. Instead, it consists of a section made up of weblogs as well as a general section, with the language in the former being more informal overall than the latter. Unlike most ICE corpora, GloWbE does not provide any biographical information on the writers of the texts in the corpus. The African varieties GloWbE includes are South African, Nigerian, Ghanaian, Kenyan and Tanzanian English and the corpus can be accessed for free for academic research.⁴ The search interface accepts queries consisting of one or several words and/or POS tags and provides the absolute and relative number of hits as well as a KWIC display of the search result. In order to keep the computational load of queries on the server at a reasonable level, the complexity of such queries is fairly limited compared to the power of regular expressions involving POS tags that can be realised in concordancing software. A full text version (with 5% of the text removed in order to not infringe copyright) can be downloaded for a fee, so that researchers with the required technical expertise and computational power to process such amounts of data can realise more complex queries. The NOW corpus extends the concept of GloWbE with a diachronic perspective. It contains material from online newspapers and magazines from 2010 to the present, currently 4.8 billion words, with more data being added regularly. The corpus can be searched online or downloaded for a fee under the same conditions as GloWbE.⁵

4. <<http://corpus.byu.edu/glowbe/>>

5. <<http://corpus.byu.edu/now/>>

A fairly recent new development in corpus research in Africa is the compilation of historical corpora. To date, work has been limited to native White and non-native Black South African English (Rossouw & Van Rooy 2012; Van Rooy & Piotrowska 2015) and Ghanaian English (Brato 2018; Chapter 1.5 this volume). Historical corpus compilation is constrained by available sources, which often present a very limited snapshot of the language. This is even more the case where published documents produced in African countries during the period of colonial rule were subject to editorial control of native speakers at missionary publishing houses (De Kock 1996). Despite these challenges, independent newspapers set up and controlled by indigenous Africans sometimes do trace their ancestry back to the 19th and early 20th century, such as the newspapers *Imvo Zabantsunda* ('The native opinion of South Africa' 1884) and *Ilanga lase Natal* ('The Natal Sun' 1903), which were bilingual papers carrying material in English alongside Xhosa and Zulu respectively. Such newspaper material, together with a diverse range of handwritten letters, representing both private and business correspondence, offer potential corpus data from printed and non-printed sources from which to construct historical corpora that are not filtered through the lens of native-speaker influence. From the middle of the 20th century, as African states became independent and colonial administrators and missionaries returned to their countries of origin, native-speaker linguistic control over printed material has been receding. Thus, materials from newspapers, fiction, non-fiction prose and letters are available from archives and can be converted into machine-readable corpora, even if a degree of opportunism has to be associated with the process and neat corpus structures cannot always be adhered to.

Table 2 lists a selection of corpora of African Englishes. For information on new corpora released after this volume goes to print, readers may consult the Corpus Resource Database (CoRD).⁶

A final development to take note of in corpus compilation in Africa is a set of specialised corpora, compiled with very particular goals in mind. One type already alluded to is student writing, where corpora are compiled to represent student writing at various levels of mastery of English. The analysis of such corpora mainly serves applied linguistic goals, as evidenced by Adeyemi (2017) or Van Rooy and Kruger (2016), but can also be used as material for linguistic analysis, as was done by Van Rooy (2006). Corpora of academic writing, such as the one compiled by Nkemleke (2011), serve as complement to the student writing corpora, to examine how African academic writers express themselves in English. A very recent type of specialised corpus compiled in South Africa is a parallel corpus of edited and unedited texts. Since editorial intervention is suspected of influencing the published written shape of texts in typical non-native English corpora, this parallel corpus

6. <<http://www.helsinki.fi/varieng/CoRD/index.html>>

enables thorough scientific investigation of exactly how far the intervention of editors affects the data, as shown by Kruger and Van Rooy (2017).

Table 2. Corpora of African Englishes and English in Africa

Name	Written/ Spoken	Size (words)	Availability
International Corpus of English (ICE): East Africa, Nigeria (Ghana, Uganda, South Africa being compiled)	W & S	1 M. per component	Free
Unofficial ICE Zimbabwe ⁷	W & S	400,000	Available from compiler (Van Rooy)
Spoken Xhosa-English corpus	S	550,000	Available from compiler (De Klerk)
Corpus of Cameroon English	W	820,000	Available from compiler (Nkemleke)
Global Web-Based English Corpus (GloWbE): South Africa, Nigeria, Ghana, Kenya, Tanzania	W	Total 1.3 B., African varieties 203 M.	Online access free, full text for a fee
News On the Web (NOW): South Africa, Nigeria, Ghana, Kenya, Tanzania	W	Currently total 5.2 B., continuously updated	Online access free, full text for a fee
Historical native White and Black South African English	W	ENL: 500,000 ESL: 200,000	Available from compilers (Van Rooy)
Historical Corpus of English in Ghana (HiCE Ghana)	W	600,000	Currently not available, but future release being considered
Tswana Learner English Corpus	W	200,000	Available as part of ICLE version 2
Nigerian Learner English Corpus	W	200,000	Available from the compilers (Adeyemi and Van Rooy)
Corpus of Nigeria and Cameroon English Learner Language (Conacell)	W	450,000	Available from compiler (Esimaje)
Parallel corpus of edited and unedited South African English	W	3 million	Not available due to confidentiality agreements

7. The corpus has not been included officially in the ICE-project, because it represents only 40% of the size of the corpus. Sadly, Dr. Thadeus Marungudzi, the principal compiler of the corpus, passed away in January 2018, after completing his PhD, so plans that were made to enlarge the corpus are currently on hold.

2.2 Tools

After presenting a range of available corpora of African Englishes in the previous section, this section discusses corpus compilation and analysis. While many researchers make use of such existing corpora, it is sometimes necessary to compile a new corpus. A number of decisions should be made beforehand, on the basis of which texts should be collected and digitised for inclusion in the corpus. One very typical approach to such questions is the design of the ICE corpora, reviewed in the previous section. The compilers need to decide on the variety of English to be represented in the corpus: which geographical area (e.g. country, region, etc.), which speakers (e.g. educational, socio-economic, and/or age), and which registers (spoken/written, and further divisions within these). Once these basic decisions have been taken, a compromise between feasibility and representative sampling of texts, and the size of individual text samples (full texts or a limit of, for example, 2,000 words per sample) must be found before individual text samples are collected. Spoken texts are collected by recording speech events or using pre-recorded speech, while written text is collected from libraries, archives, the internet or from speakers directly. In all cases, issues of informed consent, copyright and privacy must be borne in mind. These original texts must be digitised, by transcribing, retyping, scanning or other suitable means, into text files, and then checked for accuracy. While most corpora consist of texts (either derived from written documents or transcriptions of speech), recently more phonological corpora, which include audio data and often annotations of auditory events, have become available (see Gut & Voormann 2014 for an introduction).

When researchers make use of an existing corpus, they often use dedicated software to access this data. Corpora, including most of those mentioned in this chapter, generally consist in their basic form of a set of text files that can be viewed with any general text editor. A single text file often, though not always, consists of material from a single source, e.g. a single conversation or newspaper article. Even a corpus comprising 1 million words may then consist of several hundred files, and specialised software is needed to access the corpus for linguistic study. Such software, called a concordancer, allows researchers to search the corpus for a single word or a phrase or a specific tag. The search result is then displayed in context (see Figure 2), hence the name ‘keyword in context (KWIC)’. A single line from the search result is also called a ‘concordance’. The most widely used concordancers are AntConc, which is available free of charge, and WordSmith Tools, which comes at a fee.^{8,9}

8. <<http://www.laurenceanthony.net/software/antconc/>>

9. <<http://www.lexically.net/wordsmith/>>

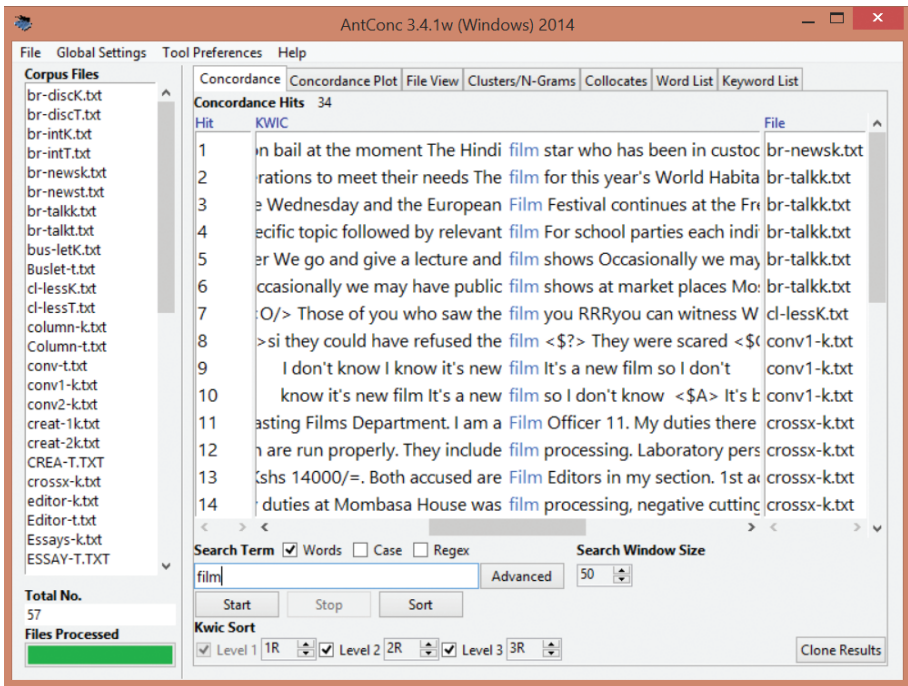


Figure 2. Keyword in Context display of the search result for the word ‘film’ in ICE East Africa

Often it will be useful to export the concordances in order to annotate them or manually remove unwanted occurrences. For example, the English progressive aspect is mainly used in the present tense active and with verbs that are durative (such as *boycott*) rather than stative (such as *reside*). There might also be differences between individual verbs (or lemmata, sg. lemma) in their propensity to take the progressive. However, the strength of these constraints may vary between varieties and registers, and their analysis requires annotation of the individual instances. An example of such an analysis table, similar to what it would look like in any spreadsheet programme, is given in Table 3. Any actual analysis would of course comprise more than just three occurrences.

Table 3. Annotation of concordance lines in spreadsheet (examples from Gut & Fuchs 2013: 250)

Concordance	Lemma	Stativity	Tense	Mode
on the eve of election you say you are boycotting	boycott	durative	present	active
Oghara and Idjerhe kingdoms were residing	reside	stative	past	active
contiguously with				
the way and manner in which teachers welfare	consider	durative	present	passive
are being considered				

The annotation of the concordances can be conducted in a spreadsheet program such as Microsoft Excel, into which the concordances can be fed. A free alternative to Microsoft Excel is Libre Office Calc, which offers all of the facilities required for the annotation of concordances. A particularly useful functionality, called ‘Pivot Table’, allows users to sum over all levels of a variable by a second variable (or more), e.g. how many occurrences of the progressive occur with various tenses and in various registers included in a corpus. This avoids error-prone and labour-intensive manual counting.

The methods of corpus analysis presented thus far can be applied to corpora that simply consist of texts. One way this textual information can be enhanced is through adding annotations, for example, on the level of parts of speech, syntax, and semantics. Of these, the former is the most widely used. Special tagging software determines, based on the context, the most likely word class for each word. The word *that* is ambiguous, for instance, as it can be used as a determiner, e.g., *that man*, but also as a subordinator that introduces a complement clause, e.g. *the minister said that they will offer drought relief to affected farmers*, or as a relative pronoun, e.g. *the doctor that I saw yesterday gave me new medicine*. Using a variety of techniques that look at aspects of the surrounding context, a part-of-speech tagger tries to disambiguate between these various uses. For example, the widely used CLAWS 7 part-of-speech tagset assigns the tag CST to both the subordinator and the relative pronoun, but the tag DD1 to the demonstrative use.¹⁰ State-of-the-art part-of-speech (POS) taggers work with a very fine-grained taxonomy that does not only distinguish, say, nouns from verbs, but includes classes such as singular count nouns (e.g. *horse*), plural proper nouns (e.g. *two Toyotas*) and pre-posed titles (e.g. *Mrs.*). The set of all POS tags used by a tagger is called its tagset, and may comprise more than 100 separate tags. Such fine-grained annotations enable analysts to formulate sophisticated queries with high recall (i.e., it captures all targeted instances) and precision (captures only targeted instances), such as the one for the present perfect, illustrated in Section 3.2 below. Most linguists that access existing corpora for their research receive them already complete with POS tagging. In AntConc and WordSmith Tools, POS tags can be searched in a corpus just like (parts of) words. More complex queries involving multiple (and also optional) words belonging to specific word classes can also be realised in this software, but require queries to be formulated as so-called regular expressions. While not immediately intuitive to those who are not familiar with it, the basics of this method of formulating queries can be learnt in a few hours’ time (for an introduction, see Weisser 2016: Chapter 6).

10. See the CLAWS tagging manual at <http://ucrel.lancs.ac.uk/bnc2sampler/guide_c7.htm#m4that>

State-of-the-art POS tagging is not error-free, but its error-rate is so small that comparatively time-consuming manual correction is rarely carried out. CLAWS, for instance, reports an accuracy in the vicinity of 96–97%.¹¹ By contrast, grammatical or syntactic annotation, also called parsing, is a computationally more challenging and error-prone annotation step. A recent version of the Stanford parser, which is freely available, achieves an accuracy of 91–92% (Chen & Manning 2014). While syntactic annotation can (like POS tagging) be represented in text files in a corpus, it is (unlike POS tagging) rather difficult for human readers to process, due to the hierarchical nature of syntactic structures. For the same reason, specialised software is required to process queries involving syntactic annotation, e.g. the proprietary software ICECUP that accompanies ICE Great Britain or the free platform ANNIS (Krause & Zeldes 2014). The main advantage of a parsed corpus is that syntactic queries can be carried out in a comparatively straightforward way (and much more accurately and/or efficiently than with a POS-tagged corpus), provided that specialised software is used. An example of this is where an investigation wishes to compare the syntactic realisation of the functional category of adverbials. Ideally, one would like to distinguish adverbs and preposition phrases, but exclude from the count all adverbs that function as premodifiers to another word (which can be retrieved by means of a POS tag), such as an adjective in *so tired*. Likewise, preposition phrases that function as noun postmodifiers, such as *the man on the moon*, need to be distinguished from adverbials of place. While most corpora are not available with syntactic annotation/parsing, it can subsequently be added.

Corpora can also be annotated semantically. The annotation consists in each word being assigned to a single semantic field, much like part-of-speech tagging assigns each word to a single word class. Since research in English corpus linguistics has so far relied more often on POS tagging than semantic annotation, the latter promises to open up new avenues for future research. First of all, studies focussing on syntactic and morphological properties of (African) varieties of English can take meaning into account as a conditioning factor. More broadly, meaning(-making) itself can become the focus of inquiry, in an investigation of what speakers of different varieties talk about and how they communicate intended meanings. An example of an older semantic study of African English that might have benefitted from semantic tagging, had it been available at the time, is De Klerk (2003). She investigated the use of English in a legal context in South Africa, focussing on a number of semantic domains, such as acts of communication (asking and telling), understanding, argumentation, and nouns denoting legal concepts. She mainly had to rely on her ingenuity as a researcher to make lists of words to look for in the corpus, and did

11. <<http://ucrel.lancs.ac.uk/bnc2/bnc2error.htm>>

not have the benefit of a semantic tagger, that could have classified at least some of the domains automatically. Like syntactic parsing, semantic annotation can be carried out automatically by dedicated computer programs (see Resources and Further Reading for more information). As useful as automatic semantic annotation might be, a word of caution is in order, given that innovative features of African Englishes (and indigenised varieties of English elsewhere) such as borrowings, coinages and shifts in meaning are unlikely to be annotated correctly by semantic taggers trained on corpora of British and American English. If they are of interest to a researcher, such innovative meanings will have to be annotated manually.

Corpora that are compiled for the study of phonological and phonetic phenomena further contain phonemic or phonetic transcriptions. In a phonemic transcription, the phonological form of a word that is assumed to be represented in the speaker's mind is transcribed, for example using the International Phonetic Alphabet (IPA), while a phonetic transcription represents the actual pronunciation by a speaker. Phonemic and phonetic transcriptions of spoken corpora can be carried out manually with tools such as Praat or ELAN (see Figure 1) or automatically with tools such as WebMAUS (Schiel 2004) and are always linked to the audio or video file (see Figure 1).

A final point concerns the validity of findings emanating from corpus-based research, which frequently requires statistical testing. Much, though not all, research relying on corpora focuses on quantification and often two or more quantities are compared with each other. For example, one syntactic construction may occur in the corpus more often than another construction, or learners belonging to one group commit a certain error more frequently than those from another group. As explained above, a corpus is a sample from the larger population. A quantitative comparison that holds for the corpus does not necessarily generalise to the population. This is why statistical tests are required to make such quantitative comparisons reliable. A statistical test answers the question of how likely it is that a quantitative comparison found in the corpus generalises to the whole population, such as all speakers of a certain variety of English or all learners of English with a specific first language. The fact that, say, the Past Tense is more frequent than the Present Perfect in a corpus is generally of little interest if we cannot conclude that a similar difference holds for the whole population, which is why statistics testing is essential in quantitative corpus-based research. Explaining what tests are appropriate in specific contexts would go beyond the scope of this chapter, and appropriate software and literature are pointed out under 'Further Reading.'

3. Corpus-based studies of African Englishes: Three case studies

3.1 Lexical differences between varieties of English

Having discussed the analysis of corpus data in general, this section illustrates practical aspects of corpus analysis with three case studies. These examples demonstrate the diverse set of linguistic areas on which corpus analysis can throw light, including lexis, phonology and syntax.

One methodological consideration that is important for research on lexical variation in particular, and more broadly for any conclusions regarding the frequency of linguistic entities, is the need for normalisation of such frequencies. Since corpora are of different sizes, one cannot compare the numbers obtained from one corpus to another directly. This problem is solved by normalising the frequencies obtained from two or more different corpora to a standard value, e.g. the number of times a particular phenomenon is observed per 1,000 or per 1,000,000 words. The calculation is done by dividing the number of observations in a particular corpus by the total number of words in the corpus, and then multiplying it with the base of normalisation. For phenomena with relatively high frequencies and for smaller corpora, a smaller base, such as 1,000 is suitable, but for low frequency phenomena (usually only observable in large corpora), a larger base such as 1,000,000 is more suitable.

Keeping this methodological consideration in mind, we embark on our first case study, which focuses on lexical variation in varieties of English. This topic has long interested researchers partly because speakers often have some degree of awareness of the lexical choices available to them. Traditional dialectology usually drew on researchers' introspection and surveys of speakers as sources of data, both of which rely on people being able to accurately judge their own language use. Corpus-based research does away with this requirement, which is often not warranted (e.g. people may say they prefer the lexical variant they consider 'proper'). Instead, it provides an objective source of evidence on lexical choices. Research on lexical items is equally useful for research on learner language, for example in order to uncover recurrent errors or other limitations on the vocabulary of learners that can then be taken into account in second and foreign language instruction, as illustrated by Adeyemi (2017) for Nigerian learners and Partridge (2012) for South African learners.

Postcolonial varieties of English often include loanwords from local languages, both also direct loans and loan-translations (also called 'calques'), as shown by Schmied (1991) for African Englishes generally, Van Rooy and Terblanche (2010) for Black South African English in journalism, and Mair (2013) and Udofot and Mbarachi (2016) for Nigerian English on the internet. In addition to contact with their input variety/ies (usually varieties of British English) and local languages, postcolonial varieties have also been in contact with other varieties of English.

For example, Indian English has exerted influence on other varieties of English through the movement, both voluntary and with various forms of coercion exerted by colonial authorities, of people from the Indian subcontinent to other British colonies in Southeast Asia, Oceania, the Caribbean as well as East and South Africa, as documented extensively by Mesthrie (1992b) for Indian South African English. Even within Africa, Nigerian English seems to serve as a reference point for the diffusion of lexical elements such as *wats*, *guyz*, *dis*, *luv* and *neva* to other varieties, as shown by Van Rooy and Kruger (2018).

Lexical variation is among the least technically demanding subject areas in corpus linguistics. Provided that the lexemes that are of interest are known, they can be found with a simple query in Antconc or WordSmith Tools. If derived forms, such as plurals or progressives, should be included, they can be searched for separately or with a slightly more complex query. For example, both *house* and *houses* can be found with a regular expression involving the ‘?’ operator that allows for a single optional character: *houses?*¹²

As a case study, a well-known property of African Englishes, the centrality of kinship (Wolf & Polzenhagen 2009: 77), is presented here. A salient kinship concept, *brother*, was selected for the first analysis, because this word, amongst others, is claimed to be used with a wider semantic range than just the meaning of a male sibling in sub-Saharan Africa, according to Mesthrie and Bhatt (2008: 112). The analysis was performed on the GloWbE corpus, where the five African countries represented there, together with Jamaica, which is known to be an important reference point for African language practices on the web (Mair 2013), were analysed separately, while results for the rest of the world were clustered together. These results are reported in Table 4. Where required, as for instance with the word *bra* below, a sample of concordance lines were inspected manually to determine how particular words were used. In this case, it emerged that only a small minority of instances represented the target meaning, and hence a refined search string that included a personal name was rather used.

The results show that, apart from South Africa, the word *brother* itself is more frequent in African countries than elsewhere in the world, with the highest frequencies observed in the two West African countries, Nigeria and Ghana. The use of *brother* as part of the appellation used to address another (male) person is of particular interest to this case study. The relevant data can be gathered from the corpus by looking at the collocation¹³ of the word *brother* and a proper noun,

12. The ‘regex’ checkbox needs to be activated to carry out this query. Unlike the regular search function in AntConc, this query is case-sensitive and also finds word embeddings such as *Parkhouse*.

13. One very typical way of analysing concordance data is to count the collocates of a particular search word: those words that occur frequently in the context of the keyword itself. Such

the name of a person. This result shows an even stronger trend of *brother* being a frequent choice in Africa, compared to the rest of the world.

Table 4. Normalised frequency per million words (pmw) of selected terms for 'brother' in the GloWbE corpus

	Rest of world	South Africa	Nigeria	Ghana	Kenia	Tanzania	Jamaica
<i>brother</i>	72	63	119	106	83	92	89
<i>brother+Name</i>	4	4	6	11	10	11	6
<i>bro</i>	4	8	12	4	8	5	6
<i>bra</i>	4	7	4	6	3	4	3
<i>bra+Name</i>	0.0	0.8	0.1	0.4	0.0	0.0	0.1
<i>brotha</i>	0.1	0.1	0.3	0.2	0.2	0.2	0.4
<i>broda</i>	0.0	0.1	1.9	0.1	0.2	0.1	0.0
<i>bru</i>	0.2	1.5	0.0	0.0	0.9	0.1	0.1
<i>boet</i>	0.0	2.8	0.0	0.0	0.0	0.0	0.0

The other searches involved informal lexical variants of *brother* that are often used in terms of address, drawing on the discussion of informal variants of *brother* in Branford and Venter (2016: 41–44). The shortened form *bro* has wide international currency, but the results in Table 3 show that it is used more frequently in Africa than in the rest of the world, with Nigeria taking the lead, followed by South Africa and Kenya. The next variant is the ambiguous form *bra*, whose principal meaning is to denote a piece of undergarment, but it is also widely used, especially in combination with a proper noun, to denote a person. This form is especially frequent in South Africa, with 0.8 pmw, while it is also used in Ghana at 0.4 pmw, and in lower frequencies elsewhere, except in Tanzania, with but a single case in Kenya. The form *brotha* is attested in all African countries and Jamaica approximately 0.2 pmw, similar to the United States, but largely unattested anywhere else in the world. The variant *broda* is absent from the rest of the world, and seems to be a very specific Nigerian form, with much lower frequencies elsewhere in Africa, while *bru* (derived from Afrikaans *broer*, meaning 'brother') and *boet* (the colloquial variant of Afrikaans *broer*) are very peculiarly South African usages (see Branford & Venter 2016: 44), although *bru* also gained some currency in Kenya.

This brief analysis demonstrates the value of a corpus such as GloWbE for investigating the use and distribution of lexical items, even more so because it

collocates are often useful pointers to an interpretation of the meanings or uses of a word. By way of extension, one can also consider grammatical constructions rather than lexical items in the context of the keyword, a variant of collocation that is sometimes called colligation.

allows the possibility to combine lexical searches with POS tag searches. Some provisional insights that emerge from the analysis is the support for claims about the centrality of the concept of kinship in Africa, the role of loanword adoption, specifically in South Africa, and the playfulness of informal spellings in the online environment, a point also made by Deumert (2014). What is interesting about the playful spelling variants is how they show that putatively Jamaican forms actually gain more traction in Africa than in Jamaica itself.

This prompted us to undertake a second analysis, of a few forms that display similar spelling adjustments, such as the use of the letter *a* for the unstressed vowel [ə], the letter ‘*e*’ for the diphthong [ɛɪ], and the alternative spelling *d* for the letters *th* where they represent the sound [ð]. These variant spellings are partly supported by phonetic adjustments that are typical of African Englishes (Mesthrie 2008: 311–315). The basic results for a number of variants, selected on the basis of an earlier study of South African English by Van Rooy and Kruger (2018), are presented in Table 5.

Table 5. Normalised frequencies per million words of selected terms showing phonetic spelling in the GloWbE corpus

	Rest of world	South Africa	Nigeria	Ghana	Kenya	Tanzania	Jamaica
<i>dat</i>	1.8	4.8	122.1	10.5	3.6	1.3	60.2
<i>dis</i>	1.6	3.2	72.9	7.4	2.7	0.9	27.1
<i>neva</i>	0.1	1.3	8.4	1.2	1.0	0.2	17.4
<i>shem</i>	0.4	2.7	0.3	0.7	2.1	0.4	0.3

The results show that these four variants are more frequent in all the African varieties than elsewhere in the world, but there are very clear local preferences. The forms *dat*, *dis*, and *neva* are extremely frequent in Nigeria, with Jamaica following at some distance, before the frequency level of the other African varieties is reached, whereas the form *shem* (= ‘shame’) is particularly frequent in South Africa and Kenya, but much less so elsewhere. Here, inspection of the concordance lines, as in Example (1) and (2) below, shows that such usages occur in passages that typically involve lots of respelling, a feature of online language use that is observed in the data from countries across the English-speaking world, but notably often selected by online communicators in the African countries. Van Rooy and Kruger (2018) argue that these forms represent a blend of phonetic imitation and similarity to forms that are used elsewhere in the world and in particular in Africa, which contribute to convey a hybrid, glocal identity by the users.

- (1) so dis prove d fact *dat* there will not b reaction if d reactant is faulty
(GloWbE, Nigeria)
- (2) I’m tired n sum ppl are just doing it wrong *shem* I don’t wan na lue!!!
(GloWbE, South Africa)

The analysis of lexical data offers opportunities to answer a range of research questions. These questions potentially relate to patterns of diffusion and adaptation of new words: where they are found most frequently (which often suggests the origin of the forms), but also to the creative adaptations those words undergo, and what these forms tell us about the functional motivation of the users for selecting these forms. The analysis can be done on the basis of normalised frequency alone, but often it is necessary to consider additional data, such as the context of use, that is accessed by means of a KWIC concordance line, or by narrowing down an analysis by limiting the data to a word in collocation with other forms, such as proper nouns.

3.2 The present perfect in African Englishes

Having presented an example of the analysis of lexical variation, we now move on to syntactic variation. This area has frequently been at the focus of research in variation linguistics and can, among others, provide a perspective on what unites all varieties of English and in what ways and to what degree they vary. Syntax is an equally important topic in research on learner English and can help shed light, for example, on the question of whether learners of English mostly follow the same order in acquiring syntactic structures or whether the presence of a similar structure in their first language helps them “jump the queue” and acquire this structure earlier than speakers of languages without said similar structure.

This case study focuses on the alternation of the Present Perfect (PP) and the Simple Past Tense (PT), which are the main syntactic ways of referring to the past in English (the Past Perfect is rare and restricted to specific contexts). While replacing the PP with the PT, or vice versa, in any given sentence changes its meaning in more or less subtle ways, it does not change whether the sentence is true. By contrast, this is not necessarily the case when replacing the PP/PT with the Present Tense. On this basis, the PP and PT are also known as truth-conditionally equivalent.

Given that the PP and PT can be replaced with each other, the next question that arises is what factors influence whether one or another is used in a specific context. While there are several such factors, one that has been highlighted in previous research is regional or dialectal differences. For example, in American English (AmE) the PP is used much less frequently than British English (BrE), and in Asian Englishes it tends to be even rarer (Yao & Collins 2012). The question then arises how frequently the PP is used in African Englishes and to what extent African Englishes differ from each other in this variable. The fact that all African Englishes emerged under British influence suggests that they might replicate the comparatively high frequency that the PP enjoys in BrE, while more recent influence from AmE (Awonusi 1994) with its low PP frequency could have led, in turn, to a decrease in its frequency in African Englishes. The PT is also a simpler form, consisting of stem and past tense marker, than the PP, consisting of auxiliary, stem and past participle

marker, and L2 varieties tend to favour simple over complex structures (Davydova 2011: 288f.). Ultimately, an analysis of the frequency of the PP and PT in African Englishes can contribute in a small way to addressing the issue raised at the start of this section, i.e. what the limits of variation among varieties of English are.

Methodologically, researching syntactic variation with the help of corpora tends to require somewhat more technical sophistication than an analysis of lexical variation. Syntactic annotation would make the task of querying a corpus for a specific construction relatively easy from a conceptual point of view. However, as such annotation is rarely available, the alternative consists of relying on POS tagging.

Where a syntactic structure comprises a single word and is represented in the tagset of a POS tagger by one or more tags, and these tags capture exclusively this syntactic category, formulating a precise query is straightforward. A case that comes close to this ideal situation is the English past tense, which is captured in the CLAWS7 tagset by the tags VVD (past tense of lexical verbs), VBDR (*were*), VBDZ (*was*), VDD (*did*) and VHD (*had*, past tense). Thus, all tags capturing the past tense start with a V, followed by another letter, a D and an optional fourth letter. This is captured by the regular expression V.D.?, where the dot operator matches any character and the question mark operator allows for the preceding character to occur zero or one time. While this regular expression captures all occurrences of the past tense (provided that the POS tagging is accurate, of course), it also includes the past perfect (e.g. *had worked*).

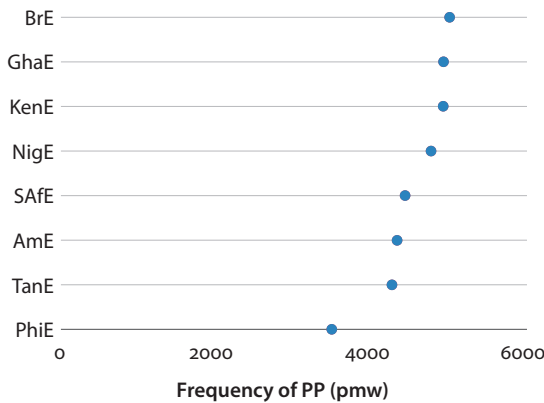
A syntactic structure that spans multiple words consequently requires a query that captures all these words. If the syntactic structure allows intervening elements, these also need to be allowed for. This is the case for the Present Perfect (PP), which consists of a form of the auxiliary *have* and a past participle, and allows intervening adverbials, as Example (3) shows. Occurrences of the PP can be captured by a regular expression search with high recall and precision, as used in Fuchs et al. (2016) and Fuchs (2016). Such search routines are usually developed iteratively which means that their performance is checked and improved in several steps, and, in this case, over the course of more than one publication. We explain in the Appendix how the regular expression used in this study works.

- (3) the spokesman of the Northern Elders' Forum *has never hesitated* to speak up
(GloWbE, Nigeria)

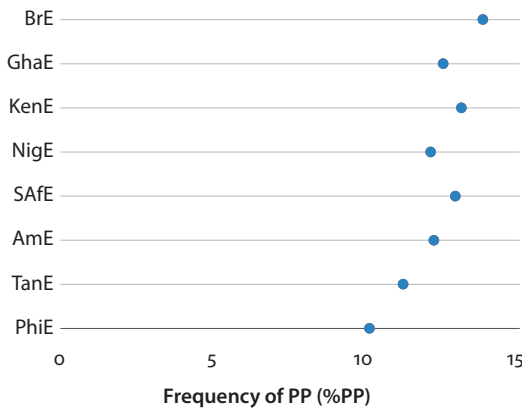
Once all occurrences of the PP and PT have been identified, the frequency of the PP can be operationalised in two ways. One is to normalise the raw frequency of the PP following the concept of normalisation introduced at the beginning of Section 3.1. The other way of operationalising the frequency of the PP is to compare it to the frequency of the PT and calculate a ratio, based on the argument, mentioned above, that they are truth-conditionally equivalent. This measure yields a number indicating what percentage of past time references in the corpus is expressed by the PP.

In Fuchs (2016), these methods were applied to all 20 national varieties of English included in the GloWbE corpus. This section discusses the results for the five African varieties included and, for comparison, three more varieties, i.e. British and American English (which have exerted some degree of influence on many other varieties of English) as well as Philippine English (PhiE), the variety with the lowest PP frequency in the corpus.

As Figure 3 indicates, BrE uses the PP the most frequently and PhiE the least frequently for both measures, and this is true not only for the varieties shown here but for all 20 considered in the original study. In AmE, the PP is less frequent than in most though not all of the African varieties. Overall, the data reveals a regional effect. First of all, it appears that neighbouring varieties might converge in PP frequency.



a.



b.

Figure 3. Frequency of the PP in African Englishes and selected other Englishes, measured per million words (left panel) and as percentage of all references to the past (right panel)

This is the case for Ghanaian English (GhaE) and Nigerian English (NigE). However, the East African varieties of Tanzanian English (TanE) and Kenyan English (KenE) provide a counter-example as they diverge to a certain extent in PP frequency, and South African English (SAfE) has no close neighbour in the present data, so does not help us in assessing the strength of the purported regional effect.

The evidence becomes more conclusive once all 20 varieties are compared on the basis of which world region they are situated in. The African varieties reveal, by and large, PP frequencies that are much higher than average – notice that PhiE uses the PP much less frequently than any of the African varieties. In fact, the regional effect, or geographic proximity, turned out to be the only robust statistical predictor of PP frequency in the 20 variety sample considered in the study. Other factors that, based on previous research, were thought to have explanatory value, turned out to be statistically insignificant. These were

- a measure of independence of the variety from its colonial input variety, operationalised as current stage in Schneider's Dynamic Model of Postcolonial Englishes (Schneider 2007) – the hypothesis was that varieties at a later stage in their development tend to diverge more from BrE than varieties at an earlier stage,
- whether a variety belongs to the Inner or Outer Circle of Kachru's (1985) Circle Model, i.e. whether English is mainly used as a first or a second language, respectively, in this country – the hypothesis was that varieties belonging to the same group are more similar to each other than varieties belonging to different groups,
- how formal a variety is, as measured by linguistic indicators of formality (e.g. a low frequency of first and second person personal pronouns and a high frequency of nouns) – the hypothesis was that more formal varieties tend to use the PP more frequently than less formal varieties,
- how much media language the subcorpus for each variety contains – the hypothesis was that subcorpora of varieties with more media language use the PP more frequently than those with less media language.

The conclusion that only geographic proximity can explain how PP frequency varies across the 20 varieties of English was reached with a linear regression analysis. This is a statistical method that assesses the influence of a number of so-called independent variables on a dependent variable (here PP frequency), the independent variables being in this case geographic proximity, stage in Schneider's model, Kachru's model, formality, proportion of media language, and the dependent variable being PP frequency.

This result has interesting implications for the present volume. It suggests that the idea that there are African Englishes is meaningful beyond just being a shorthand that refers to varieties of English spoken on this continent. The case study indicates that, at least when it comes to PP frequency, African Englishes are relatively similar. If we compare two random African Englishes, the difference in PP frequency

is likely to be smaller than if we compare any African variety with a non-African variety. In this way, the case study also contributes in a small way to answering the wider question, asked at the beginning of this section, to what extent varieties of English differ and what the limits of variability are.

This conclusion is of course limited by the restrictions of the study. The PP is just one syntactic category and the corpus consists of material gathered from the internet and presumably mostly written by people who are relatively well-educated and proficient in English. The study also provided a bird's eye view of the PP and did not take into account a number of language-internal and -external factors. One among them is potential influence from local languages, which is often considered in research on postcolonial Englishes – although, in a separate study on the PP in Nigerian English (Werner & Fuchs 2017), this factor turned out to be relatively unimportant.

3.3 /h/-deletion in Nigerian English

While lexical and syntactic variation, discussed in the first and second case studies, are traditionally at the focus of corpus analysis, the use of corpora in phonological studies is still rare and has only recently become more widespread (e.g. Gut & Fuchs 2017). The phonological process studied here is known as /h/-deletion, and takes place when the glottal fricative /h/ is not pronounced in words like *house* or *who*. Previous research has claimed that this process exists in Nigerian English, at least for Yoruba and other southern Nigerian speakers (Simo Bobda 2007: 285; Awonusi 1990, 2007). /h/-deletion does not only occur in Nigerian English but has also been observed in various British dialects such as London English (Hughes & Trudgill 1987), working class speech in Wales and England (Upton 2004) and in the North of England (Beal 2004), as well as in (working class) dialects of Norwich and Ipswich (Trudgill 2004). According to Wells (1982) it started as an innovation in London, from where it spread to Norwich and other East Anglian towns. Furthermore, /h/-deletion has been reported for Australian English (Horvath 2004), Jamaican English (Wells 1982), Bahamian English (Childs & Wolfram 2004), Malaysian English (Baskaran 2004), vernacular Newfoundland British English (Clarke 2004), South African Indian English (Mesthrie 1992a) and Sierra Leone Krio (Jones 1971).

For some English varieties, this phonological process has been suggested to vary with the speaker's gender and speaking style. Horvath (2004), for example, states that /h/-deletion is more frequent in Australian English male speech than in female speech and more common in informal conversations than in formal speaking styles. However, it has not been tested yet whether the factors speaker gender and speaking style influence the rate of /h/-deletion to the same extent in non-native varieties of English. Moreover, possible differences across non-native English speakers of different ethnicity, as suggested by Awonusi (1990) and Simo Bobda (2007), have not yet been explored.

We therefore carried out a corpus-based analysis of the frequency and variability of /h/-deletion in the speech of university-educated Nigerians. The study aimed to provide answers to the basic question of the overall rate of /h/-deletion in Nigerian English and explores whether and how the deletion rates are influenced by the speaking style, speaker gender and the speaker's ethnic group. The data for the corpus analysis was drawn from the Nigerian component of the ICE corpus, which has time-aligned annotations made with ELAN (see Figure 1) and thus allows phonetic and phonological analyses. Four speaking styles were selected, comprising 40,138 words of broadcast talks and 40,916 words of broadcast news, 61,305 words of unscripted speeches and 40,303 words of broadcast discussions. The category broadcast talks comprises scripted speeches given on radio or TV such as the President's address to the nation. The category broadcast news comprises news items read out by TV or radio presenters. The unscripted speeches consist of monologues given for example at meetings or conferences, while in the broadcast discussions several participants discuss a specific topic on the radio or TV. The data comprises speech produced by a total of 246 university-educated Nigerian speakers (68 female and 178 male) with at least 12 different ethnic backgrounds (76 Yoruba, 58 Igbo, 31 Edo, 31 Hausa, 5 Ibibio, 3 Efik, 2 Ijaw, 1 Urhobo, 1 Esan, 1 Tiv, 1 Izon, 1 Fulani, 35 unknown), who contributed different amounts of speech to the corpus.

The recordings were analysed auditorily for the presence of /h/-deletion in all words that

- contain an initial *h* in spelling that is pronounced /h/ such as *hand* (Type I) or
- contain an initial *wh* in spelling that is pronounced /h/ such as *who* (Type II).

Words with medial *h* were not included in the analysis due to their variable pronunciation: *vehicle*, for example, can be pronounced with or without word-internal /h/. A total of 3,931 words of Types I and II occurred in the data. The auditory analysis proceeded in three separate rounds and involved seven transcribers. Using ELAN, one transcriber marked for each word whether an [h] was present or not, and a second one did the same independently. In case of disagreement between the first two transcribers, who comprised both trained linguists and students of linguistics, a third one (always a trained linguist) decided on the presence of [h] or not. The speakers' gender and ethnic background was drawn from the corpus metadata.

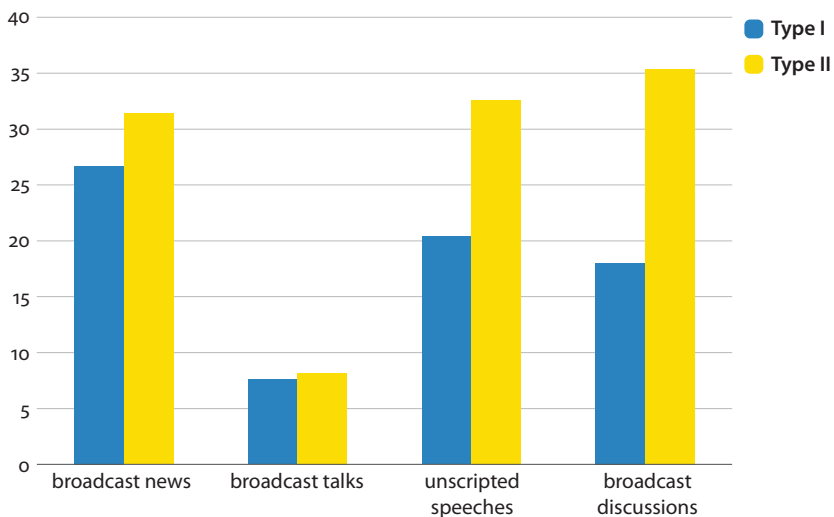
The results show that the rate of /h/-deletion in NigE lies at 19.6%, but varies considerably with the type of word. Table 6 shows that the rate of /h/-deletion is significantly higher in words that are spelled with *wh* than in words that are spelled with *h*, (27.2% vs. 18.2%, $p < 0.0001$, $G^2 = 21.8$ according to a log-likelihood test). This finding, however, needs to be interpreted with caution: Type II words form a very small lexical set in English so that lexical effects will play a role. In our data, 94.6% of these words were tokens of *who*, *whom*, *whose*, or *whoever*.

Table 6. Percentage and absolute number of /h/-deletion in Type I and Type II words and total number of words analysed

		Words with /h/-deletion	Number of analysed words
Type I:	<h> /h/	18.2% (602)	3302
Type II:	<wh> /h/	27.2% (171)	629

143 out of the 246 speakers (58%) that were investigated in this study produce /h/-deletion at least once. With regard to /h/-deletion in the two types of words, 80 speakers show /h/-deletion only in Type I words and 11 only in Type II words; 52 speakers have both processes in their speech. The speakers having /h/-deletion come from a wide range of different ethnic backgrounds including Yoruba (50 speakers), Igbo (22), Edo (21), Hausa (9), Ibibio (4), Efik (2), Ijaw (1), Esan (1), Urhobo (1) and Izon (1). The ethnicity of 31 speakers having /h/-deletion is unknown. The results suggest strongly that /h/-deletion occurs across all ethnic groups in Nigeria, including speakers from Northern regions. We would therefore like to claim that /h/-deletion is a structural feature of an emerging supraregional standard form of Nigerian English. Future research will have to substantiate this suggestion.

Figure 4 shows that the rate of /h/-deletion varies with speaking style. It is significantly lower ($p < 0.0001$ in each case) in broadcast talks than in the other speaking styles for both words spelled with *h* (Type I) and *wh* (Type II).

**Figure 4.** Percentage of Type I and Type II words with /h/-deletion across speaking styles

This confirms that specific genres differ systematically with respect to the rate of /h/-deletion.

4. Future prospects

We would like to conclude this chapter with an appeal to readers who have themselves gathered linguistic data that might be considered a corpus or database. Making data that is already digitised available to the research community is easy and fast, and contributes greatly to the reliability of research results produced by the whole community. One online publishing option is Sourceforge, another is the Open Science Framework.^{14,15} An important precondition is that informants need to consent to their data being made available and that their identity is anonymised, although information on their background (e.g. first language, age, gender, education) can be very useful and should be preserved. Both conditions can be waived if the data is already publicly available, provided that copyright issues are taken care of. On the other hand, special care needs to be taken in the case of informants who cannot, by law, give consent, such as children. Even though their legal guardians can give consent in their stead, the protection of their identity should nevertheless be an even higher priority than in the case of adult informants.

Making existing corpora publicly available is certainly motivated on altruistic grounds, as it benefits the research community (as argued in Section 1). If that were not enough, there are also clear practical benefits to the original compiler(s) of the data. As other researchers build on the existing corpus, by conducting their own studies on it and gathering parallel data from other contexts, the work of the original compiler(s) becomes known to and appreciated by a greater number of readers. With a corpus that is widely available, multiple researchers can work on understanding a variety better, drawing on the combined energy of several analysts.

5. Resources and further reading

Introductions to corpus linguistics

McEnery and Hardie (2012) provide a general introduction, Weisser (2016) a practical introduction, and Friginal and Hardy (2014) explain how corpora can be used in sociolinguistics. FutureLearn has a free (in the basic version) Massive Open Online Course on corpus linguistics.¹⁶

14. <www.sourceforge.com>

15. <www.osf.io>

16. <<https://www.futurelearn.com/courses/corpus-linguistics>>

Concordancers

Concordancers are computer programs that allow users to search a corpus and often also provide other functionalities such as the compilation of a list of all words contained in the corpus. Many linguists use AntConc (available free of charge) and WordSmith Tools (most recent version available for a fee, earlier versions for free).^{17,18}

Statistics

Many corpus linguists use the software R for statistical analysis, which is freely available from www.r-project.org, and is controlled through a command line interface. While R can be used on its own, its functionality can be further enhanced with RStudio, an Integrated Development Environment for R, and RCommander, a graphical user interface, also freely available.^{19,20} JASP is another computer program dedicated to statistical analysis and is controlled through a graphical user interface.²¹ Both FutureLearn and EdX have free (in the basic version) Massive Open Online Courses on the use of R for data analysis.²² Another free resource that focusses specifically on the use of R for the statistical analysis of linguistic data is Jeroen Claes's website, which is practically oriented and includes exercises.²³

Proprietary software for statistical analysis, SPSS and Stata, is available at many universities in high-income countries, but not necessarily in low- and middle-income countries. Both programs are controlled via Graphical User Interface. PSPP is a free alternative to SPSS and was designed to replicate many of its functions.²⁴ The proprietary Microsoft Excel and free Libre Office Calc also allow for simple statistical tests.

Introductions to statistics for (corpus) linguistics with R include Gries (2013, 2016), Levshina (2015), and Desagulier (2017).

17. <<http://www.laurenceanthony.net/software/antconc/>>

18. <<http://www.lexically.net/wordsmith/>>

19. <www.rstudio.com>

20. <<http://www.rcommander.com/>>

21. <<https://jasp-stats.org/>>

22. <<https://www.futurelearn.com/courses/data-science>, <https://www.edx.org/course/data-science-r-basics-harvardx-ph125-1x-0>>

23. <http://www.jeroenclaes.be/statistics_for_linguistics/>

24. <<http://www.gnu.org/software/pspp/pspp.html>>

Part-of-speech tagging

The CLAWS tagger is widely used and available for a fee so it can be used on a local computer, or remotely via the Wmatrix workbench.^{25,26} Texts of up to 100,000 words can be tagged for free.²⁷ Readers who use existing corpora need only be familiar with the tagsets it employs, of which there are several versions. The perhaps most widely used one is CLAWS7.²⁸

A freely available alternative is the Stanford tagger, which is run via the command line, i.e. there is no graphical user interface.²⁹ Pre-trained models use the Penn tagset, but if an existing corpus is available with a different tagset, then a new model can be trained on it.

Parsing

The Stanford parser is, like the Stanford tagger, freely available and runs via the command line.³⁰

Semantic tagging

The CLAWS semantic tagger is, like the POS tagger, available via Wmatrix and for free for up to 100,000 words.³¹

Phonological transcription

The software Praat can be used for the annotation and acoustic analysis of audio data and is freely available from <<http://praat.org>>. The software ELAN can be used for the annotation and analysis of audio and video data and is freely available from <<https://tla.mpi.nl/tools/tla-tools/elan/>>. The online tool MAUS can be used for creating automatic phonological annotations, if an orthographic transcription

25. <<http://ucrel.lancs.ac.uk/claws/>>

26. <<http://ucrel.lancs.ac.uk/wmatrix/>>

27. <<http://ucrel.lancs.ac.uk/claws/trial.html>>

28. <<http://ucrel.lancs.ac.uk/claws7tags.html>>

29. <<https://nlp.stanford.edu/software/tagger.shtml>>

30. <<https://nlp.stanford.edu/software/lex-parser.shtml>>

31. <<http://ucrel.lancs.ac.uk/usas/>>

is available. It is freely available from <<https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSGeneral>>. An additional online tool, also available from BASWebServices, generates orthographic transcriptions. The two tools can be used in conjunction.

Acknowledgements

We would like to express our gratitude to the two reviewers, whose comments greatly helped in improving the manuscript.

References

- Adeyemi, Remilkun. 2017. A Corpus-based Analysis of Cohesion in Written English Essays by Nigerian Students. PhD dissertation, North-West University.
- Awonusi, Victor. 1990. Coming of age: English in Nigeria. *English Today* 22: 31–35. <https://doi.org/10.1017/S0266078400004715>
- Awonusi, Victor O. 1994. The Americanization of Nigerian English. *World Englishes* 13(1): 75–82. <https://doi.org/10.1111/j.1467-971X.1994.tb00284.x>
- Awonusi, Victor. 2007. *Good Spoken English and national development: Sociophonology in the service of man*. Inaugural lecture held at the University of Lagos, March 14.
- Baskaran, Loga. 2004. Malaysian English: Phonology. In *A Handbook of Varieties of English*, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 1047–1063. Berlin: Mouton de Gruyter.
- Beal, John. 2004. English dialects in the North of England: Phonology. In *A Handbook of Varieties of English*, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 134–162. Berlin: Mouton de Gruyter.
- Branford, Jean & Venter, Malcolm. 2016. *Say Again? The Other Side of South African English*. Cape Town: Pharos.
- Brato, Thorsten. 2018. ‘Outdooring’ the Historical Corpus of English in Ghana. *English Today* 34: 25–34.
- Chen, Danqi & Manning, Christopher D. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014, Doha Qatar*, Alessandro Moschitti & Walter Daelemans (eds). Stroudsburg PA: ACL.
- Childs, Becky & Wolfram, Walt. 2004. Bahamian English: Phonology. In *A Handbook of Varieties of English*, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 643–656. Berlin: Mouton de Gruyter.
- Clarke, Sandra. 2004. The legacy of British and Irish English in Newfoundland. In *Legacies of Colonial English: Studies in Transported Dialects*, Raymond Hickey (ed.), 242–261. Cambridge: CUP.
- Davies, Mark & Fuchs, Robert. 2015. Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE). *English World-Wide* 36(1): 1–28. <https://doi.org/10.1075/eww.36.1.01dav>

- Davydova, Julia. 2011. *The Present Perfect in Non-native Englishes: A Corpus-based Study of Variation*. Berlin: De Gruyter. <https://doi.org/10.1515/9783110255027>
- de Klerk, Vivian. 2003. The language of truth and reconciliation: Was it fair to all concerned? *Southern African Linguistics and Applied Language Studies* 21(1–2): 1–14. <https://doi.org/10.2989/16073610309486324>
- de Klerk, Vivian. 2006. *Corpus Linguistics and World Englishes: An Analysis of Xhosa-English*. London: Continuum.
- De Kock, Leon. 1996. *Civilising Barbarians: Missionary Narrative and African Textual Response in Nineteenth-century South Africa*. Johannesburg: Witwatersrand University Press.
- Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R*. Heidelberg: Springer. <https://doi.org/10.1007/978-3-319-64572-8>
- Deumert, Ana. 2014. *Sociolinguistics and Mobile Communication*. Edinburgh: EUP.
- Friginal, Eric & Hardy, Jack. 2014. *Corpus-based Sociolinguistics a Guide for Students*. London: Routledge.
- Fuchs, Robert. 2016. The frequency of the present perfect in varieties of English around the World. In *Re-Assessing the Present Perfect*, Valentin Werner, Elena Seoane & Cristina Suárez-Gómez (eds), 223–258. Berlin: De Gruyter.
- Fuchs, Robert, Götz, Sandra & Werner, Valentin. 2016. The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In *Re-Assessing the Present Perfect*, Valentin Werner, Elena Seoane & Cristina Suárez-Gómez (eds), 297–337. Berlin: De Gruyter.
- Granger, Sylviane, Estelle Dagneux, Fanny Meunier & Magali Paquot (eds). 2009. *International Corpus of Learner English (Version 2)*. Louvain: UCL Presses Universitaires de Louvain.
- Greenbaum, Sydney & Nelson, Gerald. 1996. The International Corpus of English (ICE) project. *World Englishes* 15(1): 3–15. <https://doi.org/10.1111/j.1467-971X.1996.tb00088.x>
- Gries, Stefan. 2013. *Statistics for linguistics with R: A Practical Introduction*, 2nd edn. Berlin: De Gruyter. <https://doi.org/10.1515/9783110307474>
- Gries, Stefan T. 2016. *Quantitative Corpus Linguistics with R: A Practical Introduction*, 2nd edn. London: Routledge.
- Gut, Ulrike & Fuchs, Robert. 2013. Progressive aspect in Nigerian English. *Journal of English Linguistics* 41(3): 243–267. <https://doi.org/10.1177/0075542413492799>
- Gut, Ulrike & Fuchs, Robert. 2017. Exploring speaker fluency with phonologically annotated ICE corpora. *World Englishes* 36(3): 387–403. <https://doi.org/10.1111/weng.12278>
- Gut, Ulrike & Voormann, Holger. 2014. Corpus Design. In *The Oxford Handbook of Corpus Phonology*, Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds), 13–26. Oxford: OUP.
- Horvath, Barbara M. 2004. Regional characteristics of Australian English: phonology. In Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds.), *A Handbook of Varieties of English*, 625–644. Berlin and New York: Mouton de Gruyter.
- Hudson-Ettle, Diane M. & Schmied, Josef. 1999. *Manual to Accompany the East African Component of the International Corpus of English*. Chemnitz: Chemnitz University of Technology, Department of English. Web: <<https://www.tu-chemnitz.de/phil/english/sections/linguist/real/independent/ICE-EA/index.html>>
- Hughes, Arthur & Trudgill, Peter. 1987. *English Accents and Dialects*, 2nd edn. London: Arnold.
- Jeffery, Chris. 2003. On compiling a corpus of South African English. *South African Linguistics and Applied Language Studies* 21(4): 341–344. <https://doi.org/10.2989/16073610309486353>
- Jones, Eldred. 1971. Krio: An English-based language of Sierra Leone. In *The English Language in West Africa*, John Spencer (ed.), 66–94. London: Longman.

- Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In *English in the World: Teaching and Learning the Language and Literatures*, Randolph Quirk & Henry Widdowson (eds), 11–30. Cambridge: CUP.
- Krause, Thomas & Zeldes, Amir. 2014. ANNIS3: A new architecture for generic corpus query and visualization. *Literary and Linguistic Computing* 31(1): 118–139.
- Kruger, Haidee & van Rooy, Bertus. 2017. Editorial practice and the distinction between error and conventionalised innovation in New Englishes: The progressive in Black South African English. *World Englishes* 36(1): 20–41. <https://doi.org/10.1111/weng.12202>
- Leech, Geoffrey. 1992. Corpus linguistics and theories of linguistic performance. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*, Jan Svartvik (ed.), 105–122. Berlin: Mouton de Gruyter.
- Levshina, Natalia. 2015. *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.195>
- Mair, Christian. 2013. Corpus approaches to the New English web: Post-colonial diasporic forums in West Africa and the Caribbean. *Covenant Journal of Language Studies* 1(1): 17–31.
- Marungudzi, Thadeus. 2017. Selected Morphosyntactic Features and Variety Status of Zimbabwean English: A Corpus-based Approach. PhD dissertation, North-West University.
- McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: CUP. <https://doi.org/10.1093/oxfordhb/9780199276349.013.0024>
- Meierkord, Christiane, Minow, Verena & Morreira, Kirsten. 2009. The Corpus of Black and Coloured South African English in Contact.
- Mesthrie, Rajend. 1992a. *English in Language Shift*. Cambridge: CUP.
- Mesthrie, Rajend. 1992b. *A Lexicon of South African Indian English*. Leeds: Peepal Tree Press.
- Mesthrie, Rajend. 2008. Synopsis: The phonology of English in Africa and South and Southeast Asia. In *Varieties of English, Vol. 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 307–319. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110208429>
- Mesthrie, Rajend & Bhatt, Rakesh M. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511791321>
- Nkemele, Daniel. 2011. *Exploring Academic Writing in Cameroon English: A Corpus-based Perspective*. Göttingen: Cuvillier Verlag.
- Nkemele, Daniel. 2012. The expression of modality in Cameroon English. In *Language Contact in a Post-colonial Setting: The Linguistic and Social Context of English and Pidgin in Cameroon*, Eric A. Achimbe (ed.), 29–62. Berlin: Walter de Gruyter.
- Rossouw, Ronel & Van Rooy, Bertus. 2012. Diachronic changes in Modality in South African English. *English World-Wide* 33(1): 1–32. <https://doi.org/10.1075/eww.33.1.01ros>
- Schiell, Florian. 2004. MAUS goes iterative. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lissabon*, Vol. 3, Maria Teresa Lino (ed.), 1015–1018. Paris.
- Schmied, Josef. 1990. Corpus linguistics and non-native varieties of English. *World Englishes* 9(3): 255–268. <https://doi.org/10.1111/j.1467-971X.1990.tb00264.x>
- Schmied, Josef. 1991. *English in Africa*. London: Longman.
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511618901>
- Simo Bobda, Augustin. 2007. Some Segmental rules of Nigerian English phonology. *English World-Wide* 28 (3): 279–310. <https://doi.org/10.1075/eww.28.3.04sim>
- Trudgill, Peter. 2004. The dialect of East Anglia. In *A Handbook of Varieties of English*, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 163–177. Berlin: Mouton de Gruyter.

- Udofot, Inyang & Mbarachi, Chibuike S. 2016. Social media English in Nigeria. *Research Journal of English Language and Literature* 4(2): 775–784.
- Upton, Clive. 2004. Synopsis: Phonological variation in the British Isles. In *A Handbook of Varieties of English*, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 1063–174. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110175325>
- van Rooy, Bertus. 2006. The extension of the progressive aspect in Black South African English. *World Englishes* 25(1): 37–64. <https://doi.org/10.1111/j.0083-2919.2006.00446.x>
- van Rooy, Bertus & Kruger, Haidee. 2016. The innovative progressive aspect of Black South African English: The role of language proficiency and normative processes. *International Journal of Learner Corpus Research* 2(2): 205–228. <https://doi.org/10.1075/ijlcr.2.2.04van>
- van Rooy, Bertus & Kruger, Haidee. 2018. Hybridity, globalisation and models of Englishes: English in South African multilingual digital repertoires. In *Modelling World Englishes in the 21st Century: Assessing the Interplay of Emancipation and Globalization of ESL Varieties* [Varieties of English around the World G61], Sandra Deshors (ed.), 77–108. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g61.04van>
- van Rooy, Bertus & Piotrowska, Caroline. 2015. The development of an extended time period meaning of the progressive in Black South African English. In *Grammatical Change in English World-Wide* [Studies in Corpus Linguistics 67], Peter Collins (ed.), 465–483. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.67.19roo>
- van Rooy, Bertus & Terblanche, Lize. 2010. Complexity in word-formation processes in New Varieties of South African English. *Southern African Linguistics and Applied Language Studies* 28(4): 357–374. <https://doi.org/10.2989/16073614.2010.548022>
- Weisser, Martin. 2016. *Practical Corpus Linguistics: An Introduction to Corpus-based Language Analysis*. Chichester: Wiley & Sons. <https://doi.org/10.1002/9781119180180>
- Werner, Valentin & Fuchs, Robert. 2017. The present perfect in Nigerian English. *English Language and Linguistics* 21(1): 129–153. <https://doi.org/10.1017/S1360674316000137>
- Wells, John. 1982. *Accents of English*, 3 Vols. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511611759>
- Wolf, Hans-Georg & Polzenhagen, Frank. 2009. *World Englishes: A Cognitive Sociolinguistic Approach*. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110199222>
- Wunder, Eva-Maria, Voormann, Holger & Gut, Ulrike. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.
- Yao, Xinyue & Collins, Peter. 2012. The present perfect in World Englishes. *World Englishes* 31(3): 386–403. <https://doi.org/10.1111/j.1467-971X.2012.01756.x>

Appendix

Regular expression used in Section 3.2 (and previously in Fuchs 2016; Fuchs et al. 2016)

This regular expression (short ‘regex’) finds occurrences of the Present Perfect in a corpus tagged with the CLAWS7 (part-of-speech) tagset.

```
\w+(\VH0|VHZ)(\w+_(XX|R(\w\d)*?|MD|UH|FU)){0,4}(\w+_(AT(\w\d)*?|AP-
PGE|UH|FU)){0,2}(\w+_(JJ(\w\d)*?|N(\w\d)*?|UH|FU)){0,2}(\w+_(PPH1|PP[IH]
S(\w\d)*?|PPY|NP(\w\d)*?|D(\w\d)*?|NN(\w\d)*?|UH|FU)){0,2}(\w+_(XX|R(\
w\d)*?|MD|UH|FU)){0,4}\w+_(VBN|VHN|VVN|VVK)
```

While it might at first look intimidating, readers will find it easier to understand the inner workings of this search routine if they analyse smaller parts of it first.

- Let us assume that all instances of the Present Perfect consist of a finite form of HAVE followed by a past participle, with no or a limited number of intervening words such as time adverbials.
- We start with the expression `\w+`, which occurs repeatedly in the regex, and captures a single word. It consists of the wildcard `w`, which captures any letter, the backslash that expresses that the `w` refers to the wildcard and not the letter `w`, and the `+` operator that flexibly repeats the wildcard such that any word comprising between one and a theoretically infinite number of letters is allowed. Because the `w` wildcard does not capture spaces or punctuation, it is restricted to single words.
- The whole regex operates on texts in which the tags are separated from tagged words by an underscore. This underscore needs to be included in the query. Thus, the expression `\w+_VH0` finds all words with the `VH0` tag, which refers to finite *have* (as opposed to the infinitive *have*). In addition, we allow for finite *has*, captured by the tag `VHZ`, and express these two alternatives by using the `|` operator and surrounding them with brackets. Thus, `\w+_ (VH0|VHZ)` finds occurrences of both finite *have* and *has* in a single query.
- Moving now to the final part of the regex, the expression `\w+_ (VBN|VHN|VVN|VVNK)` is similarly constructed and targets a past participle. The first part captures the word itself, and the last part captures a tag. More specifically, *been* (`VBN`), *had* (`VHN`), past participles of lexical words (`VVN`) and past participles of catenative verbs (`VVVK`) are targeted.
- The rest of the regex between a form of HAVE and a past participle is surrounded by brackets and followed by the expression `{0,4}`. The latter expression allows for 0 to 4 occurrences of what this part of the regex targets. It thus finds both Present Perfects that just consist of HAVE and a past participle, but also those with a limited number of words in between.
- We could have simply allowed for a combination of up to four words no matter what their parts-of-speech are. Instead we opted to search for particular structures, including time adverbials and noun phrases, in order to reduce the number of false positives (instances recognised as being Present Perfects but that are actually something else). Having demonstrated the principles of how regexes are constructed, we do not explore the structure of the middle part of this regex further due to constraints of space.
- Having remarked at the start of the Appendix that this regex might look somewhat intimidating at first, we would like to stress that it was not constructed in a single attempt. Instead, Fuchs (2016) and Fuchs et al. (2016) adapted an existing regex, and tested and refined it iteratively on their data by (1) checking the instances it found to weed out those that are not Present Perfects (i.e. false positives) and modifying the regex and (2) checking the search result against the actual text in order to identify Present Perfects that are not captured by the search routine (false negatives), and modifying the regex so that such cases are captured in the future.

The purpose, design and use of the Corpus of Nigerian and Cameroonian English Learner Language (Conacell)

Alexandra U. Esimaje
Benson Idahosa University

This chapter reports on a new learner corpus project, describes its purpose and design, and demonstrates its use: the error-annotated corpus of 442,939 words of Nigerian and Cameroonian English Learner Language (Conacell) (Esimaje 2016). The aim of the project is to produce a resource for measuring learners' language development and to enhance the teaching-learning process. The corpus data comprises the language output of 998 students; 383 university and 615 secondary students. The specific uses of the corpus to explore lexical form and tense usages by learners in Nigeria and Cameroon are shown. Corpus analysis reveals, for instance, that the lexical form of capitalisation and tense are hard to learn, and therefore remain learning needs in the contexts.

1. Introduction

1.1 The learner corpus

Of all the exciting developments in corpus-based studies, the creation and use of learner corpora in language pedagogy holds the most attraction for those of us involved in language education. Notable among this research is the work of Sylviane Granger, who led the compilation of the 'most used' learner corpus – the International Corpus of Learner English (ICLE 2009). As Myles (2005) argues, SLA research depends on good quality data sets through which the mental developmental processes of learners can be observed and corpora of learners' language remain the only central evidence of such. Myles' point is illustrated in a study of tense uses by university learners in Nigeria carried out by Esimaje and Hunston (2016), which revealed a pattern of errors involving lexical BE and thus pointed to a pedagogical need of the learners in this area. Earlier, Hunston (2002) pointed out that the essence of work on learner corpora is comparison: between corpora

produced by different sets of learners, and between corpora produced by learners and those produced by native speakers. Such comparisons help students ‘notice’ the gaps between their interlanguage and the language they are learning. This point was highlighted by Granger (2012: 3237), who explained that “one of the best ways of uncovering the features that are typical of learner language is to compare it to other samples of learner language and /or to a control corpus of expert production”. A learner corpus is a systematic collection of the language produced by foreign or second language learners which is stored and processed on a computer. It consists of principled collections of learner language which give insight to how foreign or second language is acquired (Lüdeling et al. 2005).

Granger (2002) traces the history of computer learner corpora (CLC) to the period between the late 1980s to the early 1990s when investigations into non-native varieties of English language began through the collections of English usage by learners which came to be referred to as learner corpora. As she pointed out, learner corpora provide a new type of data which informs thinking in second language acquisition (SLA) and foreign language teaching (FLT), revealing the mechanisms of acquisition, and improving the learning and teaching of second and foreign languages. As she notes, formerly SLA/FLT had drawn their data from language use data, metalingual data, and self-reports but these sources notoriously excluded the learner output. Through learner corpora, this void has now been effectively filled and corpora have also illuminated the other variables of learning such as learning styles and needs. The main reason for compiling a learner corpus, Granger (1998) says, is to gather objective data for describing learner language. However, generally, learner corpora are important because they provide a record of divergence from the standard language, however that is defined, and by investigating this contrast researchers can focus on theoretical and/or pedagogical issues while educators can concentrate on the needs of learners (Pravec 2002). Granger (1998: 6) further argues that by examining a corpus annotated for error, researchers have access not only to learner errors, but also to learners’ total interlanguage (IL). Granger (2002) notes too, that approaching learner errors through learner corpora or language output rather than through experimental processes or self-reports is more reliable because this way, we can have a wider understanding of learner language and so attention can focus on more generalized errors.

Cobb (2003) affirms the significant role that learner corpora can play in second language acquisition even for the advanced user of English. He explains that; “there are practical benefits to having a better description of advanced IL. Until recently the main advice that language teachers could give advanced learners was “to get lots of practice,” assuming there was little more that focused instruction could do for them” (394f.). He goes on to say that, “these learners have become, as it were, defective native speakers, working with restricted linguistic resources and dealing

with a remainder of highly individual random errors” (ibid.). But, according to him, “if, instead, advanced learners are seen as learners nonetheless, moving systematically through acquisition sequences and overcoming shared misconceptions about the L2, then instruction can be focused more effectively throughout the learning process” (ibid.). Cobb further notes that owing to the availability of the learner corpora, the shortage of data is no longer a problem, and even the shortage of theory which remains will be resolved in tandem, because with the accumulation and study of advanced learner data, hypotheses and theories will emerge.

In addition to the production of corpora by learners, Aston et al. (2004) note that corpora interact with learners in two further ways: learners may be authors of corpora, as discussed above; learners may be beneficiaries of corpora; and learners may be users of corpora. Learners benefit from corpora through the interventions of reference book writers or curriculum designers. For example, the COBUILD series of dictionaries and grammar books is based on the Bank of English corpus. This type of study serves a primary pedagogical purpose. Finally, learners may work with and explore corpora by themselves. A term which is commonly used for this type of interaction between the learner and the corpus is Data Driven Learning (DDL). Römer (2010: 20) describes DDL as that teaching approach which puts the learner at centre stage and makes them a linguistic researcher, an interesting analogy originated by Johns (1991) who himself defines DDL as “the attempt to cut out the middleman as far as possible and to give the learner direct access to the data” (30). This means that learning can occur without the teacher and without the teaching of overt rules. Rather, the learners can explore the data/corpus by themselves and identify patterns of usage from multiple language samples. Therefore, corpora, including but not restricted to learner corpora, have made significant impact on the teaching and learning of languages.

Despite the importance of the learner corpora, only a handful exists (e.g. the Learning Prosody in a Foreign Language (LeaP) corpus; Gut 2009), and ICLE (Granger et al. 2009). Most of these do not contain samples of learner language in Nigeria and Cameroon or other African learner Englishes.

In this chapter, the learner corpus – Conacell – is introduced and the process of its compilation fully described to encourage its use and justify the findings of studies based on it. As well as creating awareness of the existence of the corpus, the chapter presents applications of the corpus in the study of lexical form and tense usage competence of learners in the SLA contexts of Nigeria and Cameroon. The remaining part of the chapter is divided into two sections. In the next Sections (2.1 to 2.5), methodological information about Conacell is provided while in the following Sections (3.1 and 3.2), the applications of the resource to investigate lexical form and tense usages by learners in Nigeria and Cameroon are shown, thereby demonstrating its usefulness.

2. Method and data

This section provides methodological information about Conacell: its design, compilation, structure as well as its challenges.

2.1 Purpose of Conacell

Conacell is a learner corpus project which set out to compile and analyse a corpus of Nigerian and Cameroonian English Learner Language, thus the acronym, Conacell. The three-year project, which was funded by The Volkswagen Foundation, Germany, has built and annotated for errors a corpus of 442,939 words. The project is motivated by two major factors: the non-existence of corpus-based studies on learner language in Nigeria and Cameroon, and the need for a systematic characterization of the types, features and patterns of students' language production in the countries. The aim of the project is to produce resources for measuring learners' development as well as for showing the current state of their learning. The corpus data comprises 998 students: 383 university learners, who produced texts amounting to 224,927 words and 615 secondary students who produced texts amounting to 218,012 words. The Nigerian sub-corpus comprises 285,456 words and represents 61% of the entire corpus while the Cameroon sub-corpus comprises 157,497 words and represents 39% of the entire corpus. The corpus has been annotated for syntactic, morphological and spelling errors and can be explored through contrastive interlanguage analyses for language teaching and learning in the various contexts. More information on the annotation scheme is provided in Section 2.4.

2.2 Design of Conacell

At the initial stage of the design of Conacell four important decisions had to be taken; these were questions relating to: (1) Size, what proportion of language would be adequate for this corpus? (2) Content, what kinds of text should be included in the corpus? (3) Sampling, how can the sample account for every part of the population? (4) Representativeness, will the corpus adequately represent the population it was drawn from? These questions have been thoroughly debated in the corpus-linguistics literature, and certain guiding principles have emerged, but no strict rule of thumb exists. Generally, however, the design criteria of Conacell are akin to those of ICLE, with respect to learner variables – age, gender, mother tongue and the assigned task – and other criteria such as that it should comprise continuous discourse, consisting of open ended data e.g. essays, as well as being stored electronically. All four issues (listed above), which were carefully considered in the design of Conacell, are now discussed.

2.2.1 *Size*

The decision on the size of the corpus was determined mainly by the amount of time available since the project was sponsored and had a limited turn-around time of three years. Therefore, intuitively, the researcher aimed to collect and fully annotate a 300,000 words corpus of learner language. However, the final size of the corpus became approximately 443,000 words. In comparison to ICLE where each sub-corpus is approximately 200,000 words, the size of the corpus for this study can be judged to be quite adequate.

2.2.2 *Content*

In considering content, the questions as to what texts to select, in what form, and when and from where to select them (Sinclair 1991; Stubbs 1996) were taken into consideration. Since this is a learner corpus, the texts must be produced by learners of English in the two countries, so we collected essay and examination texts in controlled environments; under the supervision of instructors. Most of the data comprise essays which were written specifically for the purpose of corpus collection. The production of the essay texts was timed as were the examinations. As the study is interested in the measurement of learning and in describing what is currently known to inform teaching and the development of teaching materials, the most recent productions of the students were used; these were texts by the students written as part of the corpus collection procedure in 2013 in both Nigeria and Cameroon.

2.2.3 *Balance*

To balance the data presented a real challenge. As Esimaje and Hunston (this volume) explain, the need for balance arises where a corpus consists of different sub-corpora or parts, it may then be necessary to achieve balance between the parts. This means the attempt to make a text sample of a longer text to “contain proportionately the same number and distribution of examples of particular items as the total population” (McEnery & Wilson 1996: 66); such that what holds for the population also holds for the sample. Biber’s (1993) suggestion of achieving balance is the triad processes of defining the limits of the population, that is, establishing the sampling frame (the entire population of texts from which the samples will be taken), determining the hierarchical structure of the population (i.e. the text genres), and deciding the sample sizes in terms of length and number of texts. Although this is a reasonable suggestion, it could not be applied in this project due to a number of issues: some political, some ethical and some legal. For instance, some institutions in Cameroon considered it unethical to release their students’ scripts whereas in the case of Nigeria, political crises in the North inhibited access to some areas. Therefore, the study adopted Hunston’s (2002: 28) suggestion to include all available data, covering many settings and varieties, without any attempt

at balance between text types. This provides justification for adopting this opportunistic data collection strategy. See Tables 1 to 3 for the structure of the corpus.

2.2.4 *Representativeness*

Another main challenge was to build a representative sample for the research, given the extensiveness of the research population, especially in Nigeria. In the attempt to achieve representativeness, therefore, the data collection points were subdivided according to existing recognised geo-political stratifications. These comprised six zones in the case of Nigeria: South South (6 States, namely, Bayelsa, Akwa Ibom, Edo, Rivers, Cross River and Delta); South West (6 States, namely, Oyo, Ekiti, Osun, Ondo, Lagos and Ogun); South East (5 States, namely, Enugu, Imo, Ebonyi, Abia and Anambra); North West (7 States, namely, Zamfara, Sokoto, Kaduna, Kebbi, Katsina, Kano and Jigawa); North Central (7 States, namely, Niger, Kogi, Benue, Plateau, Nassarawa, Kwara and FCT); and North East (6 States, namely, Bauchi, Borno, Taraba, Adamawa, Gombe and Yobe), and then two zones in the case of Anglophone Cameroon: North West and South West. Further, the sample population were divided according to educational levels: advanced learners, that is university learners in the third and fourth years, and intermediate learners, i.e. students in secondary schools who were in their final years of study.

2.3 Data collection

The first stage in the compilation of Conacell entailed the collection of language productions of intermediate and advanced learners of English in both countries. 'Intermediate learners' refer to final year students at the secondary level of education, aged 16–17, while 'advanced learners' are university students in their pre-final and final years of study, aged 18–28. The data elicited from these learners was in the form of essays and examinations. All the data was collected in its original form, i.e. the essays were hand written. The decision to collect written rather than oral data was informed by the interest of the research in writing because it is a problematic area for English as Second Language (ESL) learners but also because written texts are generally used in high-stakes assessment in academic environments whereas oral texts are less consequential. To date, the collection of the 442,939 words has been completed and the data annotated. As noted above, to ensure representativeness, the data was collected in six different geopolitical zones of Nigeria and two Anglophone zones of Cameroon and approximately the same amount of data was collected from each of them.

2.4 Corpus annotation

The corpus was annotated using Pacx (Platform for annotating corpora in XML; Voormann & Gut 2008; Wunder, Voormann & Gut 2010). XML is one of the main-stream data formats used in corpus data management. The XML data format is suitable for the integration of raw data as well as linguistic and non-linguistic annotations (metadata). Some of the features of Pacx which make it suitable for the Conacell project are that it enables collection and annotation of large amounts of language data (audio, video, texts), based on Eclipse <www.eclipse.org>, reduces annotation errors drastically, offers search options and easy project compilation, distribution and archiving. Corpus building in Pacx entails five main processes as shown graphically in Figure 1. Their applications in Conacell are described below.

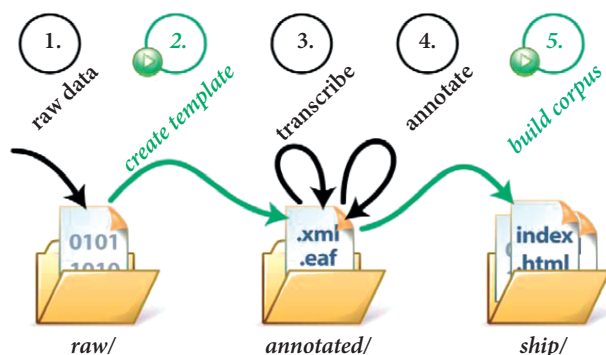


Figure 1. Main processes of corpus building in Pacx (cf. Wunder, Voormann & Gut 2010: 79)

Step 1. Collecting the raw data

A total of 442,939 words in samples of writing by 998 university and secondary students were collected. These were essays and examinations written by the students in 2013. All data was collected in their original handwritten forms. All the scripts were scanned and saved in pdf files. This was done by using an online resource, for example small pdf.com, to convert a jpg file into pdf. The data was subsequently transcribed orthographically and stored in the computer as xml files. Both files, hand written and transcribed, can be called up simultaneously. The data collection, besides being challenging due to the vastness and diversity of the population, and time consuming because essays were handwritten and collected for this specific purpose, was fraught with many other difficulties. For instance, many institutions were unwilling to grant access to students' scripts and due to the political situation of insurgency in some parts of Nigeria some regions were inaccessible to the researcher. This meant that third parties had to be used to collect some data. This

was part of the reason that data balance could not be achieved and we were left with no other choice than to work with all available data.

Step 2. Loading files and creating templates

To load a raw file (scan of an original document) into Pacx, the program first creates a folder named 'raw folder'. The created pdf files were named explicitly to explain their contents e.g. NGunivNW_001 which means data number one from a university student in the north-west of Nigeria. Once the pdfs of all the files were ready, they were dragged or copied and pasted into the folder already created for them. Then Pacx automatically creates the template for the raw files which were to be annotated in the annotated folder.

Step 3. Transcribing the data

All collected data was transcribed orthographically into the template files that are created by Pacx. To ensure the purity of the data is maintained, the typed data was aligned with the original data (the scans) to identify mistypes. This was done after at least two rounds of proof reading. Then we generated metadata from the original data. This is information about the period the data relate to, the place they were collected from, the date, the age and gender of the subjects and any other information that is important in the consideration of the data. An example of transcribed data is seen in Figure 2; the corresponding original handwritten text is shown in Figure 3.

The most Embarrassing Expirience in My Life

This is the most profound, intense painfull and at the same time embarrassing expirience i ever had, i am talking of the day i found out who my true parents are, my identity, My roots

it was more of a public show to say the least and I was just attended puberty, to say the truth, it was not some thing i was expecting i was contented where i was and still i am because I never left. But how could I live the parents that cared for me for over sixteen years how could I have left the brothers and sisters with whom i played, what on earth could have made me think this family that respected my opinion was not my root, how could i ever think that i could not be identified with this god fearing family. Be that as it may, i guess your root is where ever you want it to be.

We were happy as a family,i was contented as their own untill the ups and downs in the family resulted to the option of a divorce. But that does not speak bad of them i guess for every family has it pradicaments. My foster parents if I may say were kind of tired of themselves which they reasoned out divorce was the best for them.thg But as the divorce was been signed in court it was dicided that the children who were not up to twenty one was to be with the wife while those above were to decide for themselves. The We are seven of us and only four waswere to

decide for themselves and that's when the show came up. My father, the one I know, ~~he~~ accepted ~~an~~ the verdict with the condition that I was coming along with him, the Judge could not understand neither could I, I was seventeen and the judge was the one to make my decision for me. We were then adjourned just for me to come at the next sitting to hear the unpredictable, mother never told me, my eldest sister never did, I myself never wondered why my name was a mixture of two regions in which both my parents came from only one. The divorce pledge from my father to live with him made me understand I had been adopted. ~~He~~ The wife of his childhood friend who was short with him right beside him by thieves in Nigeria died while giving birth to me. That was the day my never ending pain began the day my seraphic life ended, the day my ever promising life looked so distant, complex and complicated than I ever dreamt. I had been the most embarrassing experience.

Figure 2. Sample of transcribed text

The Most Embarrassing Experience in My Life

This is the most profound, intense painful and at the same time embarrassing experience I ever had. I am talking of the day I found out who my true parents are, my identity, my roots. It was more of a public show to say the least ~~and~~ I ~~was~~ just attended puberty, to say the truth, it was not something I was expecting. I was contented where I was and still I am because I never left. But how could I leave the parents that cared for me for over sixteen years, how could I have left the brothers and sisters with whom I played, what on earth could have made me think this family that respected my opinion was not my root? How could I ever think that I could not be identified with this God-fearing family. Be that as it may, I guess your root is where ever you want it to be. I grew as a family. I

Figure 3. Sample of original handwritten text

Step 4. *Annotating the data*

In annotating the data, the first decision was to draw up the annotation taxonomy incorporating the types of errors we wished to target. The decision as to what errors to target was partly inductive; based on observations from the data and partly deductive, based on the findings of previous research on learner errors in Nigeria and Cameroon (see e.g. Ojetunde 2013; Alufohai 2016 in the case of Nigeria and Ayafor 2015; Njwe 2016 for Cameroon). The next thing was the creation of labels for identifying the errors in Pacx. The labels were used to build a list which pops up as a dialogue box and prompts the annotator to select a label for an identified error. The labels were coded in colours, which means that each error type was assigned a particular colour for easy identification. The annotation schema contained such labels as spelling (red), plural (brown), tense (purple), agreement (orange), article (blue), subject (green) etc. The annotation scheme is shown in the Appendix. Thus, the annotator can select any label that matches the particular error made, then Pacx marks it and prompts the annotator to insert the correct option or simply to categorise it. The correction option was enabled for only three types of errors: punctuation, spelling, and verb. These were identified as major issues in the learners' writing. Thus, the option enables a deeper insight into the error by providing details about its occurrence (Figure 4 shows what the labels look like). The annotation was done by language experts and each file was annotated at least twice by two independent annotators. At the level of annotation, two main issues arose: first, the recruitment of committed annotators was difficult; and second, there was the problem of what spelling standard to use: British or American. Although the annotators were mainly Nigerians who might be presumed to be using standard Nigerian English (which uses British English spelling), sometimes the Oxford English Dictionary Online, using British rather than Nigerian English, was consulted where the meaning of a word was unclear. Therefore, the problem of standard remains unsolved. A sample of annotated text is shown in Figure 4.

As can be seen from this screenshot (Figure 4) of the annotated page of the data from a Nigerian student, both linguistic and non-linguistic information (header on top) are shown. For the non-linguistic or metadata, for instance, information regarding the region where the data was collected, the age of the writer, the sex and the level of education are all provided. Moreover, with regard to linguistic information, the syntactic features targeted for annotation can be seen clearly in their different annotation colours. For example, while spelling errors are marked in red, tense errors are marked in purple.

Table 2. Structure of Conacell: Metadata – Nigeria

Population/ Region	Number of students		Data size		Number of males		Number of females		Number of unspecified sex		Age range	
	Sec	Uni	Sec	Uni	Sec	Uni	Sec	Uni	Sec	Uni	Sec	Uni
North Central	75	69	18,652	27,252	60	21	15	48			16–17	18–28
North East	75	24	19,747	20,219	53		22		24		16–17	18–28
North West	44	50	20,803	22,700	23	20	21	30			16–17	18–28
South East	74	37	26,021	31,306	31	11	22	12	20	04	16–17	18–28
South South	83	31	22,532	31,268	19	06	64	25			16–17	18–28
South West	76	24	24,058	20,884	31	03	45	14	07		16–17	18–28
Total	427	235	131,813	126,377	217	61	189	129	20	35		

Table 3. Structure of Conacell: Metadata – Cameroon

Population/ Region	Number of students	Data size	Number of males	Number of females	Neutral sex	Age
Secondary						
North West	95	46,040	30	65	–	16–17
South West	92	42,897	21	71	–	16–17
Total	187	88,937	51	136	–	
University						
University 1	50	25,248	13	37	–	18–28
University 2	98	46,171	13	85	–	18–28
Total	148	71,419	26	123	–	

3. Conacell-based studies

Following the completion of Conacell, a number of studies have been carried out. Summaries of two of the studies are represented below to illustrate the possible applications of the resource and to call attention to it for further exploitation. In the first case study, the stages of analysis are fully shown to lead the reader who is a newcomer to corpus linguistics through the process of corpus analysis, whereas in the second case study only a descriptive analysis and results are provided.

3.1 Case study 1 (Esimaje 2017)

3.1.1 Purpose of study

The purpose of this study was to assess the acquisition of formal aspects of lexical items, particularly spelling and capitalisation, by 383 university students in Nigeria and Cameroon. The study attempted to answer two research questions: 1. Do university students in Nigeria and Cameroon demonstrate sufficient competence in English spelling forms? 2. Are there specific patterns of spelling in the various contexts of use?

3.1.2 Data and analysis

The study analysed spelling usage by Nigerian and Cameroonian students. A sample of annotated essay written by a university student from Nigeria is shown in Figure 5. The annotations appear in different colours: 'red' stands for spelling errors. This is followed by an annotated sample essay from a Cameroonian learner in Figure 6. It will be noted that the 'spelling error' tag incorporates errors in the use/omission of capital letters as well as errors conventionally considered under the heading of spelling.

The **Embressing** Day of My Life

An adage says that, **He** that **have** life has hope, although my **embressing** made me think in **[det]** different way but I was consoled with this adage. It was **immediately** after my 400 hundred level **second** first semester resumption. On 15th January 2013, **on the early hours of** **wednesday** morning (10.00 am) **I** got up from my bed and after my normal morning activities, I set out for my morning **Pictures**, which was my first class for the year and also for my new level.

I was going to class that **they** day and was so happy because it was my first day in class as a final student of Ahmadu Bello **university**, **zaria**. I was in the **mist** of my friends and we were gisting about our last holiday, **[sub]** were **beamed** **walked** down to the gate of my hostel which **is** Ribadu hostel. We were asked to show our identity **card** cards, of which we all showed, after that I was asked to **go and** wait **I** stood there, but before then, I **give** my friends my bag to help me catch **[det]** **ait** in class and that they should not wait for me.

However, I did not know that, that was **my** the starting day of life. When **thi** I made an attempt to leave because the woman was attending to others, she told me that I should get my identity card and I told her I wasn't with it. The woman said that why **am** I on short skirt, I looked at myself and I told her, ma, please **[sub]** **am** late for my lectures **I** **pis**, **can** I go and change the cloth **I** as a being hate anything that will cause trouble or **I** being **[det]** centre of attraction. When **I** I was there many students were passing, even my **course-mates** and friends.

Not too long, I started pleading with the woman, she started shouting at me, calling me all **sort** of **name** and even involved my family in the case. Hot tears started rolling down my cheeks. People were passing and **sterling** at me. The worst of it **is** that she told me that she **will** take me to the **head quarter** for **[det]** **snap shot** and that **means** I **will** be appearing in the school bulletin. I started crying begging one man that I saw passing to help plead on my behalf, although in my sight my skirt **is** okay but the security woman there **was** saying **it** was short.

Figure 5. Annotated text: Nigerian university student writing

At first glance, items marked in red, which represent spelling mistakes, seem to be in the majority, creating an immediate impression that spelling errors predominate. However, not all spelling errors would be considered in the analysis because a particular spelling form needs to have a particular frequency to be considered systematic. In this study, the decision was made to set a threshold of four occurrences as a significant frequency of spelling errors. Therefore, a spelling form needs

An **<embarrassment>** is a situation in which you **<fine>** yourself which you less expected. This situation is always a bad one which most at times **<disgraces>** you and **<in>** public. I have had many of such challenges in life but below, we shall see the most **<embarrassing>** one **<we>** which made me almost **<lost>** my character trade of contentment and satisfaction.

On this particular day, we were **<suppose>** to have a contest for the miss of my village. I knew I could make it because I had all that was required from the **<delegates>** candidates. This particular contest, I saw it as a **<way>** **<in>** step into my career of modeling **<but>** the **<embarrassment>** I **<recieved>** made me **<lost>** that dream till today. I prepared so well that I knew nobody could meet up with me **<but>** just a small thing spoiled everything.

The night before the contest, I had invited all my friends and **<talk>** to them about it with **<elot>** of confidence. There was one thing which made me mad. I had my best dress prepared for this day but I had no shoes, not even a sister that I could get one from. I had to borrow one from my neighbour which was really nice and I practiced with it the whole night and everything was fine.

On **<The>** set day, I **<mad>** met with the other girls, we discussed and shared ideas and **<every-body>** was set to go **<in>** when time came **<Every-body>** was to enter when your name **<is>** called. The judge called about five names and they performed so well, I could hear the crowd shouting from where I was standing. I was the next person he called. I **<move>** in but I felt some strange behaviour with the shoe I had **<practice>** with. I struggled **<he>** to **<elot>** walk with it but behold when I got to the center of the stage, my heels came out. I was **<domfounded>**, I could not take any further step but the judge told me to continue. **<with>** the **<childings>** I **<recieved>** from the crowd I was **<confuse>** and more than the word **<embarrassed>**. I **<look>** at the other contestants who had already talked **<on>** my right, they were all laughing, deep in me I knew I had already lost it. I **<strugled>** to talk but all **<what>** I prepared **<out>** of the embarrassment had left my head. I could hardly say out anything clear so I was asked to go. **<while>** leaving the stage I was called to come and pick my broken heel along. **<with>** all the shame I took **<it>** out with me.

When I got home that day I could not even explain to anyone what happened, when I **<open>** my mouth tears **<while>** will drop from my eyes and I will stop. The other girls came to our **<housemy>** home that day, I **<thought>** they were coming to console me but unfortunately for me **<They>** came to mock me and left. Even my own family members could not encourage me. I was all left alone.

Figure 6. Annotated text: Cameroonian university student writing

to occur a minimum of four times to be counted. With this restriction, AntConc was used to identify the number of misspelled words in the Nigeria and Cameroon sub-corpora which had a raw frequency of 4 (our threshold) and above. In total, 2,996 spelling errors were found in the Nigerian sub-corpus and 1,070 in the Cameroon sub-corpus. Further, AntConc enabled the identification of word types, which prevented the bias that the use of token frequency would have introduced in the results. Imagine, for instance, that of the 2,996 spelling errors from the Nigeria sub-corpus, a particular word was wrongly spelled by all the students and by each person five times. This would amount to a raw frequency count of 1,175 (235×5) and would give the impression of a high number of errors whereas there is only one error repeated many times. Therefore, it was necessary to identify the word types in order to show the particular words misspelled as well as the frequency of each misspelling. For the Nigeria sub-corpus, a total of 263 word types were recognized while in the Cameroon sub-corpus, a total number of 107 word types were found. The types were not lemmatized, hence different word forms were counted separately. For each word type, the tokens or occurrences are shown in brackets following the word. These became the specific data for analysis. Although the analysis considered all words with up to four frequencies, due to space consideration, only those words with frequencies of 10 and above are shown in Tables 4 and 5 and the patterns of spelling found are shown in Tables 6 and 7.

Table 4. Words misspelled by Nigerian university students

Word types and frequency (10+)
in (140), the (114), I (67), my (59), he (50), when, (44), it (40), self (40), that (37), every (36), not (36), so (35), can (32), mr, (32), this (32), where (32), they (30), cloth (27), to (27), we (26), writing (26), there (25), faithful (24), father (24), were (24), because (23), ladies (23), land (23), thing (23), information (22), their (22), on (20), one (20), may (19), what (18), body (16), his (16), Its (16), of (16), any (15), chairman (15), embarrassing (27), if (15), she (15), some (15), student (15), female (14), with (14), writer (14), day (13), friend (13), in fact (13), living (13), mrs (13), these (13), you (13), and (12), for (12), important (12), most (12), no (12), out (12), parents (12), white (12), death (11), husband (11), king (11), kings (11), live (11), ther (11), while (11), co-debator (10), report (10), embarrassed (10), gentlemen (10), him (10), man (10), mother (10), school (10), time (10).

Word types = 263 Raw frequency (4+) = 2996 Normalised frequency = 237 per 10,000 words
 Sample size = 235 Data size = 126,377

Table 5. Words misspelled by Cameroonian university students

Word types and frequency (10+)
I (80), every (45), infront (38), in (36), my (34), embarrassing (32), body (30), one (27), where (25), it (24), self (24), the (22), faithful (20), no (20), in (19), thing (17), some (17), ashamed (16), embarrassed (16), on (13), live (12), so (12), alot (11), this (11), any (10), just (10), embarrass (10), to (10), occasion (10), until (10), were (10), what (10), when (10).

Word types = 107 Raw frequency (4+) = 1070 Normalised frequency = 149.8 per 10,000 words
 Sample size = 148 Data size = 71,419

3.1.3 Results

Table 6. Patterns of misspelling found in Nigeria

Words completely misspelled	Target words	Faulty separation/ amalgamation of words	Capitalisation errors
embarrassed (10)	Embarrassed	every thing	in, the, i, is, my, he, when, it, that, every,
embarassing (12)	embarrassing	for	not, so, can, mr, this, where, they, cloth,
[embracing (9),	embarrassing	everything	there, faithful, father, were, because,
Embrassing (7)]	embarrassing	every body	ladies, land, thing, information, their,
embarassment (4)	embarrassment	for	on, one, may, what, body, his, its, of, any,
becouse (4)	because	everybody	she, chairman, if, embarrassing, some,
commitee (23)	committee		student, female, with, day, friend, living,
writter (14)	writer		mrs, these, you, and, for, important,
writting (26)	writing		ur, most, no, out, parents, white, death,
			husband, king, kings, live, while.

Table 7. Patterns of misspelling found in Cameroon

Words completely misspelled	Target words	Faulty separation/ amalgamation of words	Capitalisation errors
imediately (4)	immediately	my self	a lot gentlemen, him, man, mother,
resoure (8)	resource	for	at, report, school, time, after, data,
distructive (4)	destructive	myself	interested, like, are, listeners, name,
clarity (5)	clarity		parent, questioners, then, who, an, work,
questioneers (6)	questionnaires	infact	a, african, audience, blacks, by, children,
ther (9)	there	for	from, had, her, ideas, microsoft, once,
nt (11)	not	in fact	people, research, be, since, them, until,
untill (8)	until		want, will, city, cloths, dad, don't, ever,
its (16)	it's	alot	first, into, later, men, other, peoples,
hight (4)	height	for	secondary, state, through, but, child.

3.1.4 Discussion of results

3.1.5 Spelling error taxonomy

The first stage of the analysis was the categorization of the spelling forms into three: completely misspelled words including omission and addition of letters e.g. untill, hight, clarity; errors where either faulty amalgamation or separation of words occurred (e.g. infact, my self, alot); and capitalisation errors which shows either the use of capitalisation where not necessary or lack of use where necessary (e.g. i, mrs.). A few examples are shown in Tables 6 and 7 above. At first glance, it is noticeable that the misspelled words comprise both function and content words. While all three patterns of misspelling occur, those involving capitalisation are the majority cases and a number of words are found to be commonly misspelled in both geographic contexts (see emboldened words in Table 5 above).

3.1.6 Frequency

The second study to be made is the frequency of occurrences of the misspellings. As can be seen from Tables 4 and 5, the words with the highest frequencies of error among Nigerian university writers are **in** (140), **the** (114), **I** (67) while in the case of Cameroonian university students, the three most frequent error types are **i** (80), **every** (45), **infront** (38). From Table 4, out of the 81 word types with frequency of 10 and above used by Nigerian students, 31 are content words while in Table 5, the 33 word types used by Cameroonian students show 13 of them to be content words. The distinction between the frequencies of function and content words misspelled is, therefore, insignificant; both are subject to misspelling. However, the form can be a distinguishing factor. For example, most of the function word misspellings are capitalisation errors while in the case of content words some are

capitalisation errors (e.g. African), and some amalgamation/faulty separation errors (e.g. a lot) others are totally misspelled (e.g. questioners). It would seem that Nigerian students committed more errors than Cameroonian students. To verify this, the raw frequencies had to be converted to relative frequencies. This was done by calculating the frequency of the errors per 10, 000 words. As the result shows, normalised frequency of misspelling in the Nigeria sub-corpus is 237 while that of Cameroon is 149.8. This means that out of every 1000 words written by a Nigerian learner of English, about 23 are likely to be wrongly spelt while for the Cameroon learner, 15 words are wrongly spelt. This translates to approximately 2% and 1.5% respectively for Nigeria and Cameroon. Therefore, the assumption stands correct. Interestingly, in both the contexts of Nigeria and Cameroon, certain words posed equal difficulty in spelling. For instance, the personal pronoun 'I', and the content word 'embarrassing' occurred with incorrect spellings in both contexts. The personal pronoun 'I' was further analysed in order to create more insight into its usage.

3.1.7 *Context analysis: Concordance of 'I'*

Examining the contexts of 'I' underlined the role of environment in spelling as well as leading to a hypothesis about the reason for the wrong spelling form. The choice of 'I' is informed by its very high frequency in both sub-corpora, in the Nigeria sub-corpus it ranked 3rd with a frequency of 67 and in the Cameroon sub-corpus, it ranked first with a frequency of 80. The sample analysis focused on the Cameroon texts. To do so, a concordance analysis was done for the word. As Sinclair (1991) explains, a concordance lists the occurrences of a word-form in its own textual environment thus enabling an examination of the word while providing its co-occurrence information. Using the KWIC (key word in context) method, the word was examined to identify its patterns of occurrence and to see if or how its environments determined its form. In the Cameroon sub-corpus, 80 instances of misspellings occurred in the use of the personal pronoun – 'I'. In all the cases, the students did not present the word in capital form as they should. Example (1) illustrates this:

- (1) 'When we finished, i hastily took a bath as i was getting set to go to church'.

The concordance analysis showed a total of 3601 instances of the personal pronoun I in a corpus of approximately 71,000 words written by 148 students. The first impression from these numbers may be that 80 wrong forms out of 3,601 uses of a linguistic item are not a cause for concern. However, the concordance further showed the contexts of occurrence and the sources of the spelling forms (i.e. the particular scripts/students) to add some details about the occurrences and enable some interesting revelations. It was shown, for instance, that three texts/

scripts accounted for a total of 56 of the errors. Moreover, as the concordance analysis also showed, a total of 16 students were responsible for the 80 misspelling forms found. In terms of the environment, a good number of the forms occurred in sentence-medial positions, as well as in many cases in subordinate clauses and thus complex sentences but especially when the clause occurred in the second part of the sentence rather than the first part. In other cases, no clear pattern was discernible and there were cases where a student used both the correct and wrong forms in a single sentence. In many cases where this happened, the student used the correct form at sentence-initial position and the wrong form in medial position. A few examples (see Figure 7) will be necessary to support the observations made and the conclusions which follow. The examples are from the Cameroon sub-corpus but the same observations were made of the Nigeria sub-corpus.

The screenshot displays the AntConc 3.4.1w (Windows) 2014 interface. The 'Corpus Files' list on the left contains 30 files from CMU01_006.txt to CMU01_030.txt. The 'Concordance Hits' window shows 3601 results for the search term 'i'. The results are displayed in a table with columns for Hit, KWIC, and File. The KWIC column shows the context of the word 'i' in various sentences. The 'Search Term' is 'i', and the 'Search Window Size' is set to 50. The 'Kwic Sort' options are checked for Level 1 (1R), Level 2 (2R), and Level 3 (3R). The 'Total No.' of hits is 148, and the 'Files Processed' count is 30.

Hit	KWIC	File
427	the truth, it was not some thing i was expecting i was content	CMU01_018
428	t some thing i was expecting i was contented where i was and	CMU01_018
429	xpecting i was contented where i was and still i am because I	CMU01_018
430	contented where i was and still i am because I never left. But	CMU01_018
431	ere i was and still i am because I never left. But how could I li	CMU01_018
432	se I never left. But how could I live the parents that cared for r	CMU01_018
433	over sixteen years how could I have left the brothers and siste	CMU01_018
434	thers and sisters with whom i played, what on earth could ha	CMU01_018
435	ion was not my root, how could i ever think that i could not be	CMU01_018
436	t, how could i ever think that i could not be identified with thi	CMU01_018
437	ring family. Be that as it may, i guess your root is where ever y	CMU01_018
438	pe. We were happy as a family, i was contented as their own un	CMU01_018
439	hat does not speak bad of them i guess for every family has it	CMU01_018
440	addicaments. My foster parents if I may say were kind of tired c	CMU01_018

Figure 7. Concordance of i

The study of spelling concludes that such errors at 2% and 1.5% respectively are negligible and as such do not pose a significant threat to learning and communication at that level. Also, the results of the investigation do not reveal any peculiar spelling patterns used by the students in Nigeria and Cameroon and so the second research question of whether there are patterns of spelling specific to the learning

contexts could not be confirmed. However, the high occurrence of capitalisation errors called for more analysis of the spelling form to identify patterns and to suggest possible remedies. The results showed that the instances where the learners used capitalisation where it was not required far outweighed instances where it was omitted where it was required. The implication is that competence in spelling form can be measured along educational levels and across educational contexts. Relative to the sample sizes of each of the corpora, it can be estimated that the extent of spelling errors is the same in both countries.

3.2 Case study 2 (Esimaje, A. U. & Hunston, S. E. 2016)

3.2.1 *Purpose of the study*

The purpose of the study was to examine tense uses by university students in Nigeria in order to identify and quantify non-standard vs standard features of tense uses in the language production of 112 university students. To achieve its objective, the study set out to answer three questions: 1. What is the frequency of tense errors in the language of the students? 2. With what kinds of verbs are these errors made? 3. Do students from different regions have similar frequencies of tense errors?

3.2.2 *Method and analysis*

In the analysis, our interest was only in the use of the present and past tense, so all instances of modals and aspect, which are also tagged as ‘tense’ in the tagger, were identified and excluded from the data. In order to enable identification of particular verb-tense categories, POS tagging of the data was done using the UCREL tagger (CLAWS). This made further subdivision of tense tags possible. Hence, tags were used to identify specific tense uses, such as VBB for base forms of “BE”, VBD for past forms of “BE”, VBZ for -s of “BE”, VDB for base forms of “DO”, VDD for past forms of “DO”, VDZ for -s forms of “DO”, VHB for base forms of “HAVE”, VHD for past forms of “HAVE”, VHZ for -s forms of “HAVE”, VVB for base forms of lexical verb, VVD for past forms of lexical verb, VVZ for -s forms of lexical verb were used to identify specific tense uses.

Using the Pacx software, searches were done to identify instances of wrong tense uses (these are usages that deviate from standard British English which is the model for teaching English in Nigeria), their frequency and patterns. The identified usages were then further examined in a wider context using corpus tools such as a concordancer by applying the keyword in context (KWIC) method. This was done by exporting the results from Pacx to AntConc (Anthony 2014). This enabled the calculation of both the standard tense uses and the non-standard uses, thus allowing the establishment of difference as well as the determination of the extent of the

difference. Since tense is a property of the verb phrase rather than verb form, these were distinguished to ensure the validity of the results. Therefore, whereas a count of verb forms (all instances of verb occurrence) found in the university corpus is 7,076, the number of verb phrases was found to be 5,605. This was estimated by manual checks of the results, discounting double counts returned by the search and removal of 2nd and 3rd verb forms in complex verbs.

3.2.3 *Discussion and conclusion*

There are a total of 5,605 verb phrases in the sub-corpus. Out of this number, the analysis reveals a total of 621 tense errors which represents about 11% of the total verb phrases. It also reveals that the tense forms do not pose equal difficulty to the students because out of the 621 errors found, 594 are of a particular kind: use of the present or unmarked form where the past tense is required. Further, it shows the proportion of errors is different depending on the type of verb carrying the tense. Lexical verbs (i.e. excluding BE, HAVE and DO) which carry tense are the most likely to be used erroneously. On closer examination, it was shown that the frequency of errors also varied along regional lines. For instance, texts from the NW showed most errors in tense marking especially in the two categories of lexical verbs and lexical BE. Texts from NC and SW had the next highest frequency of error and texts from SE and SS showed the least. This difference is statistically significant (Chi-square test $X = 16.420$; $p < 0.01$).

The study revealed that the greatest problem in tense use by Nigerian university learners of English is past tense marking and the most misused forms are shown to be the base form of the lexical verb and the BE forms. It also shows that tense marking is influenced by the learners' geographical region. It should not be forgotten that almost 90% of the verb phrases in the corpus did not show this kind of error. However, even an 11% error rate is likely to be cause for comment, and there is inconsistency between regions, meaning that the rate is substantially higher in some regions of Nigeria. To some extent, then, tense remains a learning need for university students in Nigeria despite 16 years of English medium education. This study has implications for language teaching and curriculum development.

4. Further challenges for Conacell and corpus linguistics in Nigeria and Cameroon

A great number of challenges facing corpus linguistics in especially Nigeria, but also in Cameroon which are articulated by Esimaje (2014: 9ff.) are represented here and examined in the context of current scholarship.

1. **Mode of research:** Language research in the contexts of Nigeria and Cameroon is predominantly done with the conventional method of direct observation of data: mini collections of language sample, which are then typically organised and handled manually (see Adagboyin, Aluya & Edem 2016: 3), to investigate phenomena. The use of computers is mainly for the typing of completed research and sometimes for some processing tasks such as numerical counts. The useful role of the computer to analyse and sort data using relevant software has yet to take root among academics based in the countries of Nigeria and Cameroon (see Nkemleke, this volume). As Adeya and Oyeleran-Oyeyinka (2002) confirm, the internet infrastructure in most African tertiary institutions is still inadequate and where these exist, user levels fall far below those in developed countries and although social media surveys indicate a rise in internet users worldwide, most users in Africa use the internet mainly for emails. Studies such as Agbonlahor (2006) have also identified such other challenges as access to the internet, whether the lack of basic infrastructure or even poor connectivity to the internet as well as training (the know-how of the internet).
2. **Availability of corpora:** There is also the challenge of the availability of the raw materials for computer analysis; the corpora (especially of African Englishes; see Ozon, present volume) and the software. Although, these appear now to be widely available around the globe; some are provided free online while some are available at a price e.g. The Bank of English of 400 million words and the British National corpus of 200 million words, such collections of language in use are conspicuously in short supply in the African linguistic environments: academic institutions, libraries, research institutes, and individual data banks and this may be due to poor funding of higher institutions (see Adegbite 2012: 2 and also Nkemleke 2015: 117 for the challenge with funding).
3. **Attitude to the use of computer technology:** There is yet another problem; the reluctance of Nigerians (in particular) but also Cameroonians to use computer technology for research. As there are more traditional, arm-chair linguists than modern linguists in these contexts, attitude to any computer-based methodology is more or less negative and this has caused a low level acceptance of the field of corpus linguistics and of course corpus data, its resources and applications. For instance, Ehikhamenor (2003) noted that scientists in Nigeria depend heavily on printed information sources and that besides emails very little use is made of other internet facilities; a point confirmed by some other studies such as Anadarajan, Igbaria and Anakwe (2002) who identified social pressure as a major determiner of Internet use in Nigeria. This is why despite the advances world-wide in computer-based language research, language teaching and learning in the African contexts of Nigeria and Cameroon

have remained largely the same and benefitted only little from the advances of computer technology. For this reason, the need for increased awareness of the use and potentials of computers in research and education has been stressed (Isokpehi, Thomas, Ojo, Sofoluwe & Coker 2000).

5. Chapter summary and conclusion

The chapter has introduced a new learner corpus – the corpus of Nigerian and Cameroonian English learner language (Conacell). It has described the process of its compilation, showing the steps undertaken to ensure the reliability of the corpus data and the validity of the results of studies based on it. Furthermore, it has demonstrated the application of the corpus by reports on the investigation of learners' competence with specific regard to the use of tense and spelling. Each of these studies revealed findings that either would have been very difficult to establish or completely overlooked without corpus evidence. For instance, the study on tense revealed not only that tense pedagogy in Nigeria should target the base form of the lexical verb and the BE forms but also that tense errors and geographical region are correlated. Similarly, whereas intuitive evidence would readily point to spelling as a major need of learners in ESL contexts such as Nigeria and Cameroon, the study on spelling revealed the error level of university students to lie at 3.5% and thus be negligible as it did not pose significant threats to communication. This finding was made possible by the authentic data which corpora provide and the corpus search options for 'mining' data for features that are not easily identifiable by the human eye and not easily remembered by the human brain. Therefore, Conacell makes a useful addition to the list of existing learner corpora whose general purposes are to portray the state of learning, to reveal the current needs of the learners, to enable a description of learners' competence at different stages of learning and to guide the production of teaching materials such as curricula and textbooks.

Besides, beyond the immediate significance of Conacell, there is a need to point out the consensus that language research has transited from sole reliance on human knowledge to the use of a wealth of knowledge about language which computer technology grants access to today. As Hunston (2002) notes, the slipperiness and subtlety of language is such that it confounds human intuition. This justifies the solution proffered by Leech (1991) for the partnership of humans and machine in tackling the intractable problems of language. However, to use the computer and the software presupposes, to a great extent, the presence of corpus data. Sinclair had earlier observed the problem of paucity of language data until the coming of

corpus linguistics and corpora. While the problem of data has been significantly solved by the availability of corpora in many of the places where corpus linguistics has taken root like the West, the same cannot be said about the African continent, including Nigeria and Cameroon, for instance. Therefore linguistic research continues to suffer the dearth of data and specifically, authentic data. Conacell and a few other corpora such as ICE Nig and ICE Cam (Greenbaum 1996; Wunder, Voorman & Gut 2010; Nkemleke 2008) and now ICLE (Granger 2009) (although Nigeria and Cameroon are not included yet) are attempts to solve this problem and to position linguists to contribute to the global academic scholarship towards tackling the issues in language studies.

It is my first hope that Conacell will indeed be a useful resource for language teaching, curriculum development, and textbook writing as it directs attention to what learners in Nigeria and Cameroon know and what they do not yet know, as well as facilitate contrastive studies of language learning in the multi-linguistic, multi-cultural contexts of Nigeria and Cameroon. And, it is my second hope that this study will help stimulate the interest of African linguists in corpus linguistics and the use of corpus resources.

Acknowledgements

I would like to thank The Volkswagen Foundation Germany, for funding the Conacell project. I am also grateful to the three reviewers whose comments greatly improved this work.

References

- Adagbonyin, Asomwan S., Isaiah Aluya & Edem, Samuel. 2016. A corpus-based approach to the linguistic features in Nigerian and American presidential speeches. *Journal of Linguistics, Language and Culture* 3(1): 1–27.
- Adegbite, Wale. 2012. The ICE Nigeria corpus as a data base for Nigerian English studies. *ISEL* 10(1): 1–10.
- Adeya, Catherine Nyaki & Oyelaran-Oyeyinka, Banji. 2002. *The Internet in African Universities: Case Studies from Kenya and Nigeria*. Study carried out for the Institute of New Technologies (INTECH), United Nations University. Maastricht: UNU/INTECH.
- Agbonlahor, Rosemary. 2006. Motivation for use of information technology by University Faculty: A developing country perspective. *Journal: Information Development* 22(4): 263. <https://doi.org/10.1177/0266666906072955>
- Alufohai, Peace Joan. 2016. Grammatical errors in written composition of Junior Secondary School Students in Owan West Local Government Area of EDO State. *International Journal of Academic Research in Progressive Education and Development* April 5(2): 61–66.

- Anadarajan, Murugan, Igbaria, Magid & Anakwe, Uzoamaka P. 2002. IT acceptance in a less developed country: A motivation factor perspective. *International Journal of Information Management* 22(1): 47–65. [https://doi.org/10.1016/S0268-4012\(01\)00040-8](https://doi.org/10.1016/S0268-4012(01)00040-8)
- Anthony, Lawrence. 2014. AntConc (Version 3.4.3) [*Computer Software*]. Tokyo: Waseda University. <<http://www.laurenceanthony.net/>>
- Aston, Guy, Bernardini, Silvia & Stewart, Dominic. 2004. Introduction: Ten years of Talc. In *Corpora and Language Learners* [Studies in Corpus Linguistics 17], Guy Aston, Silvia Bernardini & Dominic Stewart (eds), 1–18. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.17>
- Ayafor, Miriam. 2015. *The Grammar Problem in Higher Education in Cameroon*. Newcastle upon Tyne: Cambridge Scholars.
- Biber, Douglas. 1993. Representativeness in corpus design. *Literary & Linguistic Computing* 8: 243–257. <https://doi.org/10.1093/lc/8.4.243>
- Cobb, Tom. 2003. Analyzing late interlanguage with learner corpora: Québec replications of three European studies. *The Canadian Modern Language Review* 59(3): 393–423. <https://doi.org/10.3138/cmlr.59.3.393>
- Ehikhamenor, Fabian A. 2003. Internet facilities: Use and non-use by Nigerian university scientists. *Journal of Information Science* 29(1): 35–48. <https://doi.org/10.1177/016555150302900104>
- Esmaje, Alexandra U. & Hunston, Susan E. (to appear). What is corpus linguistics? In *Corpus Linguistics and African Englishes* [Studies in Corpus Linguistics 88], Alexandra U. Esmaje, Ulrike Gut & Bassey Antia (eds). Amsterdam: John Benjamins. (this volume)
- Esmaje, Alexandra U. 2016. The Corpus of Nigerian and Cameroonian English Learner Language (Conacell).
- Esmaje, Alexandra U. 2014. Language and computers: New directions in language research and teaching. *International Journal of Language Studies* 8(1):1–14.
- Esmaje, Alexandra U. & Hunston, Susan E. 2016. A corpus-based study of tense uses by advanced learners of English in Nigeria. *Research in English and Applied Linguistics REAL* 10: 123–136. Chemnitz: Curvillier.
- Esmaje, Alexandra U. 2017. Comparing spelling usage by advanced learners in Nigeria and Cameroon: A Conacell-based study. *Syllabus Review* 7(1): 168–189.
- Granger, Sylviane, Dagneaux, Estelle, Meunier, Fanny & Paquot, Magali (eds). 2009. *The International Corpus of Learner English Version 2*. Handbook and CD-ROM. Louvain la-Neuve: Presses universitaires de Louvain.
- Granger, Sylviane. 2012. Learner corpora. In *The Encyclopedia of Applied Linguistics*, Carol A. Chapelle (ed.), 3235–3242. Oxford: Wiley-Blackwell.
- Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In *Corpora and Language Teaching* [Studies in Corpus Linguistics 33], Karin Aijmer (ed.), 13–32. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.33.04gra>
- Granger, Sylviane. 2002. A bird's eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* [Language Learning & Language Teaching 6], Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds), 3–37. Amsterdam: John Benjamins. <https://doi.org/10.1075/llt.6.04gra>
- Granger, Sylviane. 1998. *Learner English on Computer*. London: Longman.
- Greenbaum, Sidney. 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

- Gut, Ulrike. 2009. *The LeaP Corpus*. <<http://www.phonetik.uni-freiburg.de/leap/>>
- Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP.
<https://doi.org/10.1017/CBO9781139524773>
- Isokpehi, Raphael, Thomas, B., Ojo, A. et al. 2000. Information technology literacy among Nigerian Microbiology students and professionals. *World Journal of Microbiology and Biotechnology* 16: 423. <https://doi.org/10.1023/A:1008930115533>
- Johns, Tim. 1991. Should you be persuaded: Two samples of data-driven learning materials, Classroom Concordancing. *ELR Journal* (4): 1–16.
- Leech, Geoffrey. 1991. The state-of-the-art in corpus linguistics. In *English Corpus Linguistics, Studies in Honour of Jan Svartvik*, Karin Aijmer & Bengt Altenberg (eds), 8–30. London: Longman.
- Lüdeling, Anke, Walter, Maik, Kroymann, Emil & Adolphs, Peter. 2005. *Multi-level Error Annotation in Learner Corpora*. Berlin: Humboldt-University.
- McEnery, Tony & Wilson, Andrew. 1996/2001. *Corpus Linguistics*. Edinburgh: EUP.
- Myles, Florence. 2005. Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4): 373–391. <https://doi.org/10.1191/0267658305sr2520a>
- Nkemleke, Daniel. 2008. *Manual of Information to Accompany the Corpus of Cameroon English*. Chemnitz: University of Technology.
- Nkemleke, Daniel. 2015. Computer corpora in English language research and pedagogy with reference to the corpus of Cameroon English. *Syllabus Review, Human & Social Sci. Ser.* 6(2): 93–210.
- Njwe, Amah Eyovi. 2016. Morpho-phonological misuse of English: A cause of poor performance amongst Cameroonian Secondary School Students. *Global Journal of Human Social Science: Linguistics & Education* 16(7): 4–18.
- Ojetunde, Cecilia Folasade. 2013. Lexico-grammatical errors in Nigerian English: Implications for Nigerian teachers and learners of English. *European Scientific Journal* 9(17): 252–268.
- Pravec, Norma A. 2002. Survey of learner corpora. *ICAME Journal* 26: 81–114.
- Römer, Ute. 2010. Using general and specialised corpora in English language teaching: Past, present and future. In *Corpus-based Approaches to English Language Teaching*, Mari Carmen Campoy-Cubillo, Begoña Bellés-Fortuño & MaLluisa Gea-Valor (eds), 18–38. London: Continuum.
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
- Stubbs, Mike. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
- Voormann, Holger & Gut, Ulrike. 2008. Pacx Software.
- Wunder, Eva, Voormann, Holger & Gut, Ulrike. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.

Appendix 1. Annotation scheme

Error Taxonomy and Labelling in Pacx – CONACELL Project

Error type	Examples	Pacx annotation label
Spelling and writing		
– wrong forms (1)	hapened (happened); tought	Spelling (type in correct form)
– word splitting (2)	(taught/thought); Fraudstar (fraudster)	Spelling (type in correct form)
– word merging (3)	Myself; Nobody; when ever; after ward	Spelling (type in correct form)
– capitalisation (6)	Didnot; alot; infact; ofcourse	spelling (type in correct form)
– foreign words (11)	avoid this kind of Embarrassment	spelling (type in correct form)
– deleted word	Madame	deleted
– unclear illegible words	mine me	unclear
Punctuation		
– lack of full-stop at end of sentence (6)	I did not cry, so much	punctuation (type in correct form)
– comma omission or wrong (9)		punctuation (type in correct form)
Morphology		
– Tense: wrong form (4, 13)	Lended (lent); swore (swore); becamed (became) staled (stale) felled (fell)	Tense (type in correct form)
– lack of tense marking (4)	He was so shocked and surprise....	Plural
– Pluralisation: lack of it and wrong forms (10)	Six individual (individuals); do my laundries;	
Syntax		
– no subject-verb agreement (5)	I has been (have); ... eyes was (were); Below earth (the earth); ... could not	agreement
– article omission (7)	help issue (the issue)	article
– omission of subject (12)	Actually, am Mohammed by name; when ever have something on your	subject
– wrong use of pronouns (19)	mind....	pronoun
– wrong use of progressive (8)	The boy which	progressive
– wrong complex verbs (14)	I should have being aware that; ... a language which is been spoken; were been victimized	verb
– wrong preposition in phrasal verbs (25)	I spent nearly 20 years seeking for admission; I grew up into a family of three.	preposition
Other		
– indigenous word		indigenous
– colloquial word form	coz for because	coll

Introducing a corpus of English(es) spoken in post-independence Namibia

Insights into corpus design and quantitative analyses

Helene Steigertahl

Bayreuth University

The possibility of an evolving “Namibian English” was already suggested more than 30 years ago (Chamberlain 1981: 46). However, detailed analyses of the English(es) used in Namibia were only recently initiated (e.g. Otaala 2006; Buschfeld & Kautzsch 2014; Kautzsch & Schröder 2016; Steigertahl 2017). The present chapter adds to previous research on morphosyntactic structures, introducing the Corpus of English(es) Spoken by Black Namibians post Independence (ESBNaPI). First, the Namibian linguistic situation will be introduced before procedures of data collection and methodology will be presented. Afterwards, morphosyntactic examples from the corpus will be given and compared to South African English(es) (SAE) to address potential generalizations of post-Independence English(es). The overall goal is to raise awareness of corpus resources in southern Africa.

Introduction

Several works have been published on linguistic fieldwork methods, data collection and analyses, and corpus linguistics in general (e.g. Spradley 1979; Milroy 1980, 1987; Fink & Kosecoff 1985; Kachru 1985; Bogdan & Knopp Biklen 1992; Neuman 1994; Rickford & McNair-Knox 1994; Greenbaum 1996; Nelson 1996; Kennedy 1998; Cukor-Avila & Bailey 2001; Faculty of Education University of Pretoria 2004; Feagin 2004; Tagliamonte 2006; Diekmann 2007; Turner 2010; Meyerhoff et al. 2012; Schilling 2013). However, only scarce attention has been given to compiling corpora in African countries (e.g. Allwood & Hendrikse 2003; Schröder 2004; Wolf 2004; de Klerk 2006; Pienaar & de Klerk 2009; Allwood et al. 2010; van Rooy 2013; Seidel 2015). To counter this, this chapter will discuss the compilation process of the corpus of English(es) Spoken by Black Namibians post Independence (ESBNaPI) and thus add to the description of corpora as well as structures of African Englishes.

The plural term “English(es)” is used as it is not clear (yet) if there is/are one or several varieties of English(es) spoken in post-Independence Namibia, if any at all. It is possible that different Englishes evolve similar to the South African context, based on L1 interference or regional differences. This, however, has not yet been assessed in detail. Besides, *Black* is viewed as a social construct and a self-attributive term in this chapter, never as a biological category. It is only positively connoted, as linked to empowerment, and forms part of one’s identity and identification process. Therefore, Black is always capitalized (cf. Arndt 2005: 343f.; Sow 2008; Arndt 2012: 20f.; Bologna 2013; Deutsches Institut für Menschenrechte 2017). Most significantly, Black is not used to reiterate apartheid categories of South African, South West African and Namibian citizens respectively. Neither does it refer to any shortcomings among Namibians who position themselves as Black. By only giving voice to this group, the present approach tries to distance itself from former studies that have categorized Namibians of European descent separately, into single groups such as “Germans” and “Afrikaners”, and Namibians of African descent into a so-called “indigenous” group, without taking note of ethnic and linguistic differences. This merely *white*¹ and colonial perspective might suggest a homogenous group and does not reflect the reality of a diverse group of speakers with different L1, for instance. Overall, possibilities of both corpora compilations and resources will be addressed in this chapter. Awareness of corpus application will be raised so that insights into corpus analyses are provided, particularly for the southern African region. In doing so, future researchers are given a model for applying these methods to their own corpus compilations.

The chapter is structured as follows: Firstly, the Namibian linguistic situation is outlined, with a focus on English as the sole official language since the country’s independence in 1990. Secondly, the corpus of ESBNaPI, which was constituted in four Namibian communities in 2015, is introduced. In this part, the data collection process as well as the methodology used are presented so as to prepare future linguists who intend to conduct sociolinguistic research in Namibia. Thirdly, potential morphosyntactic forms of the English(es) spoken in Namibia are assessed. A quantitative analysis is applied to see if structures are used systematically, that is if they could be considered “features” instead of “errors”. These grammatical structures are then compared to patterns found in South African English(es) (SAE) in order to examine potential similarities among southern African Englishes, as suggested by Maho (1998: 166–169), Trudgill & Hannah (2002: 28) and Stell (2014: 227), for instance.

1. The term *white* is not capitalized but put into italics to stress the way whiteness is viewed as a so-called ‘normality’ by most Europeans. This raises self-awareness and questions its self-evidence (Arndt & Ofuately-Alazard 2011; Bologna 2013).

The Namibian linguistic situation

Like the majority of African countries, Namibia is a multiethnic state with several languages spoken on a daily basis (Mendelsohn et al. 2009: 159, 164f.; Ejikeme 2011: 8). The country comprises 824,292 square kilometers, which are sparsely inhabited by roughly 2.2 million people (Mendelsohn et al. 2009: 159; Frydman 2011: 181; CIA 2014). The Namibian population is not evenly spread across the country, but predominantly lives in the north, that is in the Omusati, Oshana, Ohangwena and Oshikoto Regions, and in the center, that is, the Khomas Region in and around the capital Windhoek (Möhlig & Seidel 2007: 146f.; Mendelsohn et al. 2009: 8, 161ff.; Harris 2011: 10; Wallace & Kinahan 2011: 2, 75f.). The 2.2 million Namibian inhabitants belong to over 30 different ethnic groups (Fourie 1997: 30; Maho 1998; CIA 2014). Commonly, ethnic and linguistic identity coincide in Namibia. The predominant groups are the Owambos (50% of the population), Kavangos (9%), Hereros (7%) and Himbas (2%), Damaras (7%) and Namas (5%), Zambezians² (4%), Sans (3%), Basters (2%) and Tswanas (0.5%) (Wallace & Kinahan 2011: 11; CIA 2014). These groups are not homogeneous but comprise several sub-groups (cf. Knappert 1981: 143; Maho 1998: x, 222; Malan 2005; Nampala & Shigwedha 2006; Maho 2008; Wallace & Kinahan 2011: 12).

In total, more than 30 languages – names tend to differ between sources – are spoken in Namibia (Fourie 1997: 30; CIA 2014; Lewis et al. 2016; Steigertahl submitted) and the majority of them belongs to the Bantu language family, followed by the Khoe, San and Germanic language family (Pütz 1995b: 158; Fourie 1997: 30; Frydman 2011: 181; Wolff 2016: 211, 304). In addition, the Namibian Sign Language is used. Sources primarily name the L1 used in Namibia while figures on languages used as a second language (L2) do not exist (NSA 2001, 2011; Ejikeme 2011: 39; Lewis et al. 2016). Fourteen of the languages spoken in Namibia are codified and listed as national languages in the Constitution of the Republic of Namibia, according to the latest Population and Housing Census (NSA 2011), that is, Afrikaans, English, German, Ju|'hoansi, OshiKwanyama, OshiNdonga, OtjiHerero, Khoekhoegowab, RuKwangali, Rumanyo, SeTswana, SiLozi, Thimbukushu, and !Xuun (cf. Maho 1998: 196f.; Töttemeyer 2009: 8). These national languages can be used in education, administration and radio broadcasting (Haacke 1994: 242; Fourie 1997: 29; Harlech-Jones 1997: 227; Maho 1998: 196f.; Frydman 2011: 181f.). In contrast to Namibia, the formally multilingual language policy of the neighboring country South Africa lists eleven languages as official and not only national languages: Afrikaans, English, isiNdebele, isiXhosa, isiZulu, sePedi, seSotho, seTswana,

2. The Zambezi Region used to be called Caprivi Strip or Region until 2013 (Smith 2015). This chapter uses the current name although the sources might refer to “Caprivians” still.

siSwati, TshiVenda, and xiTsonga (Government of South Africa 1996: Chapter 1, Section 6 (1)).³ Although the South African Constitution supports multilingualism with this policy, practically, monolingualism is promoted since English is dominating in most public domains, e.g. administration and education (Adegbija 1994: 141; Mazrui 2004: 32; Frydman 2011: 180, 185). In Namibia, English is the sole official language and supposed to be used in all public domains. It was regarded the “language of liberation” in contrast to the official language used during apartheid, that is, Afrikaans, the “language of the oppressor” (Harlech-Jones 1997: 231; Fourie 1997: 30). SWAPO, the former independence movement, known as the South West Africa People’s Organization, which is now the ruling party in Namibia, suggested English as “one” official language in the 1970s (Frydman 2011: 182). Yet, English was implemented as the only official language with independence in 1990. This Indo-European language was chosen because it was considered a “neutral” language that would support unity of the Namibian people and enhance global communication (Haacke 1994: 241ff.; Webb 1994: 188; Pütz 1995b: 160; Fourie 1997: 30; Harlech-Jones 1997: 223, 238; Frydman 2011: 181ff.; Wolff 2016: 250). According to Pütz (1995b), English developed into “a marker of a new anti-colonial identity. Thus, English function[ed] as an instrument of liberation and a symbol of ideological neutrality and nation building” (163) in post-Independence Namibia.

Despite its negative connotation during apartheid, Afrikaans is still the L1 of numerous Damaras and Namas, Rehoboth Basters, Coloreds and *Whites*. Besides, it was the most frequently used lingua franca in the land that is called Namibia today prior to independence (Ejikeme 2011: 15) and still is in central and southern parts of the country (Knappert 1981: 5, 143; Fourie 1995; Töttemeyer 2009: 8; Töttemeyer 2010: 9), whereas Oshiwambo varieties have been used as a lingua franca in the north (Harlech-Jones 1995: 186; Stell 2014: 226; New Era 2015). Today, Afrikaans is characterized as “the main lingua franca in Namibia [with covert prestige], or at least as an important one” (Stell & Groenewald 2016: 2), and has experienced a boom, in the capital among young Namibians at any rate, as stated by UNAM (University of Namibia) professor of German studies, Marianne Zappen-Thomson, at the SAGV Conference in 2015. Then again, it has been argued that English is developing into Namibia’s new lingua franca nowadays, particularly among the younger and Black population. Although it is not an L1 to many Namibians, its use has increased since it was introduced as the sole official language in 1990 (NSA 2001, 2011; Ejikeme 2011: 15; Frydman 2011: 185; Buschfeld & Kautzsch 2014: 133, 139, 143; Stell 2014: 237, 239; Buschfeld & Kautzsch 2016: 6). Hence, Afrikaans and English – more overt prestige is ascribed to the latter – compete

3. Similar to the Namibian languages, the South African ones also have alternative names, e.g. without the language-indicating prefixes *isi-* and *se-* (Lewis et al. 2016).

against each other as Namibia's *linguae francae* with informal and formal functions, in the capital of Windhoek at least (Stell & Groenewald 2016: 2).

In post-Independence Namibia, the number of English L1 speakers slightly increased to 1.9% in 2001 and 3.4% in 2011, according to the Population and Housing Censuses (NSA 2001: 48; NSA 2011: 68). Furthermore, attitudes toward English seem positive as empirical research from the 1990s and 2010s has shown, e.g. Beck (1995), Pütz (1995a), Steigertahl (2010, 2015, 2017), Buschfeld & Kautzsch (2014), Stell (2014), and Pauli (2017). In addition, recent studies have assessed potential features of English(es) spoken in Namibia, e.g. Otaala (2006), Buschfeld & Kautzsch (2014), Kautzsch & Schröder (2016) and Steigertahl (2017). However, several of these potential features might involve phonological processes, such as consonant cluster reduction and simplification, and not necessarily morphological processes. Working with spoken corpora, the researcher has to be aware of the possibility that spontaneous speech contains slips of the tongue, which do not, in all circumstances, constitute potential features of a variety (de Klerk 2006: xiii). This will be further investigated in the following.

The Corpus of ESBNaPI

Data collection

Following a pilot study in 2013, the data for the corpus of ESBNaPI was collected in four Namibian communities and their surrounding areas in February and March 2015: Divundu, Gobabis, Ruacana, and Usakos. For an extensive project on educational policy and variety status of the English(es) used in post-Independence Namibia, data in the form of semi-structured sociolinguistic interviews in the tradition of Labov (1972),⁴ questionnaires on language attitudes and use, and participant observation was collected. This triangulation approach was selected in order to reduce shortcomings of each method (cf. Fink & Kosecoff 1985: 19; Schröder 2003: 31, 39; Feagin 2004: 22; Schröder 2004: 151; Gallois et al. 2009: 601; Harris 2011: 32; Meyerhoff et al. 2012: 122; Schilling 2013: 3, 19). The technique of semi-structured sociolinguistic interviews, which were used for the corpus, will be further examined here. This form of inquiry was selected to cover different registers and to receive as much natural and spontaneous speech as possible (cf. Feagin 2004: 29; Krug & Sell 2013: 73; Stell 2009: 89). However, word lists and reading passages soon had to be discarded due to some participants' insufficient reading

4. Sociolinguistic interviews generally include reading passages, word lists and an interview with the researcher (Labov 1972).

skills; hence, only the interviews between me as the researcher and Namibian participants, that is part IV of the Labovian (1972) sociolinguistic interview, were collected for the final analysis of the spoken data. It is possible that the language that was produced by the participants was influenced by the non-Namibian interview partner. For instance, other Namibian languages were not possible to switch to as is common in multilingual spaces such as Namibia. Hence, the interviews were of monolingual character, which might not reflect everyday speech in these Namibian communities. Still, the interviewer was a non-native speaker of English herself; this might have supported the interview procedure as both speakers were supposedly on similar levels with regard to language proficiency. Besides, the suggested rules for participant characteristics for ICE (International Corpus of English) corpora was not adhered to as contributors would then be obliged to have a university degree:

In general the population represented in the corpus will be adults of 18+ who have received formal education through the medium of English to the completion of high (secondary) school, but second language countries might require a university degree. (Greenbaum 1996: 3)

This does not seem appropriate or possible in Namibia if a somewhat realistic picture of the use of English(es) is aimed at. The ICE format suggested by Greenbaum (1996: 3) would have excluded the bulk of Namibians since only 5.8% of the population have a university degree (NSA 2011: 51). Thus, the interviews were of a dialogical nature between me as the interviewer and one interviewee and lasted around 20 minutes each. At times, a second or third person was present; however, the interview was only conducted between two people. Each interviewee was interviewed once during the data collection process in 2015. In order to elicit as much natural speech as possible, the content of the interviews was guided by the interviewer and the interviewee and thus includes topics of their interest, e.g. childhood memories, Namibian food and traditions, and recent events such as the inauguration of the new president Hage Geingob in 2015 and the FIFA World Cup in South Africa in 2010. In this light, the interviewer's familiarity with the local circumstances supported the development of interesting questions for a casual speech style (cf. Feagin 2004: 31). Additional questions were based on previous studies, for example Pütz (1995a), Steigertahl (2010), and Buschfeld & Kautzsch (2014). The participants of the study were then contacted through the friend-of-a-friend approach, also called *snowball sample* (cf. Meyerhoff et al. 2012: 128), in order to reduce the distance between interviewer and interviewee and to overcome the observer's paradox to a certain extent (Feagin 2004: 29). Through friends of friends, family, schools and educational organizations, contact persons were met in the respective towns and their locations; this facilitated getting acquainted with the community and finding participants. This strategy is recommended when conducting

research as an outsider to a Namibian community as matters of familiarity and respect are highly important in these settings.

The four communities were chosen for different reasons. First of all, in each of these communities, one ethnolinguistic group is dominant (cf. Scherz et al. 1981: 11 as quoted in Nampala & Shigwedha 2006: xxf.). In Gobabis, for instance, data was exclusively collected in the Herero location. Secondly, these communities were chosen as they represent average Namibian towns and locations, that is where tourism and the increasing use of European languages with foreigners seem limited (cf. Brock-Utne 2000: 200; Nampala & Shigwedha 2006: 173–230; Töttemeyer 2010: 64; Wallace & Kinahan 2011: 86–93, 201–203; Stell 2014: 226; New Era 2015; Lewis et al. 2016). By moving away from formerly applied approaches, in which university settings tend to be chosen as places of data collection and Africans of African descent have been described as merely “indigenous”, a more realistic linguistic situation could be depicted. Thirdly, the communities were chosen because of available contact persons, such as friends, friends and family of friends and educational non-governmental organizations. Contact persons are immensely important when out in the field since trust to the researcher is created then, as, for instance, a friend of a friend to the community (cf. Feagin 2004: 22; Schilling 2013: 203). The exact settings where interviews were conducted and questionnaires were distributed and filled out included very private settings such as people’s homes or the places where I stayed, but also public settings, e.g. classrooms, municipalities, museums, libraries and offices. My contact persons, who were often principals, teachers, or headmen, introduced me to the community so that solidarity was assured and the beneficial objective of the project was open and known to the participants.

The participants for the study were selected based on three factors: They had to be Namibian citizens or at least have resided in the country for the past 10 years; they had to be at least 16 years old, and sufficiently proficient in English that is to have a conversation with the interviewer without major misunderstandings. 10 years of residency was chosen because of the former apartheid rule over the country, thus people being sent into exile and returning after 1990. Consequently, some inhabitants of Namibia have, for example, South African or Angolan citizenship although they are born on the land that is today called Namibia; hence, not all of the so-called “Namibians” are Namibian by citizenship. The age level was selected so as to have mature participants who can reflect on their language use and attitudes, but also because it was easier to obtain consent from adults in contrast to parents of participants under age. The 77 interviewees include Namibians who position themselves as Black and state that they belong to different ethnic and linguistic, age groups and sex. In the following, these 77 interviewees will further be described as these participants make up the corpus of ESBNAPI. Percentage figures are rounded to whole numbers; thus, they might not always add up to 100%.

The youngest age category comprises Namibians “born-free” (Etjikeme 2011: 165), that is who were between 17 and 25 years of age by the time of data collection. This makes up the biggest category with 45% of all participants. The second youngest category, 26 to 35 years of age, includes Namibians who were born before independence and might have experienced the implementation of English as medium of instruction at school after Afrikaans had been used under apartheid. This group makes up a quarter of the 77 interviewees. The age group between 36 and 45 years comprises around a fifth of the participants and the oldest age group, that is, 46 to 54 years, includes 8% of the interviewees. Members of the oldest three age groups probably lived and were educated under apartheid rule if not in exile. The mean age of the 77 interviewees is 29. Besides, the population at hand resembles the overall age distribution of the Namibian people (NSA 2011: 8). What is more, there is a very slight gender imbalance among the participants, with 52% of them being female and 48% being male. No other categories were suggested by the interviewees. Furthermore, one third of the interviewees speak an Oshiwambo language at home, followed by around one fifth each who speak RuKavango languages respectively Damara/Nama at home. 14% of the participants said that they mostly use Otjiherero at home; 12% use Zambezi languages such as Subia, Yeyi and SiLozi.⁵ 3% of the interviewees stated that they use two languages at home and 1% each said that s/he uses Afrikaans respectively English at home. As already mentioned above, most of these 77 participants belong to the ethnic group of their respective linguistic group. Only the two interviewees who stated that they use two languages at home belong to two ethnic groups each, that is Owambo/Lozi and Owambo/Kavango. The Afrikaans L1 speaker is of Owambo ethnicity and the English L1 speaker is mixed but does not know her ethnicity.

Methodology

In total, 82 interviews were recorded with a Zoom H2N Handy Recorder. This recorder can be recommended as it is small and did not seem to disturb the participants. Of these 82 interviews, 77 were used for morphosyntactic and content-based analyses. The aim of the latter, however, was to investigate language use and attitudes, not to assess morphosyntactic structures; thus, it will not be further dealt with in the present chapter. Five interviews could not be made use of because of background noise and weak English skills, that is interviewer and interviewee did

5. For a better overview, varieties of one language are grouped together under one term, i.e. Oshiwambo languages, RuKavango languages, Otjiherero (which includes Dhemba), and Zambezi languages (including SiLozi, Yeyi, Subia). ‘Various’ includes two participants who stated that they speak two languages at home, that is, Oshiwambo/SiLozi and Oshiwambo/RuKavango respectively.

not communicate successfully, except for a few formal greetings. The remaining 77 interviews were orthographically transcribed with the program f4. Overlaps and self-repairs were also transcribed, as can be deduced from the mark-up settings in Table 1 below. Pauses were not considered as a detailed discourse analysis was not intended. When short interruptions occurred, they were transcribed in brackets, similar to comments, if possible. Seldom were interruptions in a language unknown to the researcher, e.g. Khoikhoigowab; then it was impossible to transcribe it. The same applies to code-switching and – mixing. Later, however, interruptions were deleted from the corpus as these included persons other than the interviewee himself/herself.⁶ Consequently, a manual analysis of the corpus was necessary. For better comprehension, for the content-based analysis, for instance, the interviewer’s speech was also transcribed but later deleted from the corpus and any further word counts. The program f4 was selected as it automatically inserts time stamps so that audio and text files are linked. With shortcuts, different mark ups can be easily inserted, such as laughs. The software is not as detailed as ELAN, for instance, that can be easily linked to Praat for further phonological analyses, but it can be recommended for orthographical transcriptions of one-on-one interviews that serve as a basis for morphosyntactic analyses. It is easy to understand and handle and fulfills all functions necessary for the analysis at hand. Finally, the complete transcribed corpus was saved as docx, pdf, and txt files. Both these files and the audio data is available as such. Hence, it can be searched with programs such as AntConc and WordSmith Tools, which allow for concordance analyses, for example. All in all, the corpus comprises 77 speakers and makes up 28.5 hours of speech and 190,000 words. The mark-up conventions applied to the transcriptions follow du Bois (2006). As the analyses do not include detailed discourse analyses, some mark-ups were unnecessary and excluded from transcriptions, e.g. units and pauses, as explained above. Table 1 displays the mark-up conventions used.

Table 1. Mark-up Conventions Used for ESBNaPI (cf. du Bois 2006)

Mark-up	Function	Mark-up	Function
[]	overlap	wor-	truncated word
.	terminative tone	@	laugh
,	continuative tone	@word	laughing word
!	exclamation tone	#	unintelligible syllable
?	appeal tone	#word	uncertain word
/	truncated intonation unit	~name	pseudograph

6. They are saved in an earlier version which includes the interviewer’s speech as well and is available as audio file. Thus, this version can be used for further studies. However, it does not only include speech by Namibians.

Since the corpus of ESBNaPI does not include native English(es) as a whole, with the exception of one speaker only, the data was not tagged. I agree with Mollin (2006: 102) who states “tagging [being] [...] a must for up-to-date corpora of native speech, [...] the taggability (or the accuracy rate of tagging) of non-native speech is not yet resolved”. In this case, tagging would have needed to be carried out manually. This would have proven extremely time-consuming. Moreover, some transcriptions were difficult: As mentioned above, for example, the lack of certain consonants, such as final plural or third person singular -s, could imply phonological and not necessarily morphological processes. Hence, any corpus searches have to be carried out carefully, that is always in context, so as to take note of potential misleading conclusions in the morphosyntactic analysis.

Morphosyntactic forms of English(es) spoken by black Namibians

In the following, an exemplary analysis of morphosyntactic forms detected in the corpus ESBNaPI will be presented and discussed. A brief overview of different parts of the morphosyntax used among the interviewees will be given. If possible, these grammatical forms will then be compared to SAE, in order to address potential generalizations of post-Independence English(es). This analysis does not imply that these forms do not occur in other varieties of English (cf. Siemund 2013 for an in-depth analysis of features shared by different varieties of English).

First, the corpus was searched for potential morphosyntactic features. During the transcription process, several possible grammatical structures had stood out; these were further looked into. In addition, grammatical structures described by previous studies on varieties of English, e.g. Platt et al. (1984), Mesthrie & Bhatt (2008), Siemund (2013), Buschfeld & Kautzsch (2014), were examined in the corpus. Second, these forms were searched for in AntConc, Version 3.4.4.0 (Anthony 2014), and the benchmarks put forward by Buschfeld (2013: 65–68) were used for the interpretation of the frequencies. She argues that a form is regularly used if it appears with a frequency of over 30 to 50%, that is 30% is the benchmark to indicate nativization. If a form is used more often than in 50% of all cases, it is said to be preferred as a local feature and might develop into a norm. If a grammatical form is detected with a frequency of over 70%, it is presumably used by most speakers and develops into a rule (Buschfeld 2013: 65–68, 106). Such systematic procedure is necessary if a corpus analysis aims at distinguishing between “features” and “errors”, as ascribed to interlanguages and learner Englishes (Corder 1981: 65f.; Mollin 2006: 49; Mollin 2007: 172; Groves 2010: 119; Bongartz & Buschfeld 2011: 37; Buschfeld 2013: 64). However, negative evidence should also be kept in mind; that is, the absence of certain structures does not imply its general non-occurrence or

lack of use (Trüb 2008: 140). Moreover, to the best of my knowledge, no remarks on absolute numbers for systematic analyses have been given in the literature so far, that is how many is enough to qualify for further investigation. Even if benchmarks of 30%, 50% and 70% are definitely helpful, absolute numbers have to be mentioned as well. Following this quantitative approach, the findings were exported into a Microsoft Excel 2013 file. In addition, potential differences between L1 and age groups were tackled. However, no peculiarities with regard to these variables could be found in ESBNaPI.

In the area of verb phrases, two structures were found to appear with a frequency of at least 30%, that is the benchmark to indicate nativization according to Buschfeld (2013: 65–68): the form *will* + verb for hypothetical context and *must* for *shall/should*. Hypothetical contexts include questions such as “What would you do if you were offered a job in Germany/Europe/South Africa?” and “What would you do if you won the lottery?”. Around half of the 52 hypothetical contexts found in ESBNaPI, that is 25 answers (48%), are built with a form of *will* + verb instead of the standard form *would* + verb, as shown in Examples (1) and (2).

- (1) I *will buy* myself a house, and a car
- (2) I’*ll travel* overseas, I’*ll go* to Germany!

This non-standard use of a form of *will* + verb instead of *would* + verb might be interpreted as a volition marker or an auxiliary to express wishes and refer to future time events. Clear results are difficult to obtain with this data as detailed background knowledge about the participants does not suffice to analyze the probability of their desires. What is more, the use of this structure has not been mentioned with regard to so-called SAE, to the best of my knowledge (cf. Branford & Branford 1991; Branford 1996; Chisanga 2001; Lass 2004; de Klerk 2006; Bowerman 2008; Mesthrie 2008a, 2008b; McCormick 2008; Minow 2010).

Illocutionary force of *must* appears with a frequency of 51%, that is 187 times, of all 374 instances of the three verb forms *must*, *shall* and *should*, in the corpus. This is shown in Examples (3) and (4).

- (3) If you are eating nicely, I *must* also eat nicely
- (4) I also believe that a person *must* first, *must* try your mother tongue.

It can be assumed that the use of *must* for *shall/should* derives from Afrikaans *moet* (cf. Bowerman 2008: 477; McCormick 2008: 523). As Afrikaans is a de facto official language in South Africa, this structure has also been found in what has been termed White South African English (WSAE) (Bowerman 2008: 477) and Cape Flats English (CFE) (McCormick 2008: 523).

Noun and pronoun structures found in ESBNaPI, which appear with a frequency of at least 30%, include pluralization of collectives, that is nouns that refer to discrete objects (Wiese 2012: 54f.), and the use of quantifier *one of* without number agreement. Overall, several instances of non-standard pluralization of non-count nouns, mass nouns and collectives were found. For instance, *stuffs*, *foods*, *peoples*, and *childrens* were detected between 6 to 18 times in ESBNaPI. However, they do neither appear commonly nor with a higher frequency than the standard form, that is, only with a frequency between 5 and 18%. Yet, two plural forms of collectives make an exception here as they appear with a higher frequency: *cattles* and *sheeps*. The form *cattles* instead of standard *cattle* was detected with a frequency of 38%, that is 23 of all 60 standard and non-standard hits in the corpus. Two examples are included in Examples (5) and (6).

- (5) we just look after, *cattles* also and goats
- (6) and the other jobs are like, looking after *cattles*, and, and working in a bar

Another instance of a frequent non-standard pluralization is the form *sheeps* for plural *sheep*. This non-standard form appears with a frequency of 56%, that is 14 of the 25 standard and non-standard occurrences, as demonstrated in Examples (7) and (8).

- (7) The weather becomes too cold and, the *sheeps* and the goats end up dying in the south.
- (8) in Wamboland we don't really have *sheeps*.

It is possible that final *-s* is applied because the interviewees learned this regular pluralization pattern, thus apply it more regularly than in so-called Standard English. Another possibility might be the structural discrepancy between language families. In Bantu languages, certain noun class markers distinguish between mass and count nouns (Ghomeshi & Massam 2012: 6). Number of nouns is commonly marked with affixes (Williamson & Blench 2000: 11–13). Then again, as described above, 20% of the interviewees are not Bantu L1 speakers, that is, interviewees who speak Damara/Nama, Afrikaans or English as L1. Consequently, this explanation cannot be sufficient for all occurrences. Furthermore, these pluralization structures were found among other varieties of English as well (Mesthrie & Bhatt 2008: 53; Wiese 2012; Mohr 2016). In what has been called Black South African English (BSAE) and Indian South African English (ISAE), non-count nouns, collectives or mass nouns frequently occur with plural *-s*. Then again, plural *-s* has also been omitted in other cases so that number has to be guessed from the context, if possible (Mesthrie 2008a: 497; Mesthrie 2008b: 513).

Moreover, the quantifier *one of* was found 78 times in the corpus. In 27 cases of all occurrences (35%), number agreement is missing and a noun in its singular form is used, as shown in Examples (9) and (10).

- (9) one can say it's *one of the biggest primary school* in the eh region
- (10) so maybe that can be *one of the problem*

It could be assumed that *one of* cognitively triggers the use of singular forms because only a single item is referred to. Another possibility is that consonant clusters are reduced, hence plural *-s* omitted. This would imply a phonological and not a morphological process then (cf. Platt et al. 1984: 117f.; Percillier 2016: 46). Besides, the structure of *one of* + singular nouns was not discussed in the research literature found on what has been termed SAE; thus, comparisons cannot be drawn here.

With regard to third person singular agreement, two items were detected that appear with a frequency of at least 30%, however with very few absolute occurrences: 3rd person singular pronoun *she* without number agreement of *to do* and 3rd person singular pronoun *he* without number agreement of *to have*. The form *she don't* appears with a frequency of 38%, that is 6 out of 16 times. This is depicted in Examples (11) and (12).

- (11) *she don't* speak English or Afrikaans
- (12) *she don't* even try to speak other languages with her kids.

3rd person singular pronoun *he* without number agreement of *to have* appears with a frequency of 36%, that is 8 out of 22 instances. Examples (13) and (14) show this lack of number agreement.

- (13) *he have* to go and attend
- (14) *He have* to come with his uncle

In what has been labelled BSAE, ISAE, and CFE, third person singular present tense *-s* is not always used (McCormick 2008: 525; Mesthrie 2008a: 488; Mesthrie 2008b: 501). In contrast, it is common in so-called WSAE and might be influenced by Afrikaans, which does not have third person agreement (Bowerman 2008: 480f.). All in all, this low number of occurrences shows that number agreement tends to be applied in its standard form. A high frequency of lack of number agreement was not found in the corpus.

Conclusion

This chapter addressed the compilation process of the corpus of ESBNaPI and presented morphosyntactic structures found in the corpus as examples of the Englishes used in southern Africa. Generally, awareness of corpora of African Englishes was raised as literature in this field is scarce; the same applies to English(es) used in Namibia, which only recently attracted attention in academia. In this context, the linguistic situation of Namibia was described, where two languages are used as *linguae francae*, that is English and Afrikaans. Overall, the number of L1 speakers of English is still low, despite figures increasing since its implementation as the sole official language in 1990. As figures on L2 speakers do not exist, only assumptions can be made.

The corpus of ESBNaPI was introduced in detail, including a description of the data gathering which was carried out through semi-structured sociolinguistic interviews conducted with Namibian participants mainly in four different communities. Afterwards, light was shed on the methodology applied, that is, one-on-one recordings, orthographic transcriptions, corpus searches and qualitative and quantitative analyses of potential morphosyntactic features of the English(es) spoken by Namibians who position themselves as Black. Hence, the use of this corpus was highlighted. Some of these potential features were then presented, discussed, and compared to so-called SAE, such as *will* + verb for hypothetical context, illocutionary *must*, pluralization of collectives, quantifier *one of* without number agreement, and lack of third person singular agreement. All of these occur with a frequency of at least 30%, thus, they indicate nativization according to the thresholds coined by Buschfeld (2013: 65–68) and might develop into a norm in the future if more frequently used and if accepted by the speakers. Differences with regard to the variables of age and L1 could not be found. Yet, it is possible that a larger corpus might reveal different findings. In general, it is important to not only search corpora qualitatively and list non-standard forms of English(es) but also to conduct a quantitative analysis to investigate systematic use of grammatical items. In this way, potential features of a possibly new variety of English can be detected and distinguished from “errors”. However, it remains to be seen if these potential features are accepted by the Namibian English(es) speakers themselves. This might be interesting for future studies on language use and attitudes in Namibia.

Overall, Namibia as an officially English-speaking country serves as an example of the southern African region. It is hoped that this chapter adds to the understanding of corpus compilation, particularly data collection and methodology, so that future researchers can gather and receive somewhat authentic data for language corpora. All in all, this exemplary study offers insights into morphosyntactic

structures used in four Namibian communities in comparison to the neighboring country South Africa. Consequently, this chapter does not aim to provide a holistic view on English(es) used in all of Namibia but represents a starting point for future sociolinguistic research and corpus linguistics in southern African countries.

References

- Adegbija, Efurosibina. 1994. The context of language planning in Africa: An illustration with Nigeria. In *Language Contact and Language Conflict*, Martin Pütz (ed.), 139–163. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.71.o8ade>
- Allwood, Jens & Hendrikse, Andries. 2003. The spoken language corpora for the nine official African languages of South Africa. *Southern African Linguistics and Applied Language Studies* 21(4): 189–201. <https://doi.org/10.2989/16073610309486343>
- Allwood, Jens, Hammarström, Harald, Hendrikse, Andries, Ngcobo, Mtholeni N., Nomdebevana, Nozibele, Pretorius, Laurette & van der Merwe, Mac. 2010. Work in spoken (multimodal) language corpora in South Africa. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC) 2010*, Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds), 885–889. Valletta, Malta: European Language Resources Association (ELRA).
- Arndt, Susan. 2005. Mythen des weißen Subjekts: Verleugnung und Hierarchisierung von Rassismus. In *Mythen, Masken und Subjekte. Kritische Weißseinsforschung in Deutschland*, Maureen Maisha Eggers, Grada Kilomba, Peggy Piesche & Susan Arndt (eds), 340–362. Münster: Unrast.
- Arndt, Susan. 2012. *Die 101 wichtigsten Fragen: Rassismus*. Munich: Beck. <https://doi.org/10.17104/9783406638862>
- Arndt, Susan & Ofuately-Alazard, Nadja (eds). 2011. *Wie Rassismus aus Wörtern spricht. (K)Erben des Kolonialismus im Wissensarchiv deutsche Sprache. Ein kritisches Nachschlagewerk*. Münster: Unrast.
- Beck, Anke. 1995. Language and nation in Namibia: The fallacies of modernization theory. In *Discrimination through Language in Africa? Perspectives on the Namibian Experience* [Contribution to the Sociology of Language 69], Martin Pütz (ed.), 207–222. Berlin: Mouton de Gruyter.
- Bogdan, Robert C. & Knopp Biklen, Sari. 1992. *Qualitative Research for Education. An Introduction to Theory and Methods*. Boston MA: Allyn and Bacon.
- du Bois, John W. 2006. Basic symbols for discourse transcription. Level 3+ by topic. In *Representing Discourse*. Department of Linguistics University of California, Santa Barbara. <<http://www.linguistics.ucsb.edu/projects/transcription/A02asymbols.pdf>> (3 December 2014).
- Bologna, Anna. 2013. Rezension: Wie Rassismus aus Wörtern spricht. *MDÜ – Fachzeitschrift für Dolmetscher und Übersetzer* 4/13(59): 58–65.
- Bongartz, Christiane M. & Buschfeld, Sarah. 2011. English in Cyprus: Second language variety or learner English? In *Exploring Second-Language Varieties of English and Learner Englishes. Bridging a Paradigm Gap* [Studies in Corpus Linguistics 44], Joybrato Mukherjee & Marianne Hundt (eds), 35–54. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.44.o3bus>

- Bowerman, Sean. 2008. White South African English: Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 472–487. Berlin: Mouton de Gruyter.
- Branford, William. 1996. English in South African society: A preliminary overview. In *Focus on South Africa* [Varieties of English around the World 15], Vivian de Klerk (ed.), 35–52. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g15.o4bra>
- Branford, Jean & Branford, William. 1991. *A Dictionary of South African English*. Cape Town: OUP.
- Brock-Utne, Birgit. 2000. *Whose Education for All? The Recolonization of the African Mind*. London: Falmer Press.
- Buschfeld, Sarah. 2013. *English in Cyprus or Cyprus English? An Empirical Investigation of Variety Status* [Varieties of English Around the World G46] Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g46>
- Buschfeld, Sarah & Kautzsch, Alexander. 2014. English in Namibia. A first approach. *English World-Wide* 35(2): 121–160. <https://doi.org/10.1075/eww.35.2.o1bus>
- Buschfeld, Sarah & Kautzsch, Alexander. 2016. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* [First View].
- Chamberlain, R. 1981. *Toward a Language Policy for Namibia: English as the Official Language; Perspectives and Strategies* [Namibia Studies Series 4]. Lusaka: United Nations Institute for Namibia.
- Chisanga, Teresa. 2001. Current studies in black South African English: A brief survey. In *Guardian of the Word. Literature, Language and Politics in SADC Countries*, Brian Harlech-Jones, Ismael Mbise & Helen Vale (eds), 141–149. Windhoek: Macmillan.
- CIA (Central Intelligence Agency). 2014. *The World Factbook. Africa: Namibia*. Washington, DC. <<https://www.cia.gov/library/publications/the-world-factbook/geos/wa.html>> (19 August 2015).
- Corder, Stephen Pit. 1981. *Error Analysis and Interlanguage*. Oxford: OUP.
- Cukor-Avila, Patricia & Bailey, Guy. 2001. The effects of the race of the interviewer on sociolinguistic fieldwork. *Journal of Sociolinguistics* 5(2): 254–270. <https://doi.org/10.1111/1467-9481.00150>
- Deutsches Institut für Menschenrechte. 2017. Schwarz. *Glossar – Deutsches Institut für Menschenrechte*. <http://www.institut-fuer-menschenrechte.de/service/glossar/?tx_contagged%5Bsource%5D=default&tx_contagged%5Buid%5D=764&cHash=34e11defee4be8fd29ddb5858ce8ec37> (9 May 2017).
- Diekmann, Andreas. 2007. *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*, 17th edn. Reinbek: Rowohlt.
- Etijkeme, Anene. 2011. *Culture and Customs of Namibia*. Santa Barbara CA: Greenwood.
- Faculty of Education University of Pretoria. 2004. Guidelines for creating questionnaires. <www.ais.up.ac.za/edu/research/Vraelys.Eng.doc> (10 October 2013).
- Feagin, Crawford. 2004. Entering the community: Fieldwork. In *The Handbook of Language Variation and Change*, Jack K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds), 20–39. Malden MA: Blackwell.
- Fink, Arlene & Kosecoff, Jacqueline. 1985. *How to Conduct Surveys. A Step-by-Step Guide*. Beverly Hills CA: Sage.
- Fourie, David J. 1997. Educational language policy and the indigenous languages of Namibia. *International Journal of the Sociology of Language (IJSL)* 125: 29–42.

- Fourie, Katie. 1995. Afrikaans – The unwanted lingua franca of Namibia. In *Discrimination through language in Africa? Perspectives on the Namibian Experience* [Contribution to the Sociology of Language 69], Martin Pütz (ed.), 315–325. Berlin: Mouton de Gruyter.
- Frydmann, Jenna. 2011. A critical analysis of Namibia's English-only language policy. In *Selected Proceedings of the 40th Annual Conference on African Linguistics: African Languages and Linguistics Today*, Eyamba G. Bokamba, Ryan K. Shosted & Bezza Tesfaw Ayalew (eds), 178–189. Somerville MA: Cascadilla Proceedings Project.
- Gallois, Cindy, Watson, Bernadette & Brabant, Madeleine. 2009. Attitudes to Language and Communication. In *Handbook of Language and Communication: Diversity and Change*, Marlis Hellinger & Anne Pauwels (eds), 595–618. Berlin: De Gruyter.
- Ghomeshi, Jila & Massam, Diane. 2012. The count mass distinction: Issues and perspectives. In *Count and Mass across Languages*, Diane Massam (ed.), 1–8. Oxford: OUP.
- Government of South Africa. 1996. *The Constitution of the Republic of South Africa*. Pretoria: Government Printer.
- Greenbaum, Sidney. 1996. *Comparing English Worldwide. The International Corpus of English*. Oxford: Clarendon Press.
- Groves, Julie. 2010. Error or feature? The issue of interlanguage and deviations in non-native varieties of English. *HKBU Papers in Applied Language Studies* 14: 108–129.
- Haacke, Wilfrid H. G. 1994. Language policy and planning in independent Namibia. *Annual Review of Applied Linguistics (ARAL)* 14: 240–253. <https://doi.org/10.1017/S0267190500002919>
- Harlech-Jones, Brian. 1995. The role of English in Namibia: A sociocultural and linguistic account. In *Discrimination through Language in Africa? Perspectives on the Namibian Experience* [Contribution to the Sociology of Language 69], Martin Pütz (ed.), 224–243. Berlin: Mouton de Gruyter.
- Harlech-Jones, Brian. 1997. Looking at means and ends in language policy in Namibia. In *Language Choices: Conditions, Constraints, and Consequences* [Impact: Studies in Language and Society 1], Martin Pütz (ed.), 223–249. Amsterdam: John Benjamins. <https://doi.org/10.1075/impact.1.16har>
- Harris, Priscilla G. 2011. Language Schools in Namibia. The Missing Link in Educational Achievement? In *The Urban Trust of Namibia* [Monograph No 1]. Windhoek: Solitaire Press. <http://www.osisa.org/sites/default/files/language_in_schools_in_namibia_-_the_missing_link_in_educational_achievement.pdf> (8 March 2013).
- Kachru, Braj B. 1985. Standards, codification, and sociolinguistic realism: The English language in the outer circle. In *English in the World: Teaching and Learning the Language and Literature*, Randolph Quirk & Henry G. Widdowson (eds), 11–30. Cambridge: CUP.
- Kautzsch, Alexander & Schröder, Anne. 2016. English in multilingual and multiethnic Namibia: Some evidence on language attitudes and on the pronunciation of vowels. In *Anglistentag 2015 Paderborn. Proceedings*, Christoph Ehland, Ilka Mindt & Merle Tönnies (eds), 277–288. Trier: WVT.
- Kennedy, Graeme D. 1998. *An Introduction to Corpus Linguistics*. London: Longman.
- de Klerk, Vivian. 2006. *Corpus Linguistics and World Englishes. An Analysis of Xhosa English*. London: Continuum.
- Knappert, Jan. 1981. *Namibia. Land and Peoples. Myths and Fables*. Leiden: Brill.
- Krug, Manfred & Sell, Katrin. 2013. Designing and conducting interviews and questionnaires. In *Research Methods in Language Variation and Change*, Manfred Krug & Julia Schlüter (eds), 69–98. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511792519.007>

- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia PA: University of Pennsylvania Press.
- Lass, Roger. 2004. South African English. In *Legacies of Colonial English. Studies in Transported Dialects*, Raymond Hickney (ed.), 363–386. Cambridge: CUP.
- Lewis, M. Paul, Simons, Gary F. & Fennig, Charles D. 2016. *Ethnologue: Languages of the World*, 19th edn. Dallas TX: SIL International. <<http://www.ethnologue.com>>
- Maho, Jouni Filip. 1998. *Few People, Many Tongues. The Languages of Namibia*. Windhoek: Gamsberg Macmillan.
- Maho, Jouni Filip. 2008. NAMLOB – Namibian Languages Online Bibliography. <<http://goto.glocalnet.net/eball/namlob.pdf>> (25 August 2014).
- Malan, Johan S. 2005. *Die Völker Namibias*. Göttingen: Klaus Hess.
- Mazrui, Alamin M. 2004. *English in Africa. After the Cold War*. Clevedon: Multilingual Matters.
- McCormick, Kay. 2008. Cape Flats English: Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 521–534. Berlin: Mouton de Gruyter.
- Mendelsohn, John, Jarvis, Alice, Roberts, Carole & Robertson, Tony. 2009. *Atlas of Namibia. A Portrait of the Land and its People*. Cape Town: Sunbird Publishers.
- Mesthrie, Rajend. 2008a. Black South African English: Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 488–500. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110208429>
- Mesthrie, Rajend. 2008b. Indian South African English: Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 501–520. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110208429>
- Mesthrie, Rajend & Bhatt, Rakesh M. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511791321>
- Meyerhoff, Miriam, Adachi, Chie, Nanbakhsh, Golnaz & Strycharz, Anna. 2012. Sociolinguistic fieldwork. In *The Oxford Handbook of Linguistic Fieldwork*, Nicolas Thieberger (ed.), 121–146. Oxford: OUP.
- Milroy, Lesley. 1980. *Language and Social Networks*. Baltimore MD: University Park Press.
- Milroy, Lesley. 1987. *Observing and Analysing Natural Language*. Oxford: Blackwell.
- Minow, Verena. 2010. *Variation in the Grammar of Black South African English* [European University Studies: Reihe 21, Linguistics, 362]. Frankfurt: Peter Lang.
- Mollin, Sandra. 2006. *Euro-English. Assessing Variety Status*. Tübingen: Narr.
- Mollin, Sandra. 2007. New variety or learner English? Criteria for variety status and the case of Euro-English. *English World-Wide* 28(2): 167–185. <https://doi.org/10.1075/eww.28.2.04mol>
- Mohr, Susanne. 2016. From Accra to Nairobi: The use of pluralized mass nouns in East and West African postcolonial Englishes. In *Aspects of (Post)Colonial Linguistics*, Daniel Schmidt-Brücken, Susanne Schuster & Marina Wienberg (eds), 157–188. Berlin: De Gruyter.
- Möhlig, Wilhelm & Seidel, Frank. 2007. Namibian languages. In *Atlas of Cultural and Environmental Change in Arid Africa*, Olaf Bubenzer, Andreas Bolten & Frank Darius (eds), 146–147. Cologne: Heinrich-Barth Institute.
- Nampala, Lovisa T. & Shigwedha, Vilho. 2006. *Aambo Kingdoms, History and Cultural Change. Perspectives from Northern Namibia* [Basel Namibia Studies Series 8/9]. Basel: P. Schlettwein.
- Nelson, Gerald. 1996. The design of the corpus. In *Comparing English Worldwide. The International Corpus of English*, Sidney Greenbaum (ed.), 27–35. Oxford: Clarendon Press.
- Neuman, W. Lawrence. 1994. *Social Research Methods. Qualitative and Quantitative Approaches*, 2nd edn. Boston MA: Allyn and Bacon.

- New Era. 2015. Is Afrikaans now Namibia's second official language? *New Era*, May 22, 2015. <<https://www.newera.com.na/2015/05/22/afrikaans-namibias-official-language/>> (15 April 2015).
- NSA (Namibia Statistics Agency). 2001. *Namibia 2001. Population and Housing Census Main Report*. <http://www.nsa.org.na/files/downloads/a5d_Namibia%202001%20Population%20and%20Housing%20Census%20Main%20Report.pdf> (1 June 2015).
- NSA (Namibia Statistics Agency). 2011. *Namibia 2011. Population and Housing Census Main Report*. <<http://cms.my.na/assets/documents/p19dmn58guram30ttun89rdrp1.pdf>> (1 June 2015).
- Otaala, Laura Ariko. 2006. Consonant devoicing by Namibian speakers of English. In *The Study and Use of English in Africa*, Arua E. Arua, Mompoloki M. Bagwasi, Tiro Sebina & Barolong Seboni (eds), 165–176. Newcastle upon Tyne: Cambridge Scholars.
- Pauli, Julia. 2017. 'Of course I can speak English more than my mother tongue!' Sprache und Identität aufstrebender Mittelklassen im postkolonialen Namibia. In *Form und Funktion. Festschrift für Angelika Redder zum 65. Geburtstag*, Arne Krause, Gesa Lehrmann, Winfried Thielmann & Caroline Trautmann (eds), 409–421. Tübingen: Stauffenburg.
- Percillier, Michael. 2016. *World Englishes and Second Language Acquisition. Insights from Southeast Asian Englishes* [Varieties of English around the World G58]. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g58>
- Pienaar, Leela & de Klerk, Vivian. 2009. Towards a corpus of South African English: Corraling the sub-varieties. *Lexikos* 19: 353–371. <https://doi.org/10.4314/lex.v19i1.49135>
- Platt, John, Weber, Heidi & Lian, Ho Mian. 1984. *The New Englishes*. London: Routledge and Kegan Paul.
- Pütz, Martin. 1995a. Attitudes and language: An empirical investigation into the status and use of English in Namibia. In *Discrimination through Language in Africa? Perspectives on the Namibian Experience* [Contribution to the Sociology of Language 69], Martin Pütz (ed.), 245–284. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110906677>
- Pütz, Martin. 1995b. Official monolingualism in Africa: A sociolinguistic assessment of linguistic and cultural pluralism in Namibia. In *Discrimination through Language in Africa? Perspectives on the Namibian Experience* [Contribution to the Sociology of Language 69], Martin Pütz (ed.), 155–173. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110906677>
- Rickford, John R. & McNair-Knox, Faye. 1994. Addressee and topic-influence style shift: A quantitative sociolinguistics study. In *Sociolinguistic Perspectives on Register*, Douglas Biber & Edward Finegan (eds), 235–276. Oxford: OUP.
- van Rooy, Bertus. 2013. Corpus linguistic work on black South African English. *English Today* 113(29/1): 10–15. <https://doi.org/10.1017/S0266078412000466>
- Scherz, Anneliese, Scherz, Ernst Rudolf, Taapopi, G. & Otto, Antje. 1981. *Hairstyles, Head-Dresses and Ornaments in South West Africa/Namibia and Southern Angola*. Windhoek: Gamsberg Uitgewers.
- Schilling, Natalie. 2013. *Sociolinguistic Fieldwork*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511980541>
- Schröder, Anne. 2003. *Status, Functions, and Prospects of Pidgin English: An Empirical Approach to Language Dynamics in Cameroon*. Tübingen: Gunter Narr.
- Schröder, Anne. 2004. Deeper insights through triangulation: Experiences from a sociolinguistic study on Pidgin English in Cameroon. In *Crossing Borders. Interdisciplinary Approaches to Africa* [Afrikanische Studien 23], Anne Schröder (ed.), 151–181. Münster: LIT.

- Seidel, Frank. 2015. Describing endangered languages. Experiences from a PhD grammar project in Africa. In *Language Documentation and Endangerment in Africa* [Culture and Language Use 17], James Essegbey, Brent Henderson & Fiona Mc Laughlin (eds), 277–312. Amsterdam: John Benjamins. <https://doi.org/10.1075/clu.17.12sei>
- Siemund, Peter. 2013. *Varieties of English. A Typological Approach*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139028240>
- Smith, David. 2015. Lüderitz v !Nami≠nūs: Dispute over Town's Name Divides Namibia. *The Guardian*, February 26, 2015. <<https://www.theguardian.com/world/2015/feb/26/luderitz-v-naminus-dispute-over-towns-name-divides-namibia>> (15 April 2016).
- Sow, Noah. 2008. *Deutschland Schwarz weiß: Der alltägliche Rassismus*. Munich: Bertelsmann.
- Spradley, James P. 1979. *The Ethnographic Interview*. Belmont CA: Wadsworth.
- Steigertahl, Helene. 2010. The Role of the Official Language English in a Namibian Community: An Empirical Study on Language Use and Language Attitude in Lüderitz. State Examination Thesis, University of Heidelberg.
- Steigertahl, Helene. 2015. Language use and attitudes. How do they contribute to identity-formation? In *Acta Germanica* [German Studies in Africa 43], Carlotta von Maltzan (ed.), 90–98. Frankfurt: Peter Lang.
- Steigertahl, Helene. 2017. English(es) Spoken by Black Namibians: Linking Variety Status with Educational Policy in Post-Independence Namibia. PhD dissertation, University of Bayreuth.
- Steigertahl, Helene. Submitted. Similarities and differences between varieties of speech in Namibia – Expressions of language ideologies and attitudes in four Namibian communities. *SOAS Working Papers in Linguistics*.
- Stell, Gerald. 2009. Is there a Namibian Afrikaans? Recent trends in grammatical variation in Afrikaans varieties within and across Namibia's borders. *Stellenbosch Papers in Linguistics PLUS* 39: 85–105.
- Stell, Gerald. 2014. Uses and functions of English in Namibia's multiethnic settings. *World Englishes* 33(2): 223–241. <https://doi.org/10.1111/weng.12082>
- Stell, Gerald & Groenewald, Gerald. 2016. 'n Perseptuele verslag van Afrikaans in Namibië: Tusien lingua franca en sosiaal-ekslusiewe taal. A perceptual account of Afrikaans in Namibia: Between lingua franca and socially exclusive language. *Tydskrif vir Geesteswetenskappe* 56(4–1): 1–21. <https://doi.org/10.17159/2224-7912/2016/v56n4-2a3>
- Tagliamonte, Sally. 2006. *Analysing Sociolinguistic Variation*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511801624>
- Töttemeyer, Andree-Jeanne. 2009. *Multilingualism/ Multiculturalism in Africa and its Impact on a Reading Culture: The Namibian Experience*. <<http://www.kzneducation.gov.za/Portals/0/ELITS%20website%20Homepage/IASL%202009/totemeyer.pdf>> (14 January 2014).
- Töttemeyer, Andree-Jeanne. 2010. *Multilingualism and the Language Policy for Namibian Schools* [PRAESA (= Project for the Study of Alternative Education in South Africa) Occasional Papers No. 37]. Cape Town: PRAESA. <<http://www.praesa.org.za/files/2012/07/Paper37.pdf>> (14 January 2014).
- Trudgill, Peter & Hannah, Jean. 2002. *International English. A Guide to Varieties of Standard English*, 4th edn. London: Arnold.
- Trüb, Regina. 2008. American English Impact on South African English: An Empirical Analysis of its Manifestations and Attitudes towards it. PhD dissertation, University of Regensburg.

- Turner III, Daniel W. 2010. Qualitative interview design: A practical guide for novice investigators. *The Qualitative Report* 15(3): 754–760.
- Wallace, Marion & Kinahan, John. 2011. *A History of Namibia. From the Beginning to 1990*. Johannesburg: Jacana Media.
- Webb, Vic. 1994. Revalorizing the autochthonous languages of Africa. In *Language Contact and Language Conflict*, Martin Pütz (ed.), 181–203. Amsterdam: John Benjamins.
<https://doi.org/10.1075/z.71.10web>
- Wiese, Heike. 2012. Collectives in the intersection of mass and count nouns: A cross-linguistics account. In *Count and Mass across Languages*, Diane Massam (ed.), 54–74. Oxford: OUP.
<https://doi.org/10.1093/acprof:oso/9780199654277.003.0004>
- Williamson, Kay & Blench, Roger. 2000. Niger-Congo. In *African Languages: An Introduction*, Bernd Heine & Derek Nurse (eds), 11–42. Cambridge: CUP.
- Wolf, Hans-Georg. 2004. Cultural approaches to second language varieties of English: A call for new methodologies and a review of some findings on (West) African English. In *Crossing Borders. Interdisciplinary Approaches to Africa* [Afrikanische Studien 23], Anne Schröder (ed.), 133–149. Münster: LIT.
- Wolff, H. Ekkehard. 2016. *Language and Development in Africa. Perceptions, Ideologies and Challenges*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781316105023>

The historical corpus of English in Ghana (HiCE Ghana)

Motivation, compilation, opportunities

Thorsten Brato

University of Regensburg

This chapter discusses various aspects relevant to the compilation of the Historical Corpus of English in Ghana (HiCE Ghana), a 600,000-word corpus of written Ghanaian English (GhE) from the period 1966 to 1975. The corpus captures written GhE in the early years of its nativization phase and by comparison to the relevant sections of the Ghanaian component of ICE makes it possible to trace structural nativization in real time. The paper addresses the motivation for such a corpus and shows how methodological and theoretical concerns and challenges have affected the final corpus design. The paper is rounded off with contemporary examples of GhE.

1. Introduction

Over the last 30 years the study of the so-called *New Englishes* (Platt, Weber & Ho 1984) has gained momentum. Several, sometimes competing models of varieties of and variation in English have since been put forward. Kachru's (1985) *Three Circles* model was ground-breaking for the classification of Englishes but proved inflexible and too static so that more flexible models were brought forward, some focussing on diachronic developments in colonial and postcolonial contexts (Schneider 2003, 2007; Trudgill 2004), others on centre and periphery and spheres of influence (Mair 2013). Most recently Buschfeld and Kautzsch (2017) proposed an integrative model that can be applied to (post-)colonial Englishes¹ and those without a colonial history alike.

While the evolution of New Englishes has been theorised extensively in recent years, there is still a considerable lack of empirical evidence of early developments of

1. The terms *New Englishes* and *Postcolonial Englishes* are used interchangeably in this paper.

these varieties so that scholars often have resorted to using synchronic data from varieties having progressed to different stages in their evolution to model the underlying linguistic processes. Only very recently have linguists begun to compile historical or diachronic corpora of New Englishes to overcome these limitations (see Section 3). One such corpus is the Historical Corpus of English in Ghana (HiCE Ghana), a corpus of printed Ghanaian English (GhE), which consists of about 600,000 words and covers the period 1966 to 1975. It is based on the design of the written-printed sections of the International Corpus of English (ICE) and is therefore largely compatible with the respective genres of ICE Ghana (Huber & Dako 2013).

In the present paper, I discuss various theoretical and methodological aspects that have been relevant to the compilation of HiCE Ghana. The following section will provide a very brief outline of the sociolinguistic and historical evolution of English in Ghana. A good understanding of these processes is an important step in the creation of a historical corpus, as these bear directly on the design, structure, and contents of HiCE Ghana. In Section 3, I discuss the motivation for the compilation of the corpus. Section 4 describes the compilation process and shows how both theoretical considerations and the (non-)availability of data for certain periods have affected the final corpus design. In Section 5, I briefly talk about some of the issues regarding the digitisation of the material and how the corpus was annotated. Section 6, Opportunities, provides the reader with some examples from the corpus to illustrate GhE at the time. I also summarise some of the completed and ongoing research based on the corpus. The paper is rounded off with a short conclusion and outlook showing the potential for expanding the corpus at a later stage and for the creation of comparable corpora of other Englishes.

2. English in Ghana

Ghana is a country in West Africa with a population of about 28 million people (2016 estimate; Ghana Statistical Service 2016). It is located on the Gulf of Guinea and borders the Ivory Coast, Burkina Faso and Togo. Ghana is a highly multilingual country, with over 80 languages from two language families, the Kwa languages in the south and the Gur languages in the north. The most widely spoken languages in the south are Akan (45%), Ewe (14%) and Ga-Dangme (7%); in the north Dagbani (4%) and Dagaare (4%) are most common (cf. Huber 2014: 87). About two-thirds of the population is literate in English (Ghana Statistical Service 2012: 6).

Schneider (2007) suggests that as a variety develops it goes through five incremental stages based on changes in the identity constructions of the British settlers and the indigenous population. According to Huber (2014: 87–91), Ghana currently falls between the nativization and endonormative stabilization phases.

However, at the same time Huber makes an important point about the divergence of GhE from the prototypical path laid out by Schneider, who suggests (2007: 31f.) that one of the most important driving forces in the development of a new variety of English is the interplay between colonisers (the settler or STL strand) and colonised (the indigenous or IDG strand). Ghana was an exploitation colony in the sense of Mufwene (2001: 8, fn.3) and administration was largely in the hands of local rulers, which means that there were hardly any British in the country over the whole colonisation period and those who were there would not lose their ties with the mother country. This leads Huber (2014: 88) to conclude that “convergence and identity construction [...] did and does take place not so much *between* the STL and IDG groups but rather *within* the IDG strand” (emphasis in the original).

The first contact between locals and Europeans dates to the late 15th century, when, in 1471, Portuguese traders first set foot on the coast of today’s Ghana. They set up a first permanent post near today’s Elmina in the Central Region 11 years later and called the place *Costa do ouro* (Gold Coast) because of the easy availability and abundance of gold. The first English ship arrived in 1553, followed by the Dutch in the 1590s and Danish, French, Prussian and Swedish merchants over the following decades, who all built forts along the coastline. The Dutch built their first settlement in 1612, the British followed 20 years later when they set up a fort at Kormantin east of today’s Cape Coast. 1632 therefore constitutes the onset of the foundation phase (Huber 2014: 87–88). It is important to note that during this early phase the competition was largely about commercial interests and Europeans and Africans were equal partners. Also, none of the European powers were interested in colonisation (Bourret 1960: 13). Between the mid-1600s and late 18th century the slave trade flourished, in which the Fante, a local tribe inhabiting the south-western area of today’s Ghana, were involved as collaborators. By the turn of the century most of the other European nations had lost interest in the trade so that only the British, Danish, and Dutch were left when the slave trade was finally abolished in 1807.

During this period, there were only very few speakers of English in the IDG strand, and those who spoke English would mainly live near the forts in the coastal areas. Schools were far and few and were mainly attended by the children of English traders and local women and, albeit in very small numbers, children of wealthier urban African traders (Foster 1965: 45).

Formal colonisation started in 1844 and constitutes the onset of the exonormative stabilization phase. At this time, local chiefs in the western areas yielded some of their juridical power to the English and accepted this foreign influence (Huber 2004: 842–44). After 1874, when the Dutch had finally left, there were only British settlements left and the coastal strip (up to about 100km inland) was proclaimed the Gold Coast Colony. The capital, which used to be Cape Coast, where the English

had built a large castle in the 17th century, was moved to Accra in the Ga-speaking region in 1877. The following years saw several wars between the British and the Ashanti, a powerful tribe to the north of the Gold Coast Colony. The English were defeated several times but by 1901 Ashantiland and the Northern Territories were integrated into the colony as protectorates, followed by British Togoland in 1922.

During the second phase, there were still overall only very few locals who could speak English. Cape Coast became a centre of English education and some of the prestigious schools in the country are still located in or near the city today. Overall, schools were concentrated in the urban centres of the south and most of the locals would speak Fante (a dialect of Akan), Ga (the dominant language in Accra) or Ewe (spoken mainly in the Eastern Region) as their first language. While the actual number of speakers of English compared to the rest of the local population was still low overall, they were the ones who filled the key positions in the political, administrative and social spheres or became teachers of English. As they spoke English first and belong to the most influential ethnic groups, they may be considered the founder population (Mufwene 2001), who provided some covert norms when the colony was expanded at a later stage (cf. Brato in prep. for a more detailed discussion).

The onset of the third phase – nativization – is the declaration of independence in 1957. This phase “is the most interesting and important, the most vibrant one, the central phase of both cultural and linguistic transformation.” (Schneider 2007: 40). It is characterised by large-scale linguistic change as in the IDG strand bilingualism becomes (even) more common. English is often the only co-official language and there is a need to be able to communicate in it more strongly. The speakers transform the English language and make it their own. Linguistic peculiarities, which may have already developed before this time, become more widespread and acceptable. They surface most clearly in a distinct accent, but there are also morphosyntactic innovations, which so far, however, have often either gone unnoticed or were just considered deviances from the exonormative standard or learner errors. Language use and expression become an issue. This shows most strongly in the discussion about linguistic norms. With the absence of an STL strand of any sizable number, these discussions were mainly restricted to the small, highly educated elite. Many, if not most of the people in the higher positions in administration, society and education had spent considerable time in the UK or US, either as students or for their PhDs. They usually held the view that GhE was erroneous and that only the exonormative, that is British standard would be acceptable. Letters to the editors complaining about the decline of English in the country or pieces addressing the “infectious [...] disease” (Mensah October 22, 1971: 6) that is GhE newspaper writing abound. These views are even held today. In a book about which accent should be taught in Ghanaian schools Odamtten and Laryea (2009: 5ff. cited in Huber 2014: 89) claim that only Received Pronunciation should be the target.

Structurally, the nativization phase is characterised by innovations on all levels of linguistics, some of which clearly set Ghanaians apart from their neighbours, such as the [ɛ] variant in words of the lexical set STRUT (Huber 2004: 851–852), the ubiquitous and sociolinguistically stratified T-affrication (Brato and Huber 2008) or request forms (Anderson 2006). Others are shared with and commonly found across varieties of English in Africa (cf. Brato and Huber 2012). See Section 6 for examples of lexical and grammatical innovations drawn from HiCE Ghana.

Ghana is currently on the verge of moving into the endonormative stabilization phase (Huber 2014: 90). While many Ghanaians still struggle to accept anything but Received Pronunciation and Standard British English as the norm (Anderson 2009: 30–33), actual practice draws a very different picture. Codification of the variety is ongoing and several dictionaries (Kirby 1998; Dako 2003; Blench 2006)² have been published, highlighting so-called Ghanaianisms on the semantic, grammatical and pragmatic levels (see Brato, this volume for a corpus study on the lexical expansion of GhE). Ghana has a vibrant literary scene of writers publishing their work in English and a very recent development is the growing number of children acquiring English as their first or only language.

3. Motivation for compiling a historical corpus of GhE

There exists today a wealth of linguistic data from Englishes all over the world, most notably thanks to the International Corpus of English (ICE; Greenbaum 1991) project, which currently counts 27 national components being available or compiled (The ICE Project 2016). As all components are compiled using the same design, ICE allows for a direct comparison across varieties and genres to uncover patterns shared by most varieties and, of course, those not shared. Since then, several other corpora of or including New Englishes, mostly varieties in which English used to be spoken as a second language following colonisation by the British, have been compiled, of which the Corpus of Global Web-based English (GloWbE; Davies and Fuchs 2015) is by far the largest. There are also more specific corpora, for example on news language, such as the News on the Web (NOW; Davies 2016) corpus or the Corpus of Namibian Online Newspapers (CNamOn; Kautzsch this volume) or ‘Cyber-Jamaican’ (Mair 2011).

All these corpora have in common that they have been designed for synchronic analyses, that is describing patterns of variation at a single point in time, and cannot be applied directly to uncover diachronic developments in a variety or test

2. Incidentally, all these are written by linguistic outsiders, i.e. non-Ghanaians.

assumptions laid out in models of the linguistic evolution of Postcolonial Englishes such as Schneider's (2007) Dynamic Model. Several studies have therefore compared varieties of English which have progressed to different stages in Schneider's model. Mukherjee and Gries (2009), for example, compare verb complementation patterns in Hong Kong (between phases 2 and 3), Indian (between phases 3 and 4) and Singaporean English (phase 4) and find that the further a variety has developed, the more dissimilar it is from British English. This is mirrored by Loureiro-Porto (2016), who studied (semi-)modals in Hong Kong, Indian and British Englishes. She argues that as Indian English progressed to phase 3 about 100 years before Hong Kong English more distinct features could be expected. While this method can help uncover general trends in linguistic change along the different phases of the Dynamic Model, it cannot be used for testing assumptions about how during one phase the English of a single variety has developed. One way out was proposed by Fuchs and Gut (2015) in their paper on the progressive aspect in Nigerian English. In the absence of real-time data, they applied the apparent-time approach to model how this feature has developed over time.

To overcome these limitations and allow for real-time comparisons of different states of one or more varieties, scholars have recently started compiling diachronic or historical corpora of New Englishes. Two diachronic corpora are currently being compiled. The Corpus of Historical Singapore English aims to provide 10-year interval snapshots from 1951 to 2011 (Hoffmann 2013), the Diachronic Corpus of Hong Kong English (Biewer et al. 2014) takes snapshots from around 1931, 1961, 1991 and 2011, following the design laid out by Francis and Kučera (1964) for the Brown corpus and thus allowing for the comparison to British and American Englishes from the same time frame. Diachronic analyses of Black South African English have recently made use of newspaper and fiction data from several periods covering the 1880s to 2012 (van Rooy & Wasserman 2014; van Rooy & Piotrowska 2015). Phil-Brown is a historical corpus of Philippine English. It covers the early 1960s and has been used recently to model diachronic change in modality (Collins, Borlongan & Yao 2014), the progressive aspect (Collins 2015) and expanded predicates (Borlongan & Dita 2015) by comparing it to relevant sections of the Philippine component of ICE.

When in early 2013 I first began to think about the possibility of exploring the historical dimension of GhE, two main ideas were taken into consideration. I was made aware of a project underway at the Ghana Broadcasting Corporation (GBC) aiming at digitising parts of their archive of speeches, announcements, interviews and broadcasts mainly by Kwame Nkrumah, one of the leaders of the Independence Movement and first Prime Minister and President of Ghana. Also, there is a project referred to as the 'Gramophone Archive', led by a musicologist, for which parts of the music collection have been digitised. These recordings also include spoken data

from as far back as the 1940s (also cf. Huber 2017: 446–51). However, for various reasons, mainly time and financial constraints, this idea was not pursued further.

Instead I decided to create a corpus of historical written GhE to provide an empirical basis for the analysis of structural nativization in this variety. Specifically, I am interested in finding answers to the following research questions: Is there linguistic evidence that Ghanaians have nativized English as proposed by Schneider (2007), that is has it followed the prototypical path suggested in the model?³ (How) does the virtual absence of an STL strand influence structural nativization and the related question of who provides the norms in this case? Does the orientation remain exonormative or are writers more likely to accept a local form? How does the massive increase in speakers of English influence the variety?

4. The compilation of HiCE Ghana

After the decision to compile a historical corpus of English in Ghana two questions had to be considered. What should be the design of the corpus and from which period or periods should data be collected? I was aware that a diachronic corpus of Singapore English (Hoffmann, Sand & Tan 2012) was being compiled at that time, which, as mentioned above, follows the Brown design and may therefore be used to directly compare British, American and Singaporean English from the 1960s (Brown/LOB) and the 1990s (Frown/FLOB). On the other hand, the compilation of the Ghanaian component of ICE (Huber and Dako 2013) was already well underway at the time and most of the written sections had already been completed. ICE Ghana mainly covers the end of the nativization period (most texts are from the early 2000s) and could therefore either be integrated into a diachronic corpus of GhE or serve as a reference for a historical corpus from an earlier period of English in Ghana.

An initial idea was to compile a diachronic corpus of GhE either following the Brown design or trying to mirror the written-printed categories of ICE. Snapshots of 200,000 words were to be collected in ten-year intervals from around 1960–62 to 2010–12, thus totalling about 1,200,000 words. The first step therefore was trying to locate possible samples to include in the corpus and making sure that I could get hold of them without encountering the “corpus compiler’s nightmare” that

3. There is some evidence that this may not be the case for phonetic-phonological features. Brato and Huber (2008) show that social variation is clearly present in the STRUT vowel and regarding r-affrication. My own analyses of the progressive aspect (see Section 6 for a brief discussion) indicate that innovative forms were already well established in the early nativization phase.

Hoffmann (2013) reports in the context of Singapore, that is running into major copyright and availability issues. Several archives and libraries were identified as probably providing a good range of material, amongst these the Balme Library at the University of Ghana, the Public Records and Archives Administration Department in Accra, the British Library and the library of the African Studies Centre Leiden in the Netherlands as well as a comparatively large number of libraries and archives in Germany. In addition, I included a range of online sources, such as archives of journals published in Ghana. For the sake of convenience, the German libraries were tapped in first. In the end, it turned out that these would provide enough data on almost all categories and that a trip to Ghana, London, or Leiden would not be necessary as material for categories for which the data basis was thin would not be available in sufficient numbers there either.

Three major tools were used to find possible material for the inclusion in the corpus: WorldCat (OCLC Online Computer Library Center 2017), which is a meta library catalogue, several library catalogues of German universities and archives, particularly those specialising in African studies. Most of the catalogues allow for very fine-grained search queries. To make sure that the results were as exhaustive as possible I used a combination of the year range I was interested in and place of publication, for which I compiled a list of about 40 towns and cities in Ghana. A tool that proved invaluable for locating newspapers, journals and other periodical titles is the *Zeitschriftendatenbank* (journal database; Staatsbibliothek zu Berlin – Preußischer Kulturbesitz 2017), the world-largest database of periodicals. It provides information about which German library stocks the periodical and which volumes are available. All books, journals, magazines, pamphlets and other printed material possibly of interest for inclusion in the corpus were entered into a reference manager to keep track.

The largest collection of Africana in Germany and a corpus compiler's dream is the special collection on 'Africa South of the Sahara' held by the university library in Frankfurt. It contains about 200,000 units collected since 1964 and covers a wide range of subjects mainly from the humanities, social sciences as well as government publications (Universitätsbibliothek J. C. Senckenberg 2016). In addition, it contains several issues of contemporary magazines aimed at a broader reading public and administrative writing such as university handbooks or business guides and travel literature. This collection proved invaluable for the corpus compilation and provided me with about a third of the material. Most of the newspaper data was collected from two libraries. Several libraries, most notably the *Staatsbibliothek zu Berlin: Preußischer Kulturbesitz*, which hosts the special collection on newspapers from around the world, have comprehensive newspaper archives. Almost complete volumes are held for two of the three national daily newspapers, the *(People's) Daily Graphic* (since 1950 to today) and the *Ghanaian Times* for the years 1964 to 1979

as well as the (*Sunday*) *Mirror*, a weekly newspaper, since 1964 to today) between four libraries. In addition, I had access to several volumes of the *The Spokesman* and the *Star*, two nation-wide contemporary newspapers. For many of the other newspapers of the time, at least individual issues were located and included in the corpus. Comprehensive archives or at least several issues were also found for many of the magazines and periodicals published at the time. For the year 1972–73 the Ghana Who's Who (Bartels 1972: 256f.) lists three daily newspapers, 14 weeklies, three fortnightlies, six English-language monthlies, four bi-monthlies and two quarterlies. For some of these several volumes were available, for other at least individual issues could be located.

Following this comprehensive search, I conducted a 'pilot collection' to see whether the goals mentioned above were realistic. It quickly became clear this was not the case. The Brown design proved impractical as it was rather too fine-grained and not always appropriate for the Ghanaian context (there is no such thing as detective stories or science-fiction and there were very few biographies or memoirs). The idea to create the envisaged diachronic corpus could not be pursued further either. There is a severe paucity of published material in the years after independence so that the corpus would not contain much data from the 1960–62 period. Just how small the potential data basis is, becomes obvious in this quotation: "[I]n 1966, Ghana produced a total of 225 books and pamphlets" (Apponsah 1968: 4).

From this already small number we must deduct all the works published in a Ghanaian language. While overall the situation was much better for the 1970–1972 period, some of the categories were still virtually impossible to fill to any substantial degree.

For this reason, I took the decision to do away with the idea of a diachronic corpus and compile a synchronic corpus of about 600,000 words of GhE covering the period between 1966 and 1975 based on the design of the ICE corpora. From a sociohistorical point of view, 1966 marks an important break in the history of Ghana. In February that year, Kwame Nkrumah was ousted in a coup d'état and exiled from the country. Nkrumah was one of the leading politicians long back into the colonial era. So while the declaration of independence in 1957 marks the onset of the nativization phase, the coup can be regarded as a kind of (minor) Event X (cf. Schneider 2007: 49), in which the ties to the colonial days were finally cut off. It was therefore deemed an appropriate starting point for the collection period. Another reason is the assumption that structural nativization follows the S-curve pattern of language change (as Schneider 2003: 248 suggests). Therefore, this period should cover the initiation and expansion stages, that is we should be able to find both conservative and traditional forms along with new innovative forms creeping in.

1975 as the cut-off point was not so much determined by theoretical, but rather methodological considerations as it was felt that a ten-year period could still be

considered largely synchronic and this way I could also ensure that I would have enough data to fill most of the genre categories substantially. The corpus was to be large enough to allow for synchronic analyses across genres, but at the same time should be usable for diachronic comparisons to ICE Ghana. Wherever possible, as for ICE, text samples of 2,000 words were included. The authors had to be Ghanaian and at least have had their primary and secondary education in the country. Also, all texts included in the corpus were to be published in Ghana so that the effect of non-Ghanaian editing could be minimised. It was clear that only printed material could be included in the corpus as any non-printed material would be impossible to get hold of. As the 300,000 words that make up the written-printed sections of ICE was deemed too small to find more than anecdotal evidence for many of the lower-frequency lexical and morphosyntactic peculiarities, the size of the corpus was going to be considerably larger.⁴ However, as I show in Brato (this volume), even at 600,000 words the corpus does not allow for detailed lexical investigations of a more quantitative nature.

While I was going to keep the basic structure of the ICE corpus design, it soon became clear that I would not be able to just double or triple the number of words in each category because too little material exists. This was especially true for the *Natural Sciences* and *Technology* categories in both academic and popular writing. On the other hand, for some of the categories there was so much data that it was deemed appropriate to go for a more fine-grained, more representative division. Press reports were readily available, so I decided to enlarge this category and break it down into four sections: political, cultural, regional and sports to better mirror the distribution I found. In a similar fashion, administrative writing was divided into government and non-government and creative writing into novels and stories. I also decided to strongly increase the opinion category, which in ICE only consists of editorials, and included letters to the editor as an additional category. Counting 50,000 words, the latter is rather substantial in HiCE Ghana. As including non-printed material was not viable, I felt that these letters would come as close as possible to a rather informal style of writing. In addition, this of course meant including also writers who would not be on the extreme upper end of the social spectrum. To make the comparison between the two corpora more straightforward, I therefore added a small number (10,000 words) of letters to the editor to the ICE data. Table 1 shows the corpus designs and word figures for HiCE Ghana and ICE Ghana.

4. Of course, this problem is also encountered in ICE Ghana. However, here we can resort to GloWbE or NOW if we want to check the occurrence of a particular low-frequency item.

Table 1. Comparison of the Corpus Designs and Word Counts of HiCE Ghana and ICE Ghana

Category	HiCE Ghana	ICE Ghana
Academic Writing	120,000	80,000
Humanities	30,000	20,000
Social science	30,000	20,000
Natural science	30,000	20,000
Technology	30,000	20,000
Popular Writing	120,000	80,000
Humanities	30,000	20,000
Social science	30,000	20,000
Natural science	30,000	20,000
Technology	30,000	20,000
Press reportage	100,000	40,000
Political	50,000	
Cultural	10,000	
Regional	20,000	
Sports	20,000	
Instructional Writing	80,000	40,000
Administrative (Government)	20,000	20,000
Administrative (non-Government)	20,000	
Skills & Hobbies	40,000	20,000
Opinion	100,000	30,000
Press editorials	50,000	20,000
Letters to the editor	50,000	10,000
Creative writing	80,000	40,000
Novels	40,000	
Stories	40,000	
Total	600,000	310,000

A major issue in the compilation of a corpus of a postcolonial variety of English is that of representativeness. Who should be included in the corpus and which sources can be tapped to provide a balanced sample? Speakers or writers to be included in the ICE corpora should at least be 18 years of age, born in the country in question (or have moved there at an early age) and have been educated to at least secondary level using English as a medium of instruction (Nelson 1996: 28).

There are several problems applying these criteria in the context of HiCE Ghana, which I will address in turn. The corpus covers the period between 1966 and 1975, that is to be included in the corpus authors would have to be born in or before 1948 and 1957 respectively and have had their secondary education

completed. In 1966, the population of Ghana stood at about 7.9m, by 1975 this figure had risen to about 9.8m (The World Bank Group 2017). According to Sey (1973: 1f.), in 1966 of the adult population of about 4.3m about 20% were educated, whereby educated is defined as

having at least completed a course of formal instruction in the primary and middle schools in Ghana. This course is ten years in duration and the medium of instruction is mainly English. A very small proportion of middle school pupils trickle out into secondary, technical and commercial schools, as well as teacher training colleges. Only a few of these eventually continue their education in universities and other institutions of higher education.

He goes on to state that

[e]very educated Ghanaian speaks some English, the degree of proficiency ranging from a sub-Basic type (i.e. with a very limited range of vocabulary and structures) to the most sophisticated Standard English varieties of the few highly educated people. (Sey 1973: 1)

This means that there are about 800,000 potential users of English⁵ at the beginning of the relevant period. However, given his comments the number of proficient writers was much lower. Besides, particularly writers in the academic genres would be considerably older than 18 as they will have at least completed (large parts of) their tertiary education before publishing. Here, taking into account the education statistics for the 1948 census (The Government of the Gold Coast 1950: 18) may give us a better picture. Only about 5% of Gold Coasters aged 10 or above (approx. 99,000) were considered educated, whereby educated means having attended school at least until the age of ten. The enrolment in secondary schools was about 4,150 (Sey 1973: 1, fn. 1).

Therefore, an issue that is relevant in the context of HiCE Ghana is the question about authors having been educated abroad. Prior to 1948, when the University College of the Gold Coast (the forerunner of the University of Ghana) was founded, studying or doing a PhD automatically meant having to leave the country for a certain period of time. The author's sociobiographical data is usually very sketchy, so I can only assume that most of the younger Gold Coasters and Ghanaians included in the corpus have had their secondary in the country. Many have a degree from one of the three Gold Coast/Ghanaian universities, but the large majority of PhDs

5. Despite the name 'middle school' and a separate 'secondary school' in the Ghanaian education system at the time, ten years of formal education is here considered equivalent to secondary level education.

were awarded by British universities. For this reason, one cannot exclude these authors from the corpus. For one, this would have meant that – particularly in the academic genres – the already small data base would have been diminished by about 80–90%. Furthermore, these authors should be included because – as mentioned above – they held key positions in the country. Being educated abroad, they would usually have had prolonged contact with Received Pronunciation (RP) and may even have picked up some RP features and often spoke with little or no trace of a Ghanaian accent (cf. Huber 2017 for a discussion about this issue).⁶ Whether this also holds for morphosyntactic features is not clear.

Due to the low number of book publications, most of the data in the academic genre comes from a small range of journals with very specific foci. In the Academic-Humanities section these are *Asemka* (a literature journal) and the *Research Review* of the Institute of African Studies at the University of Ghana. The technology category mainly revolves around agriculture as much of the data comes from *The Ghana Farmer* and the *Legon Journal of Agricultural Science*. In the Academic-Natural Sciences category the clear majority of data stems from the *Ghana Journal of Science* and describes various aspects in the field of plant biology. Data for the popular genres comes from a larger range of sources, but it proved very difficult to find enough data for the natural sciences section so that two samples are only marginally relevant to this category.

As mentioned above, there are comprehensive archives of Ghanaian newspapers, so news reports from a wide range of sources and topics were included in the corpus. This section therefore is clearly representative of newspaper writing in Ghana at the time. Care was taken to not include any material from outside Ghana, such as from major agencies like Reuters or European newspapers and from the Ghana News Agency, which also relied on agency reports. These items were clearly marked in all newspapers included in the corpus. For many of the newspaper reports, particularly the shorter ones, no author information was provided. Where this was the case, I tried to minimise the chances of a non-Ghanaian author by mainly including items with a clear Ghanaian context, for example by focusing on reports on Ghanaian internal politics or news reports from the different regions of the country.

6. A case in point is Robert Kweku Atta Gardiner (1914–1994), a leading economist and later Deputy Executive Director at the UN Economic Commission for Africa (UNECA). He was educated in Ghana and Sierra Leone and later went on to do his BA and MA in London and Oxford. He can be heard in this interview from 1957, available on YouTube <<https://www.youtube.com/watch?v=XY2SrD1zAwM>, 2:28–3:31>.

Administrative writing from a wide range of sources and on various topics is readily available. Government material, for example, includes economic surveys, commission reports and policy outlines, non-government material comes from student handbooks, notes for contributors or the constitution of the Ghana National Youth Council. Most of the data in the skills and hobbies genre are shorter items gained from wide a range of sources such as recipes in magazines, horoscopes from newspapers or guide books.

Editorials and opinion pieces from all available newspapers were included, supplemented by data from magazines and journals. As mentioned above, there is an additional category of 50,000 words of letters to the editor. Including these items makes the corpus more representative as many more writers could be included and as these letters come as close to a more informal style of writing as possible. In total, 181 letters of different lengths on a wide of range of topics and with very different levels of English proficiency were included in the corpus.

Novels and stories also cover a good range of sources and levels of proficiency in English. They mainly stem from lesser-known writers as novels from authors like Ama Ata Aidoo and Kofi Awoonor were published by the major European and American publishing companies and therefore were excluded from the corpus to minimise any non-Ghanaian editing.

5. Digitisation and annotation

Most of the material included in the corpus was photographed, scanned or copied from the original sources. Only very few items could be collected directly in digital format, for example from digitised books or journal articles. Therefore, before the data could be annotated it was necessary to digitise the data and apply optical character recognition (OCR). This brought with it several additional challenges and proved extremely time-consuming. Whenever possible the material was directly scanned or in the case of the newspaper photographed, but in many cases the intermediate step of photocopying had to be taken, which affected the quality of the digital copy. Another major issue was the quality of the print, particularly for the newspapers but also for many of the other sources. This made the OCR extremely error-prone and required a lot of manual post-editing and several correction cycles. An example of the typical quality of a newspaper article and the initial OCR output are shown in Figure 1. As can be seen there are several issues, some of which are obvious and will be recognised by a spell checker such 'Reaion' (line 4) or 'Iirand' (line 9). Also, the software may be trained on these rather typical OCR errors. Many of the other errors are relatively clear to the human reader and may be corrected rather straightforwardly, such as the extra spaces in line 11. Others, particularly those in which the sentence remains grammatical may go unnoticed.

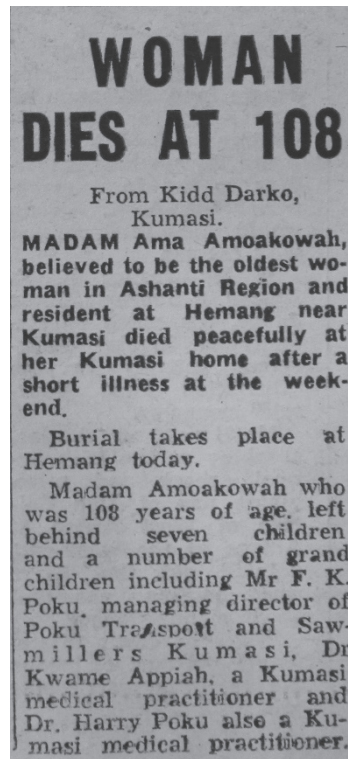


Figure 1. Example of newspaper article and OCR output

1. WOMAN DIES AT 108
2. From Kidd Darko, Kumasi.
3. MADAM Ama Amoakowah, believed to be the oldest
4. woman In Ashanti Reaion and resident at Hema”” near
5. Kumasi died peacefully at her Kumasi home after a short
6. illness at the weekend.
7. Burial takes place at Hemang today.
8. Madam Amoakowah who was 108 years of age. left be-
9. hind seven children and a number of Ilrand children in-
10. cluding Mr F. K. Poku managing director of Poku’
11. Trl’lIsoo,t and Saw· mill e r s K u m a s i, Dr Kwame
12. Appiah. a Kumasi medical practitioner and Dr. Harry
13. Poku also a Xumasi medical practitioner.

The annotation of the corpus was carried out in Pacx (Platform for Annotated Corpora in XML). Pacx is an annotation tool developed in the context of the compilation of ICE Nigeria and has since been used in the creation of several other corpora (Wunder, Voormann & Gut 2010: 79). Pacx uses the developer environment Eclipse and extends it with several plugins, like an XML editor and image

viewer. It uses an annotation schema that can be set up or adapted in a template file. One can, for example, create tags for marking items from particular languages or quotations. The annotation process is aided by a consistent and intuitive folder structure. The raw data, such as image or pdf files, is loaded into the *raw* folder. Pacx will then automatically create a corresponding xml file in the *annotated* folder based on the template, which consists of a header and placeholders for the title and main text. The header can be used to add metadata on the text. It was adapted from the original file and can include the following information: Name of the transcriber, date and place of publication, the author's name, birthplace, gender, age and ethnic group. The annotation itself is carried out by highlighting words or passages and selecting the appropriate tag, which is then visible in the annotation and stored in a well-formed xml file. There are several other advantages, such as a word count or the TODO mark-up, which allows keeping track of the revision cycles. Furthermore, during all stages Pacx also provides txt-formatted output of the corpus so that initial analyses can be carried out more easily.

6. Opportunities

HiCE Ghana allows the researcher to tap into a previous state of English in Ghana. It can be used for the synchronic analysis of a range of qualitative and quantitative lexical and grammatical features and in combination with ICE Ghana it allows tracing how GhE has changed in real time. It will therefore also provide the opportunity for testing evolutionary models of language development and showing if and how the variety has developed in the way suggested.

During the compilation of HiCE Ghana, I constantly took notes on linguistic phenomena that struck me as peculiar. A first qualitative analysis confirms that many of the features that have been commented on in the current literature on GhE are also found in the historical data, such as the different use of articles as in (1) and (2), clear examples of learned set phrases as in (3) or constructions, which are likely due to L1 transfer as in the resumptive adverb *here*⁷ as in (4), usually placed at the end of a syntactic unit. This sentence also exemplifies another quite typical feature of Ghanaian writing in the earlier data, using a local term (*Ya-Na*) and providing an English gloss straight afterwards (*Paramount Chief*).

- (1) Here's $\emptyset$ seat for you. (HiCE-Story026)
- (2) It all happened at 1.05 p.m. when three armed robbers entered the bank and at *a* gun point made away with the money. (HiCE-PrReg011)

7. This term is used here in analogy to resumptive pronouns.

- (3) At festivals, durbars, etc., where *every Tom Dick and Harry* is enjoying himself [...]
(HiCE-SkiHob001)
- (4) Colonel Iddisa made this statement when he called on the *Ya-Na* Mahama Abudulai VI, *Paramount Chief* of the Dagomba Traditional Area, at his palace *here*.
(HiCE-PrPol102)

Several quantitative linguistic analyses using HiCE Ghana have already been carried out or are currently underway, some of which are briefly outlined below. One of the key processes in the evolution of New Englishes is lexical expansion, that is the increase in localised vocabulary as a variety progresses further away from the colonial input (Schneider 2007: 78–82). Borrowings are no longer restricted to flora, fauna and cultural items as was mainly the case before and all kinds of word-formation processes are commonly employed to expand the vocabulary (cf. Biermeier 2008 for a comprehensive study).

(5) is an example from HiCE Ghana that has not been attested before, the semantic widening of the term *inmate* referring not only to prisoners but any other person who happens to be in the house or neighbours living in the same house.

- (5) It is customary for the guest to greet the host and the *inmates* as well.
(HiCE-SkiHob001)

Based on more than 4,000 previously attested so-called Ghanaianisms (Dako 2001) and by applying several other techniques I carried out a semantic and structural analysis of lexical expansion by comparing data from HiCE Ghana to ICE Ghana. While the corpora are comparatively small for lexical research, the results still provide a useful starting point for our understanding of how GhE has evolved over 40 years (see Brato this volume for a more comprehensive analysis).

Many studies have commented on differences in the usage of the progressive aspect between New Englishes and the traditional L1 varieties. This includes the general overuse of the progressive and the extension to stative verbs or habitual contexts as in (6). Interestingly, this is not mentioned in some of the earlier research on English in Ghana (Mensah 1971; Sey 1973; Tingley 1981), but is found comparatively frequently in HiCE Ghana, thus suggesting that not only it was already relatively common in the early stages of nativization but also that it was likely below the radar of contemporary (local) linguists. This could be an indication of it already being so entrenched that it was no longer considered an error but an acceptable feature of GhE.

- (6) Let's do something about the plight of these typists in the civil service, especially about those who are stationed in the cities and *are renting* houses costing not less than NC8.00 per month.
(HiCE-LetEd100)

A feature that is often commented upon in World Englishes is the nonstandard usage of modals. In the Ghanaian context, Huber and Dako (2004: 856) note that “*Would* commonly expresses definite future”, as in (7).

- (7) While this match goes on in Accra, four very interesting upper league matches *would* also be played in Koforidua, Sunyani, Kumasi and Sekondi.
(HiCE-PrSpo047)

While the coding and the analysis of the modal data is still ongoing, a first trend that shows in the corpus data is that this feature is in fact not very frequent. Another is the use of *will* in contexts in which British Englishes uses *would* as in (8). Sey (1973: 35) suggests that this may be because the past tense functions of modals are taught only relatively late, by which time the pupils have often developed fixed systems for the uses of present and past tense forms.

- (8) I *will* like to mention here that... (HiCE-AcHum-010)

New verb-complementation patterns are a common innovation on the lexis-grammar interface, as in (9) where the otherwise intransitive phrasal verb *sell out* is used transitively, meaning ‘sell’.

- (9) About 6,500 seeds of the plant were *sold out* or distributed from Aburi in 1899.
(HiCE-AcTech010)

A wealth of new fixed expressions enters the language and become established. It is very common, for example, to read the construction *allow/permit me a space* in letters to the editor from the 1960s and 1970s as in (10).

- (10) Would you please *allow me a space* to write about the problems facing the lower income group in your popular paper?
(HiCE-LetEd169)

Another example is the overuse of fixed expressions such as *all and sundry* (‘everyone/everybody’) as in Example (11).

- (11) United – a club whose strategy in “power” soccer once baffled *all and sundry* – has given way to inexplicable weakness.
(HiCE-PrSpo032)

There is also a tendency, particularly in the news reports, to use syntactically overly complex sentences as in (12), which Mensah (1971: 9) attributes to linguistic showmanship, “a hang-over from our school-days when the measure of good English was the number of ‘idiomatic’ phrases with which we seasoned our essay.”

- (12) The schools, to be set up in Tema, Accra and Takoradi, will afford the seamen the chance of knowing how to read and write so that their presence might be felt wherever they might be.
(HiCE-PrPol003)

An exploratory analysis using a web-based syntactic complexity analyser (Lu 2010) will hopefully shed some light on whether this is a more general trend towards complex sentences on a more general level and across the two corpora or if this is just anecdotal evidence.

7. Conclusion and outlook

In this chapter, I have discussed the major theoretical and methodological considerations and challenges in the compilation of a historical corpus of a New English. The empirical study of the development of postcolonial Englishes is still a very recent phenomenon and there clearly is a desideratum for authentic data from historical states of postcolonial Englishes. Corpora like HiCE Ghana therefore are a much-needed addition to the linguist's toolkit.

Scholars have taken different approaches to compiling historical and diachronic corpora. I hope to have shown that these approaches complement one another and that the decision for or against using Brown, ICE or some other format is determined by various factors. Following the Brown design allows for synchronic and diachronic comparisons across varieties of English. The approach taken in HiCE Ghana was motivated by the wish to provide a data basis for a diachronic comparison between the early and late stages of the nativization phase in GhE, that is variety-internal. My aim is to illustrate whether and how the Dynamic Model can cater for the specific linguistic situation in Ghana and how well it predicts the linguistic developments based on a real-time analysis.

Although initially the design was not meant to be adopted for other varieties of English, there is certainly potential for doing so. This is most likely true for Nigeria, which shares with Ghana the insignificant STL strand, a similar colonial history and the close proximity of independence and the onset of the nativization phase but may also work for countries like Tanzania and Kenya. Some modifications may be necessary to cater for specific situations in these countries (e.g. by trying to make sure to include a balanced sample of speakers from the major L1s) or because, as in Ghana, there may just not be enough data for a certain period.⁸ In the Ghanaian context, corpora based on HiCE Ghana could be created for one or more later periods (e.g. around 1980 and around 1990) to allow more fine-grained analyses of diachronic trends.

8. This may of course influence the comparability across corpora. However, fine-tuning the compilation process to cater for the local needs is very common both regarding the ICE corpora (cf. e.g. Bolt 1994; Biewer, Hundt & Zipp 2010) and those based on the original Brown design (e.g. Collins, Borlongan & Yao 2014).

References

- Anderson, Jemima Asabea. 2006. Request forms in English in Ghana. *Legon Journal of the Humanities* 17: 75–103.
- Anderson, Jemima Asabea. 2009. Codifying Ghanaian English: Problems and prospects. In *World Englishes – Problems, Properties and Prospects: Selected Papers from the 13th IAWC Conference* [Varieties of English around the World G40], Thomas Hoffmann & Lucia Siebers (eds), 19–36. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g40.05and>
- Apponsah, Kwamina. 1968. African writers must be up and doing. *The Evening News*, February 20.
- Bartels, Charles. 1972. *Ghana Who's Who, 1972–73: A Fully-Comprehensive Annual Biographical Dictionary of Prominent Men and Women in the Country, Includ. an Encyclopaedia of Useful Information*. Accra: Bartels Publishing.
- Biermeier, Thomas. 2008. *Word-Formation in New Englishes: A Corpus-Based Analysis*. Munster: LIT.
- Biewer, Carolin, Bernaisch, Tobias, Berger, Mike & Heller, Benedikt. 2014. Compiling the diachronic corpus of Hong Kong English (DC-HKE): Motivation, progress and challenges. *ICAME Journal* 35. Nottingham, April 30.
- Biewer, Carolin, Hundt, Marianne & Zipp, Lena. 2010. 'How' a Fiji Corpus? Challenges in the Compilation of an ESL ICE Component. *ICAME Journal* 34: 5–23.
- Blench, Roger. 2006. A Dictionary of Ghanaian English. <<http://www.rogerblench.info/Language/English/Ghana%20English%20dictionary.pdf>> (13 July 2016).
- Bolt, Philip. 1994. The International Corpus of English project: The Hong Kong experience. In *Creating and Using English Language Corpora: 14th International Conference on English Language Research on Computerized Corpora: Selected Papers*, Udo Fries, Gunnel Tottie & Peter Schneider (eds), 15–24. Amsterdam: Rodopi.
- Borlongan, Ariane Macalinga & Dita, Shirley N. 2015. Taking a look at expanded predicates in Philippine English across time. *Asian Englishes* 17(3): 240–47. <https://doi.org/10.1080/13488678.2015.1069965>
- Bourret, F. M. 1960. *Ghana: The Road to Independence, 1919–1957*, 3rd edn. London: OUP.
- Brato, Thorsten. In preparation. Extra- and intraterritorial forces in the development of Ghanaian English. In *Modelling Current Linguistic Realities of English World-Wide: The Extra- and Intra-Territorial Forces Model Put to the Test*, Sarah Buschfeld & Alexander Kautzsch (eds).
- Brato, Thorsten & Huber, Magnus. 2008. The emergence of social varieties in Ghanaian English. *Accents 2008: II International Conference on Native and Non-native Accents of English*, Łódź, December 12.
- Brato, Thorsten & Huber, Magnus. 2012. English in Africa. In *Areal Features of the Anglophone World*, Raymond Hickey (ed.), 161–85. Berlin: Mouton de Gruyter.
- Buschfeld, Sarah & Kautzsch, Alexander. 2017. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36(1): 104–26. <https://doi.org/10.1111/weng.12203>
- Collins, Peter C. 2015. Recent diachronic change in the progressive in Philippine English. In *Grammatical Change in English World-Wide*, edited by Peter C. Collins, 271–96. *Studies in Corpus Linguistics* 67. Amsterdam: John Benjamins.
- Collins, Peter C., Borlongan, Ariane Macalinga & Yao, Xinyue. 2014. Modality in Philippine English: A diachronic study. *Journal of English Linguistics* 42(1): 68–88. <https://doi.org/10.1177/0075424213511462>

- Dako, Kari. 2001. Ghanaianisms: Towards a semantic and formal classification. *English World-Wide* 22(1): 23–53. <https://doi.org/10.1075/eww.22.1.03dak>
- Dako, Kari. 2003. The sound system of Ghanaian English. *Exploration* 1(1): 107–119.
- Davies, Mark. 2016. NOW Corpus (News on the Web). <<http://corpus.byu.edu/now/>> (22 June 2016).
- Davies, Mark & Fuchs, Robert. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word global web-based English corpus (GloWbE). *English World-Wide* 36(1): 1–28. <https://doi.org/10.1075/eww.36.1.01dav>
- Foster, Philip. 1965. *Education and Social Change in Ghana*. Chicago IL: The University of Chicago Press.
- Francis, W. Nelson & Kučera, Henry. 1964. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers (Brown)*. Providence RD: Brown University.
- Fuchs, Robert & Gut, Ulrike. 2015. An apparent time study of the progressive in Nigerian English. In Collins (ed.), 373–87.
- Ghana Statistical Service. 2012. 2010 Population & Housing Census: Summary Report of Final Results. Ms, last modified July 13, 2016. <http://www.statsghana.gov.gh/docfiles/2010phc/Census2010_Summary_report_of_final_results.pdf>
- Ghana Statistical Service. 2016. *Population Statistics*. <http://www.statsghana.gov.gh/pop_stats.html> (9 August 2017).
- Greenbaum, Sidney. 1991. ICE: The International Corpus of English. *English Today* 7(4): 3–7. <https://doi.org/10.1017/S0266078400005836>
- Hoffmann, Sebastian. 2013. *The Corpus of Historical Singapore English: Practical and Methodological Issues*. Trier: University of Trier.
- Hoffmann, Sebastian, Sand, Andrea & Tan, Peter K. W. 2012. The Corpus of Historical Singapore English: A first pilot study on data from the 1950s and 1960s. Paper presented at ICAME 33, Leuven, 30 May.
- Huber, Magnus. 2004. “Ghanaian English: phonology.” In *A Handbook of Varieties of English: A Multimedia Reference Tool. 1: Phonology*, edited by Edgar W. Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie, and Clive Upton. 3 vols, 842–865. Berlin: Mouton de Gruyter.
- Huber, Magnus. 2014. Stylistic and sociolinguistic variation in Schneider’s nativization phase: The case of Ghanaian English. In *The Evolution of Englishes: The Dynamic Model and Beyond* [Varieties of English Around the World G49], Sarah Buschfeld, Magnus Huber, Thomas Hoffmann & Alexander Kautzsch (eds), 86–106. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g49.06hub>
- Huber, Magnus. 2017. Early recordings from Ghana: A variationist approach to the phonological history of an outer circle variety. In *Listening to the Past: Audio Records of Accents of English*, Raymond Hickey (ed.), 444–463. Cambridge: CUP. <https://doi.org/10.1017/9781107279865.019>
- Huber, Magnus & Dako, Kari. 2004. Ghanaian English – Morphology and syntax. In *A Handbook of Varieties of English: A Multimedia Reference Tool*, Edgar W. Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 854–865. Berlin: Mouton de Gruyter.
- Huber, Magnus & Dako, Kari. 2013. *International Corpus of English, Ghana Component*. Giesen: University of Giessen.
- Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In *English in the World: Teaching and Learning the Language and Literatures*, Randolph Quirk & Henry G. Widdowson (eds), 11–30. Cambridge: CUP.

- Kirby, Jon P. 1998. *A North American's Guide to Ghanaian English*. Tamale: Tamale Institute of Cross-Cultural Studies.
- Loureiro-Porto, Lucía. 2016. (Semi-)modals of necessity in Hong Kong and Indian Englishes. In *World Englishes: New Theoretical and Methodological Considerations* [Varieties of English Around the World G57], Elena Seoane & Cristina Suárez Gómez (eds), 143–72. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g57.07lou>
- Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4): 474–96. <https://doi.org/10.1075/ijcl.15.4.02lu>
- Mair, Christian. 2011. Corpora and the New Englishes. In *A Taste for Corpora* [Studies in Corpus Linguistics 45], Fanny Meunier, Sylvie de Cock, Gaëtanelle Gilquin & Magali Paquot (eds) 209–36. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.45.13mai>
- Mair, Christian. 2013. The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34(3): 253–78. <https://doi.org/10.1075/eww.34.3.01mai>
- Mensah, A. N. 1971. The language of Ghanaian newspapers. *Legon Observer*, October 22.
- Mufwene, Salikoko S. 2001. *The Ecology of Language Evolution*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511612862>
- Mukherjee, Joybrato & Gries, Stefan T. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1): 27–51. <https://doi.org/10.1075/eww.30.1.03muk>
- Nelson, Gerald. 1996. The design of the corpus. In *Comparing English Worldwide: The International Corpus of English*, Sydney Greenbaum (ed.), 27–35. Oxford: Clarendon Press.
- OCLC Online Computer Library Center. 2017. *WorldCat*. <<https://www.worldcat.org>> (11 August 2017).
- Odamtten, Helen & Larya, E. B. 2009. *Steps to Effective Spoken English (SESEN)*. Accra: Asempa Publishers.
- Platt, John Talbot, Weber, Heidi & Ho, Lian-Mian. 1984. *The New Englishes*. London: Routledge and Kegan Paul.
- Schneider, Edgar W. 2003. The dynamics of New Englishes: From identity construction to dialect birth. *Language* 79(2): 233–281. <https://doi.org/10.1353/lan.2003.0136>
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511618901>
- Sey, Kofi Abakah. 1973. *Ghanaian English: An Exploratory Survey*. London: Macmillan.
- Staatsbibliothek zu Berlin – Preußischer Kulturbesitz. 2017. *Zeitschriftendatenbank*. <<http://zdb-opac.de>> (10 August 2017).
- The Government of the Gold Coast. 1950. *The Gold Coast: Census of Population 1948 – Report and Tables*.
- The ICE Project. 2016. *International Corpus of English*. <<http://www.ice-corpora.net/ice/>> (28 July 2017).
- The World Bank Group. 2017. *Ghana Population, 1960–2016*. <<http://data.worldbank.org/indicator/SP.POPTOTL?locations=GH>> (16 August 2017).
- Tingley, Christopher. 1981. Deviance in the English of Ghanaian newspapers. *English World-Wide* 2(1): 39–62. <https://doi.org/10.1075/eww.2.1.04tin>
- Trudgill, Peter. 2004. *New-Dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: EUP.
- Universitätsbibliothek J. C. Senckenberg. 2016. *Africa South of the Sahara*. <https://www.ub.uni-frankfurt.de/afrika/afrika_en.html> (10 August 2017).

- van Rooy, Bertus & Piotrowska, Caroline. 2015. The development of an extended time period meaning of the progressive in black South African English. In Collins (ed.), 465–83.
- van Rooy, Bertus & Wasserman, Ronel. 2014. Do the modals of black and white South African English converge? *Journal of English Linguistics* 42(1): 51–67.
<https://doi.org/10.1177/0075424213511463>
- Wunder, Eva-Maria, Voormann, Holger & Gut, Ulrike. 2010. The ICE Nigeria Corpus Project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.

Addressing a coverage gap in African Englishes

The tagged corpus of Cameroon Pidgin English

Gabriel Ozón¹, Sarah FitzGerald² and Melanie Green²

¹University of Sheffield / ²University of Sussex

This paper illustrates the uses of a tagged corpus of spoken Cameroon Pidgin English (CPE), which has recently been finalised (Ozón et al. 2017) and made available online (Green et al. 2016). The corpus consists of 240,000 words, with mark-up and part-of-speech-tagging. Text categories and proportions of monologue/dialogue are guided by those of the ICE project (Nelson 1996), making the CPE corpus comparable with existing corpora of post-colonial Englishes. This tagged corpus offers an invaluable resource for the investigation of CPE, particularly in addressing issues of multifunctionality in pidgin or creole languages. We introduce the dataset and present case studies illustrating its potential uses, in order to highlight the usefulness of this freely accessible resource for research on African languages.

1. Introduction

The shortage of corpora representing African varieties of English (compared to other regional varieties of English) has been pointed out (see e.g. Nelson & Ozón 2017). However, steps have been taken recently to redress this imbalance. Within the ICE project itself, corpora of African varieties are either already available (Nigeria and East Africa, which includes Kenya and Tanzania) or are being compiled (Ghana, Uganda, Namibia). Other corpora have also been constructed, including the Historical Corpus of English in Ghana (Brato 2018), and the Corpus of Cameroon English (Tiomajou 1993; Nkemleke & Mbangwana 2007).

Most recently, a spoken corpus of Cameroon Pidgin English (CPE) has been finalised (Ozón et al. 2017) and made available on-line (Green et al. 2016). CPE is a widely-used, yet stigmatised and largely uncodified pidgin/creole variety. The corpus consists of 240,000 words, with mark-up and part-of-speech-tagging. The text categories and the proportions of monologue and dialogue are in line with those of the ICE project, making the CPE corpus directly comparable with existing corpora of post-colonial Englishes.

Our two main objectives in this paper are first to provide an overview of this resource, and secondly to illustrate its potential uses for linguistic research by means of a set of case studies. The paper is organised as follows: Section 2 offers some background on CPE, and Section 3 provides an overview of the CPE corpus, including some information about the process of constructing the corpus and the guiding principles behind that process. Section 4 offers some insights into the potential uses of the corpus, focusing on two main approaches, which we illustrate with case studies: corpus-based and corpus-driven approaches. Section 5 concludes the paper.

2. Background to CPE

Cameroon Pidgin English (CPE) is an expanded pidgin/creole spoken in some form by an estimated 50% of Cameroon's population of 22,000,000 (Simons & Fennig 2017), primarily in the Anglophone west regions, but also in urban centres throughout the country. As a primarily spoken language, CPE is largely uncodified, but enjoys a vigorous oral tradition, not least through its presence in the broadcast media. CPE is a stigmatised variety in the face of French and English, prestige languages of Cameroon, where it also co-exists with an estimated 280 indigenous languages (Simons & Fennig 2017).

A number of authors have published research on the sociolinguistic situation in Cameroon, including Todd (1982); Koenig, Chia & Povey (1983); de Féral (1989); Wolf (2001); Simo Bobda & Wolf (2003); Menang (2004) and most recently; Anchimbe (2013). Although it is not the most recent publication, Schröder (2003) remains the most comprehensive study of the sociolinguistic context of CPE.

Schröder (2003) found the highest concentration of proficient CPE speakers in the anglophone regions but established, nevertheless, that CPE is used in some form by a substantial proportion of speakers across Cameroon. Schröder (2003: 85) also found that the 0–15 age group had the lowest proportion of CPE speakers (50%), while the 50+ age group had the highest proportion (83.3%). In terms of attitudes to CPE, Schröder (2003: 54–58) reports a widespread view that CPE is detrimental to the acquisition of 'good English', a view expressed by both anglophone and francophone CPE speakers in her study. Schröder (2003: 64–70) also reports participants' views that CPE is contributing to the endangerment of Cameroon's indigenous languages, despite some ambivalence to multilingualism.

Schröder (2003) distinguishes the varieties of CPE according to a number of parameters: regional (anglophone vs. francophone), urban vs. rural, social varieties and situational varieties. With respect to anglophone and francophone varieties, Schröder (2003: 90–98) reports that CPE speakers can often distinguish the two varieties, and mentions phonological, lexical and morphosyntactic differences

reported by her participants. With respect to urban and rural varieties, she comments that urban varieties are more likely to show influence from the official European languages, while rural varieties are more likely to show influence from indigenous languages (2003: 101). Indeed, some of Schröder's participants described the rural CPE as 'unadulterated', recognising that urban CPE is more likely to be influenced by English. Urban varieties thus tend more towards the acrolect, while rural varieties are more basilectal. The urban-rural distinction also overlaps with social variation, particularly in relation to education. Education in rural areas tends to be limited to the primary level, while secondary and tertiary educational establishments are located in urban areas. With respect to situational variation, Schröder (2003: 115) observes that the key factor is accommodation to the interlocutor's language preferences and ability, and that educated CPE speakers may have command of a range of lects and use whichever is most appropriate to the situation.

Schröder concludes her study with a discussion of the pros and cons of a national language for Cameroon, and it is striking to note that 29.1% of her participants identify CPE as the most suitable candidate, although the majority (34.2%) responded that there was no Cameroonian language with potential national language status (Schröder 2003: 196). Schröder attributes the position of CPE in this survey to its relative ethnic neutrality (although many francophones consider it an anglophones' language) and its geographical spread, but also points out a number of drawbacks, among them the absence of a standardised orthography and, most significantly, the low social status of CPE, which is widely considered a form of 'broken English', not a 'proper' language, and a medium of communication for the uneducated (Schröder 2003: 206f.). We refer the reader to Anchimbe (2013) for a recent discussion of the complexities of language policy and identity in Cameroon.

3. Building the CPE Corpus

In this section, we present a brief overview of the design and construction of the CPE corpus, highlighting the guiding principles behind this process. We focus here on proportions of text types (Section 3.1), audio files (Section 3.2), participant metadata and research ethics (Section 3.3), the development of an orthography for a primarily spoken variety (Section 3.4) and annotation and tagging (Section 3.5).

The compilation of the spoken CPE corpus, the procedures employed in its tagging, and the spelling challenges and decisions adopted (including homogenisations) are discussed in greater detail in Ozón et al. (2017). In addition, the corpus files deposited in the Oxford Text Archive <<http://ota.ox.ac.uk/desc/2563>> include (i) a spelling guide, as well as (ii) a tagging guide, with detailed discussion of problematic tagging decisions, profusely illustrated with examples from the corpus.

3.1 Text types

The corpus consists of transcriptions of private and public dialogues and monologues and is freely available online <<http://ota.ox.ac.uk/desc/2563>>. Text categories and the proportions of monologue and dialogue were guided by those of the International Corpus of English (ICE) project <www.ice-corpora.uzh.ch/en.html>, which makes the corpus comparable with existing corpora of post-colonial varieties of English.

Table 1. Proportions of text types in the CPE corpus

CPE		Texts	Words	%
Dialogues	private	26	78,000	33
	public	21	63,000	26
Monologue	unscripted	18	54,000	23
	scripted	15	45,000	19
		TOTAL	240,000	

3.2 Audio

The corpus contains 80 sound recordings of monologues (scripted and unscripted) and dialogues (public and private). The recordings were conducted in five different locations in Cameroon (Bamenda, Buea, Douala, Kumba and Yaoundé). The collected data are equally distributed across the five locations, which allows for potential insights into regional variation.

Each sound file is 10–15 minutes in length and is available in two formats: high quality .wav format, which allows for detailed phonetic/acoustic analyses, and lower quality .mp3 format, which requires less storage space and is adequate for listening purposes. The sound files have so far been used to inform aspects of the description of the grammar of CPE (Ayafor & Green 2017), and to investigate the applicability of the concept of tone to CPE (Devlin et al., in prep.). As Green & Ozón (2018a) point out, the corpus also offers a rich resource for an investigation into the prosodic features that accompany the morphosyntactic marking of topic and focus in CPE. However, this valuable resource is to date underexploited.

3.3 Participant metadata and research ethics

Background information about the participants was collected by means of a questionnaire completed before each recording session. The categories of information are summarised in Table 2, and this information has been made available in the corpus release, allowing researchers to take into account not just regional but sociolinguistic variation in their research, as well as possible language contact phenomena.

Table 2. Participant information available

Birthplace	Education
Anglophone/Francophone	Occupation
Age	First language(s)
Gender	Other languages
Age first spoken CPE	Parents' languages

In accordance with the requirements of good research ethics, participants were provided with information and consent forms, and all personal information was fully anonymised in the transcriptions and in the metadata, with each participant identified by a unique code. All personal and place names were substituted in the transcriptions and bleeped out in the sound files.

3.4 Spelling

Ensuring a consistent orthographic representation of CPE was the earliest, most significant challenge. Since there is no standardised orthography for CPE, the orthography adopted for this project was based on that developed by Ayafor (2014), which was kept under review during the course of the project. At the end of the project, a spelling list was included with the corpus files.

One example of a spelling challenge was homophones. Given that a considerable number of high-frequency expressions are multifunctional in CPE, it was necessary to strike a balance between the distinction between homophones on the one hand, and a faithful transcription-based orthography on the other. In some cases, therefore, we decided against distinguishing homophones in the spelling, while in others we were able to introduce distinct spellings that were still reasonably faithful to our transcription system. Some examples follow.

- (1) a. *bi* /bi/ 'be', anterior tense particle
- b. *dey* /de/ pronoun (3PL.SBJ)
- c. *dei* /de/ 'day'
- d. *deiy* /de/ locative copula, 'there'
- e. *dem* /dem/ pronoun (3PL.OBJ/TOP/FOC), plural particle
- f. *foe* /fo/ preposition, infinitival particle
- g. *fo* /fo/ 'four'
- h. *goe* /go/ 'go', irrealis mood particle
- i. *na* /na/ non-verbal copula, focus particle
- j. *sey* /se/ 'say'
- k. *sei* /se/ complementiser
- l. *wey* /we/ 'way'
- m. *wei* /we/ relativiser
- n. *yi* /yi/ pronoun (3S.OBJ/TOP/FOC)
possessive determiner (3s)

3.5 Annotation and tagging

The recordings were transcribed (each sound file resulting in approximately 3,000 words of transcribed text) and simultaneously annotated by the transcriber. Based on the International Corpus of English (ICE) guidelines for the annotation of spoken texts (Nelson 2002), standard mark-up symbols were used to identify speakers, overlapping speech and uncertain transcriptions. Mark-up symbols were also used to identify the use of foreign and/or indigenous words in CPE. Words or phrases from European languages (i.e. English or French) were marked as ‘foreign’ where the transcriber judged that the expression was a loanword for which a near-synonym exists in CPE, or where the speaker was code-switching into English/French. Words from indigenous African languages were marked as ‘indigenous’ according to the same criteria. All these mark-up symbols are automatically searchable and allow fast retrieval of e.g. all instances of codeswitching (Section 4.3.5).

Since this was the first time that a corpus of CPE had been compiled and tagged, it was necessary to find a trainable, automatic tagger for the language. The Tree Tagger (Schmid 1994) was trained and tested using manually tagged CPE data, and reached a 94% accuracy rate. All 240,000 words in the CPE corpus have been tagged, and 60,000 have also been manually post-checked.

A tagged corpus is particularly useful for automatic retrieval of language phenomena above the level of the lexicon. For example, the tagging allows researchers to address issues of multifunctionality, such as the case of the form *foe*, which can function either as a preposition (tagged PRF in our corpus) or as an infinitive marker (tagged INF in our corpus) in CPE. Searching the corpus for the word form *foe* will not allow the category distinction to be made; however, the tagging allows for a combination of lexical and tagset searches, which makes it easy to extract all the *foe* forms that function as e.g. prepositions (illustrated in Table 3 below).

Table 3. Multifunctional distribution of *foe* in the CPE corpus

Word	Tag	N	%
<i>foe</i>	PRF	9,117	83
	INF	1,915	17
		11,032	

As this brief example illustrates, the tagged corpus can thus be used to quickly and easily generate large data sets of specific parts of speech rather than individual words.

4. Searching the CPE corpus

In this section, we begin by providing an overview of the different types of research allowed by the corpus-linguistic method, drawing an important distinction between corpus-based and corpus-driven approaches (Section 4.1). We then present a number of case studies based on the CPE corpus aimed at illustrating the corpus-driven approach (Section 4.2) and the corpus-based approach (Section 4.3).

4.1 Corpus methods

In languages that remain understudied, available datasets are often small and typically consist of elicitation pairs, Bible portions and/or orally narrated folk tales. A larger dataset, designed and compiled in order to be representative of the variety in question, allows many more questions to be addressed. The analysis of patterns present in such dataset/corpus serves the dual purpose of confirming not only what is possible in the language, but also what is preferred. Even before information is added to the raw language data (in the form of mark-up or grammatical tagging), recurrent patterns can be extracted from such a dataset.

Tognini-Bonelli (2001) draws the distinction between corpus-driven and corpus-based approaches to the study of language. The corpus-driven approach attempts to use corpus data free from the influence of existing theoretical frameworks (e.g. McEnery & Gabrielatos 2006). Research is carried out on raw, unannotated corpora, since annotation (including tagging) imposes an interpretation on the data. In contrast, the corpus-based approach relies on the full range of corpus data in order to explore a theory or hypothesis, aiming to validate it, refute it or refine it. Annotation (often based on a pre-conceived theory or taxonomy) is added to the raw corpus in order to reduce data to “a set of orderly categories which are tractable within existing descriptive systems” (Tognini-Bonelli 2001: 68).

We provide illustrations of both approaches in the following sections. The CPE corpus release contains both a raw corpus and an annotated corpus, so researchers can choose which approach suits their purposes best. Besides the actual words uttered by the speakers, the corpus contains different types of information that can be thought of as ‘handles’ on the data. Using freely available software, such as word processors and AntConc (Anthony 2014), researchers can conduct different types of searches for different types of patterns/information:

- Lexical searches (frequency information): corpus-driven
- Mark-up searches (e.g. monologue/dialogue, public/private, overlaps, code-switching): corpus-driven
- Metadata searches (e.g. region, age, gender, education): corpus-driven

- Combined searches (for example, searching for a word form with a particular grammatical function, e.g. forms functioning as serialising verbs): corpus-based

4.2 Corpus-driven approach: Case studies

In the corpus-driven approach, the absence of mark-up or POS-tagging is not detrimental but is seen as an advantage. This approach involves extracting and making sense of frequency information about the patterning of lexical items. In a sufficiently large corpus, patterns (words, phrases, constructions) will emerge, the significance of which is identified by the researcher.

4.2.1 Word frequency

In the CPE corpus, a simple frequency search reveals the ratio of lexical-to-grammatical words in the top-20 most frequent words (Table 4). This immediately shows a high proportion of multifunctional items e.g. *foe* (< Eng. *for*), which can be either a preposition or an infinitival marker, or forms such as *goe* ‘go’, which can function as (a) the lexical verb ‘go’, (b) a motion serial verb, or (c) a preverbal particle marking irrealis mood (Ayafor & Green 2017). CPE also lacks inflected forms, a feature which allows more space in the top twenty for distinct lexical items (rather than inflected forms of the same lexical item). As this simple search shows, the CPE list is heavily populated by grammatical elements such as determiners, pronouns and preverbal particles, and this offers a snapshot of the typological character of the language.

Table 4. Top 20 most frequent lexical items in the CPE corpus

Rank	Item	Gloss	Rank	Item	Gloss
1	<i>foe</i>	preposition; infinitival PVP	11	<i>noe</i>	negation PVP; indefinite determiner
2	<i>i</i>	3s subject pronoun	12	<i>dat</i>	demonstrative determiner
3	<i>di</i>	imperfective PVP	13	<i>dem</i>	plural marker; 3PL object/topic/focus pronoun
4	<i>yu</i>	2s subject/object pronoun	14	<i>deiy</i>	locative copula; ‘there’
5	<i>goe</i>	lexical/serial ‘go’; irrealis PVP	15	<i>soe</i>	‘so’
6	<i>na</i>	copula; focus particle	16	<i>dey</i>	3PL subject pronoun
7	<i>wei</i>	relativiser	17	<i>bi</i>	anterior PVP; ‘be’
8	<i>sei</i>	‘say’; complementiser	18	<i>wi</i>	2PL subject/object pronoun
9	<i>som</i>	indefinite determiner	19	<i>wan</i>	specific indefinite determiner; ‘one’
10	<i>de</i>	definite determiner	20	<i>don</i>	perfective PVP

4.2.2 Lexical diversity and richness

Type-token ratio (TTR), which measures the proportion of distinct words in a text relative to the total number of words in the text (Johansson 2008) can be used as a quick measure of lexical diversity. Written texts tend to have higher TTRs than spoken texts, because the latter tend to involve more repetition. In addition, the larger the text, the lower the TTR is likely to be, again because of the increased likelihood of repetition. The TTR for our CPE corpus is 0.03, a very low ratio even for spoken language, especially considering the small size of the corpus and the lexical diversity of the speakers. This indicates that our corpus does not show a high degree of lexical richness, which is consistent with the high degree of multifunctionality in the CPE lexicon. For comparison, the TTR of the London-Lund Corpus of spoken British English is 0.395 (Taylor et al. 2001).

4.2.3 High frequency verbs

By taking the tagging into account, it is possible to retrieve all the verbs in our corpus and compile a frequency list. Table 5 shows the top ten most frequent verbs in the CPE corpus.

Table 5. Top ten most frequent verbs in the CPE corpus

Rank	Verb	Gloss	% of all verbs
1.	<i>goe</i>	lexical/serial 'go'; irrealis pvp	6.9
2.	<i>meik</i>	lexical/serial 'make'	5.0
3.	<i>get</i>	lexical/light 'have'	4.8
4.	<i>kam</i>	lexical/serial 'come'	4.7
5.	<i>tok</i>	'talk; say'	4.1
6.	<i>si</i>	'see'	3.2
7.	<i>wan</i>	'want'	3.1
8.	<i>nou</i>	'know'	3.0
9.	<i>du</i>	lexical/light 'do'	2.4
10.	<i>teik</i>	lexical/serial 'take'	2.3

The most frequent verbs in this list (*goe*, *meik*, *get*, *kam*, as well as *teik*) are also multifunctional, in that these semantically general verbs are not simply used to convey lexical meanings, but also participate in larger (lexico-grammatical) constructions such as serial verb constructions (SVCs) and light verb constructions (LVCs). The position of *goe* at the top of the list is accounted for by its additional use as a preverbal marker of irrealis mood. Some illustrative examples follow, taken from Ayafor and Green (2017). Example (2) illustrates the lexical use of *goe*, as well as the verb *meik* in its light verb function, where it forms a complex predicate (light verb construction) with its nominal complement *kongosa* 'gossip'. In this light verb

construction, the verb lends minimal meaning to the complex predicate, while the main lexical burden is on the complement.

- (2) Lexical *goe*, light *meik*¹

a noe di goe mi foe som-man yi haus foe
 1S.SBJ NEG IMPF go 1S.FOC PREP PRO.INDF 3S.DET.POSS house INF
 goe meik kongosa! <Ku-MU-01:48>
 go make gossip
 ‘I don’t go to somebody’s house to gossip!’

Example (3) illustrates the serial use of *goe*. In this example, the main event verb in the embedded clause (in square brackets) is *ron* ‘run’, while the serial verb adds ‘motion away’ semantics within the monoclausal serial verb construction.

- (3) Serial *goe*

a wan [jos ron **goe**] <Be-DPr-03-A:3>
1s.SBJ want just run go
'I just want to run away.'

Example (4) illustrates the use of *goe* as a marker of irrealis voice, which in the given context is used to refer to future time.

- (4) *goe* as irrealis preverbal particle

wi goe stat foe luk insait Baibel <Do-MS-02:28>
lPL.SBJ IRR start INF look PREP Bible
'We'll start looking in the Bible.'

1. Gloss abbreviations:

1, 2, 3	first/second/third person;	NEG	negation;
ANT	anterior tense;	OBJ	object/prepositional
DEF	definite determiner;		complement pronoun;
DEM	demonstrative;	PF	perfective aspect;
DET	determiner;	PL	plural;
FOC	focus pronoun;	POSS	possessive;
IMPF	imperfective aspect;	PREP	preposition;
INF	infinitival particle;	PRO	pronoun;
INDF	indefinite;	REL	relativiser;
INT	interjection;	S	singular;
IRR	irrealis;	SBJ	subject pronoun;
LOC	locative;	TOP	topic pronoun.

4.3 Corpus-based approach: Case studies

The case studies in this section take a corpus-based approach. While the first three case studies rely on tagging as added (theory-bound) information (Sections 4.3.1 to 4.3.3), the final two case studies rely on a combination of lexico-grammatical and sociolinguistic metadata (Sections 4.3.4 and 4.3.5).

4.3.1 *Lexical vs. functional uses of verbs*

As explained above, certain CPE verbs can function as lexical verbs as well as participating in serial verb constructions. Without POS tagging, an investigation of SVCs in CPE would require the corpus to be searched for each word separately, and further laborious work to separate and code the instances in which each verb functions as a serial verb, and those in which it functions as a lexical verb. In the CPE corpus, however, this process is simplified. The tagged corpus makes a distinction between the different uses of these verbs, allowing automatic retrieval with a simple search. Lexical uses of verbs are identified with the VB0 tag, whereas their serialised counterparts carry the tag VBS. A simple search allows us to use the tagging to gain an insight into the distribution of serial and lexical verbs in the CPE corpus; the results are presented in Table 6.

Table 6. Lexical vs. serial uses of multifunctional verbs

VB0 (lexical)	39,204	90%
VBS (serial)	4,330	10%
TOTAL	43,534	

To refine the enquiry further, combining lexical and tagset-based searches allows the researcher to investigate how this multifunctionality patterns in the case of individual verbs. Table 7 shows the five verbs most frequently tagged as serial verbs, *goe* ‘go’, *meik* ‘make’, *kam* ‘come’ and *stat* ‘start’ as a percentage of the total number of instances of each verb. As this table shows, certain CPE verbs (those with figures in bold in Table 7 below) are more frequently used as serial verbs than as lexical verbs.

Table 7. Multifunctionality in verbs

	<i>goe</i>	<i>kam</i>	<i>meik</i>	<i>stat</i>
VB0 (lexical)	52%	45%	43%	28%
VBS (serial)	48%	55%	57%	72%

4.3.2 *The case of meik: Causative and deontic*

In CPE, the serial verb *meik* ‘make’ is used to express causativity, as illustrated by Example (5).

- (5) Serial *meik*: causative
 i goe **meik** tings dem wos <PP-Ba-DPu-01>
 3s.SBJ IRR make things PL be.worse
 ‘He’s going to make things worse.’

Next, consider Example (6).

- (6) Serial *meik*: deontic modality
meik yu noe kam set konfushon <PP-Ba-DPu-01>
 make 2s.SBJ NEG come set confusion
 ‘Don’t come and cause confusion.’

While Example (5) illustrates straightforward causativity, Example (6) illustrates a grammaticalised use of this construction, which by extension expresses deontic modality (see also Green & Ozón forthcoming 2018b). Inspecting the corpus results for *meik* SVCs (by means of the VBS tag) allows the identification of this meaning of *meik* in SVCs. The observed high frequency of this meaning hints at grammaticalisation of the *meik* causative.

In the context of pidgin/creole linguistics, an important caveat worth making here rests on the distinction between genuine grammaticalisation (genuine internal language change over time) and ‘apparent’ grammaticalisation in the sense of Bruyn (2009). Bruyn argues that grammatical patterns that look like grammaticalisation in a creole language are often cases of substrate influence without grammaticalisation, a process that Heine and Kuteva (2005) call ‘polysemy copying’. This process involves copying the beginning and end stages of the grammaticalisation chain in the model (substrate) language(s) and mapping them onto corresponding categories in the pidgin/creole language. This does not involve grammaticalisation as a process within the contact language, but rather a type of calquing or reanalysis that maps a superstrate form onto a substrate function (Bruyn 2009: 323). See FitzGerald (2017) for further discussion.

4.3.3 *Refining existing descriptions*

Testing, improving, refining and correcting grammatical descriptions is an essential task of linguistics, which is even more important when these grammatical descriptions and/or generalisations are based on reduced datasets, as is often the case with pidgin/creole languages. In what follows, we test existing descriptions of CPE as listed in the *Atlas of Pidgin and Creole Language Structures Online* (APiCS; Michaelis et al. 2013). We select three linguistic features and investigate the claims contained in the APiCS description.

4.3.3.1 Nominal plural meanings <<http://apics-online.info/parameters/23#1/30/10>>

From a cross-linguistic perspective, plurals can be expressed using overt marking (prefix, suffix, stem change, tone or stress change, reduplication or a plural word preceding or following the NP) or may be unmarked (Dryer 2013). Schröder (2013) states that the favoured strategy in CPE (i.e. in 70% of cases) to express plural is through the use of plural word *dem* following the noun. Schröder states her confidence level on this observation as ‘very certain’ (Schröder 2013, see <<http://apics-online.info/valuesets/18-23>> and cites Bellama et al. (1983) and Ayafor (2008) as support.

A quick search in the corpus can be used to verify Schröder’s claims. Searching for all instances of plural particle *dem* (tag NPL) as well as for all instances of acrolectal plurals, i.e. inflected forms like *tings* ‘things’ (tag NN2) in the CPE corpus does indeed confirm *dem* as the preferred strategy for marking plural NPs (3,445 NPL vs. 562 NN2, almost 86%).

4.3.3.2 Ditransitive constructions with ‘give’ <<http://apics-online.info/parameters/60#1/30/10>>

Cross-linguistically, there are three main types of ‘give’ constructions (Haspelmath 2013). The first is the indirect object construction, in which the theme argument (direct object) is coded like the monotransitive object, while the recipient argument is coded differently as an argument, e.g. by a distinct case or by the presence of an adposition. The second is the double object construction, in which both the theme and recipient arguments are coded in the same way as the monotransitive object. The third is the secondary object construction, in which the recipient argument is coded like the monotransitive object and the recipient argument is coded differently. Many languages allow more than one of these constructions, as do CPE and English. Schröder (2013) claims that, in GIVE ditransitives, CPE speakers favour the indirect-object construction (7) 70% of the time over the double-object construction (8), which occurs in the remaining 30%.

(7) Indirect object construction

i don gif moni foe yi pikin
3s.SBJ PF give money PREP 3s.DET.POSS child
‘She’s given (some) money to her child.’

(8) Double object construction

i don gif yi pikin moni
3s.SBJ PF give 3s.DET.POSS child money
‘She’s given her child (some) money.’

Searches in our CPE corpus, however, do not confirm Schröder’s observations. In GIVE ditransitives, the double object construction is in fact the preferred pattern (74%), with the indirect object construction being selected considerably less

frequently (26%). This result confirms similar findings. Green & Ozón's (forthcoming 2018b) looked at all the tokens of 'give' in a different CPE dataset (Ayafor & Green 2014) and reported that the double object construction is favoured (77%) over the indirect object construction (23%).

These findings are consistent with Michaelis' (2014) claim that it is the double object construction that is favoured by African pidgin and creole languages, whereas the indirect object construction is favoured by their European lexifiers.

4.3.3.3 *Instrumental expressions* <<http://apics-online.info/parameters/69#3/30.30/10.02>>

Cross-linguistically, languages may code instrumental arguments in a range of ways including case, adposition, or serial/non-serial verb (Maurer et al. 2013a). Schröder (2013) states that instrumental arguments may be expressed in CPE by means of two strategies: the serial verb *teik* (which in her reckoning would account for 70% of instrumental expressions), and the use of adposition *wit* (which accounts for the remaining 30%). Her confidence level is 'certain' (Schröder 2013, see <<http://apics-online.info/valuesets/18-69>>, and this is supported by the intuition of her CPE informant.

The CPE corpus shows instances of *teik* tagged as a serialising verb (VBS, that is, serial verb or light verb). In order to discard any false positives (e.g. cases where *teik* might have been mistagged), we conducted searches for *teik* followed by noun phrase/pronoun followed in turn by a second verb and inspected the results manually. There were 33 cases of *teik* participating in an instrumental SVC such as (9).

- (9) Instrumental with serial verb *teik*
 de pikin **teik** de machis goe bon faya deiy
 DEF child take DEF matches go burn fire there
 'The child went and set fire (to the place) with matches.' <PP-Ba-DPr-01-B>

A separate search for *wit* was conducted in the CPE corpus; the results were manually inspected to verify that they introduced an instrumental expression. There were 87 such instrumentals in our corpus, illustrated by Example (10).

- (10) Instrumental with preposition *wit*
 noe meik=am kam shut yu wit gon o!
 NEG make=3s.OBJ come shoot 2s.OBJ PREP gun INT
 'Don't make him come and shoot you with a gun!' <Ba-DPr-01:263>

The results thus confirm the proportions identified by Schröder (2013).

4.3.3.4 'Give' serial verb constructions <<http://apics-online.info/parameters/86#3/30.15/9.84>>

In serialising languages, 'give' verbs can be used in SVCs to increase valency by introducing a recipient or beneficiary argument (Maurer 2013b). Schröder states that

no ‘give’ SVCs exist in CPE. Instead, the beneficiary/recipient argument either occurs immediately following the verb, or in an adverbial argument introduced by the preposition ‘for’. She states her confidence level as ‘very certain’ (Schröder 2013, see <<http://apics-online.info/valuesets/18-86>>, her claim is supported by her informant, a native speaker of CPE.

A quick search in the corpus for *gif* tagged as a serial verb (VBS) yields instances which might have been misidentified by the automatic tagger. We can however counter any false positives by conducting a search for any verbs followed by any noun phrase/pronoun followed by *gif* and inspecting the results manually. Even discarding erroneous tagging, these searches confirm that *gif* SVCs are attested in CPE: there are 39 instances of this construction in the corpus, illustrated by Example (11).

- (11) Benefactive with serial verb *gif*
 moni wei dey bi kari gif mi foe orphani
 money REL 3PL.SBJ ANT carry give 1s.OBJ PREP orphanage
 ‘money that they brought me at the orphanage’ <Ku_MU_01:25>

This confirms previous work (Ozón et al. 2017), which also found evidence (using a different dataset of CPE) for the *gif* benefactive serial verb construction in CPE.

4.3.4 Combining grammatical and speaker information: SVCs

Lexical or grammatical searches can be further broken down by region, allowing the identification of regional variation.

Table 8. Distribution of verb function by region

Region	VBS (serial)	VB0 (lexical)	% VBS per region
Bamenda	1,081	8,147	13
Bertoua	838	8,019	10
Douala	772	7,714	10
Kumba	940	7,814	12
Yaoundé	699	7,510	9
TOTAL	4,330	39,204	11

These figures suggest that SVCs are not that frequent. Lexical uses of verbs are, perhaps unsurprisingly, much more numerous. The figures in Table 8 also show that the relative proportions of lexical and serial verbs are quite consistent: in all the five investigated regions, serial verbs (VBS) occur once every ten clauses containing a lexical verb (VB0). This finding is consistent with the status of SVCs as constructions that are more grammatical than lexical in nature (in many cases, offering the only means of expressing certain meanings), and therefore less likely to be susceptible to regional variation.

4.3.5 Codeswitching

The language of the texts is CPE, with codeswitching into English, French, and indigenous Cameroonian languages. As mentioned above (Section 3.5), instances of codeswitching can be identified by searching for the mark-up identifiers <foreign> and <indigenous>, as illustrated by Examples (12) to (14).

- (12)

ya oun difen <foreign>definition</foreign> foe dem
2S.DET.POSS own different definition PREP 3PL.OBJ
'your own different definition for them' <Ba_DPr_02:7>
- (13)

soe yu stil deiy foe dat <foreign>école</foreign>?
so 2S.SBJ still be.LOC PREP DET.DEM school
'So, are you still at that school?' <Ku_DPr_05:14>
- (14)

pleis wei dem di wok <indig>njamanjama</indig>
place REL 3PL.TOP IMPF work vegetables
'A place where they grow vegetables' <Ba_DPr_01:52>

By combining mark-up information with metadata pertaining to social variables (e.g. age, gender, education), it is possible to approximate the social meaning of codeswitching, and to answer questions about who codeswitches, where is code-switching more frequent, which languages are codeswitched into, and others. For example, Figure 1 shows a breakdown of codeswitching by region. The results are not surprising, and point towards higher occurrences of codeswitching in larger, metropolitan areas, where a greater number of different languages come into contact. Results correlate by and large with the size of those cities' populations.

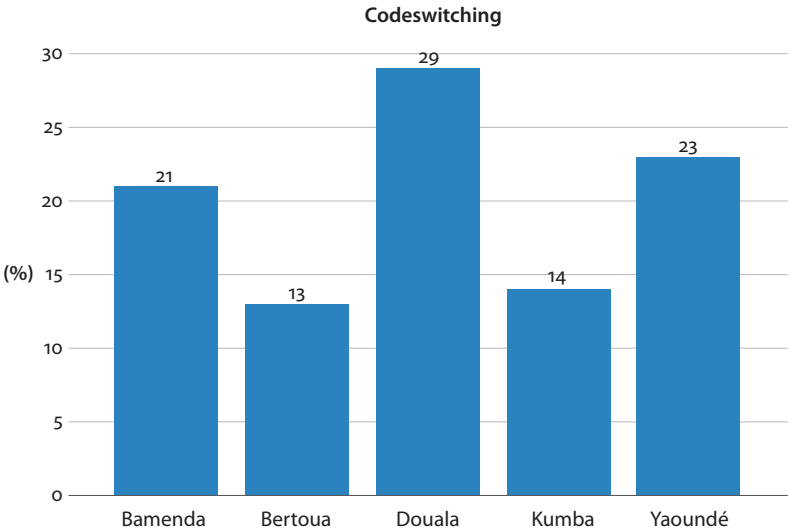
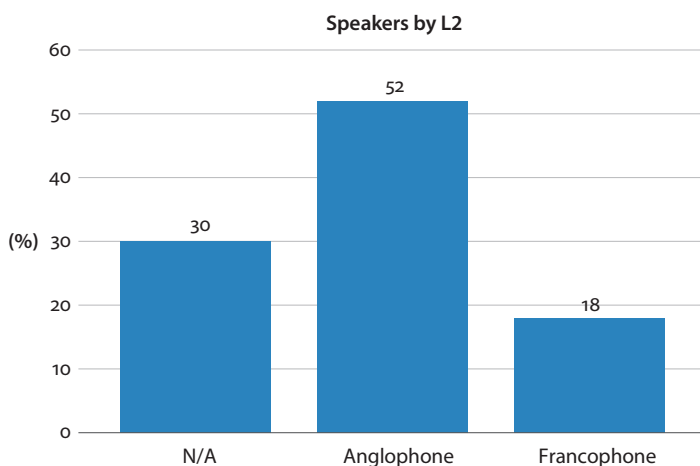
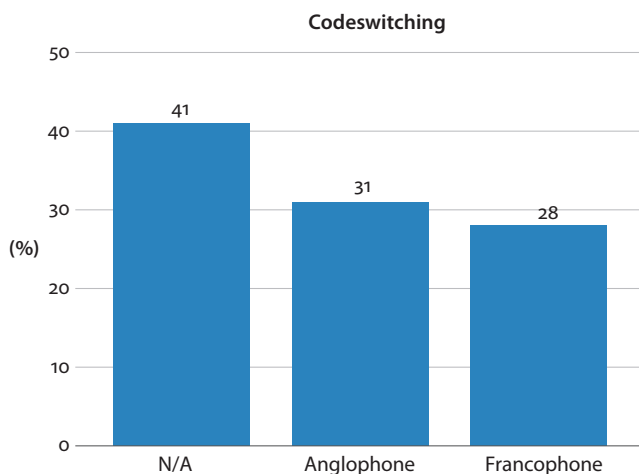


Figure 1. Codeswitching by region

Figure 2 shows the breakdown of codeswitching between anglophone and francophone speakers. The results show that although francophones are fewer in the corpus, they codeswitch more than anglophones. According to those who claim CPE is an anglophones' language, francophones might be expected to codeswitch more frequently because they have less familiarity with the language. To investigate such



a.



b.

Figure 2. Anglophones, francophones, and codeswitching²

2. The category 'N/A' in the figures refers to speakers for whom we have no participant data, e.g. participants in radio broadcasts. While it is possible in most cases to identify gender from the audio files, we preferred to take a consistent approach to cases where participant data is absent.

claims thoroughly, researchers could look at proportions of CPE native speakers in the anglophones' and francophones' participant data. Another possible explanation points to a methodological challenge: since CPE is predominantly English-lexified, it is a much more subjective judgement for the transcriber to identify cases of code-switching into English than cases of codeswitching into French.

Figure 3 shows a breakdown of codeswitching by gender. These findings suggest that males appear to codeswitch proportionally more than females. One avenue for investigation could be the hypothesis that this difference reflects a greater proportion of males working in the domains of science and technology where lexical gaps in CPE are more likely to occur than in e.g. domestic contexts. That hypothesis could be investigated by looking at the speakers' occupations in the participant information.

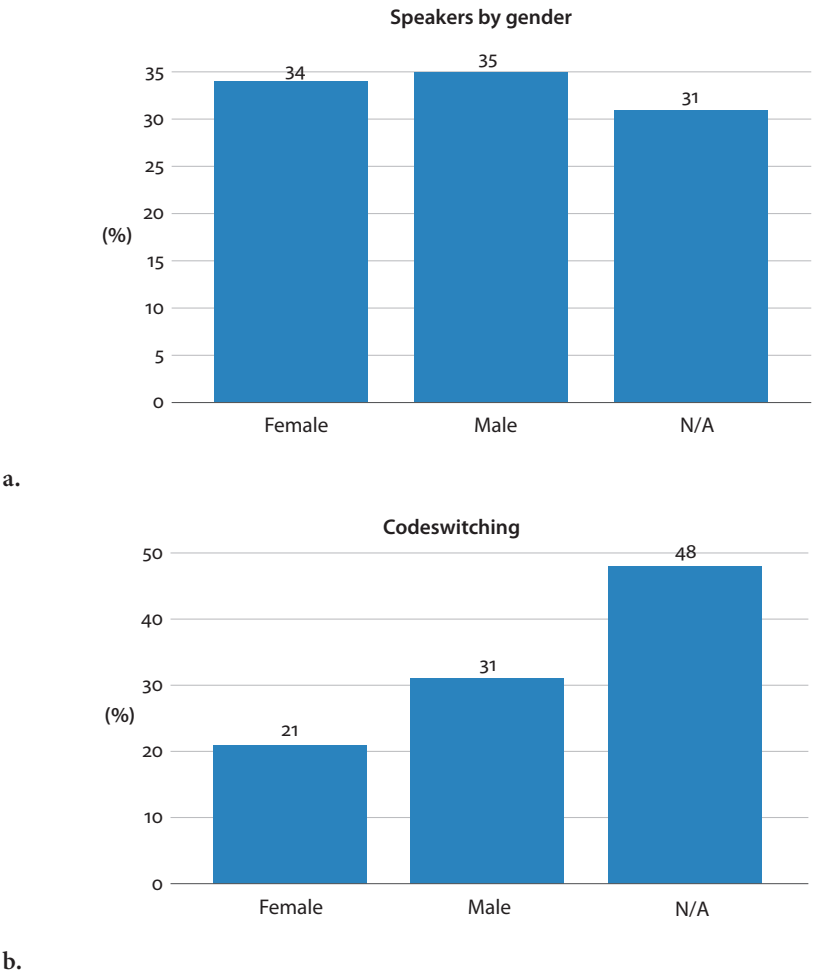
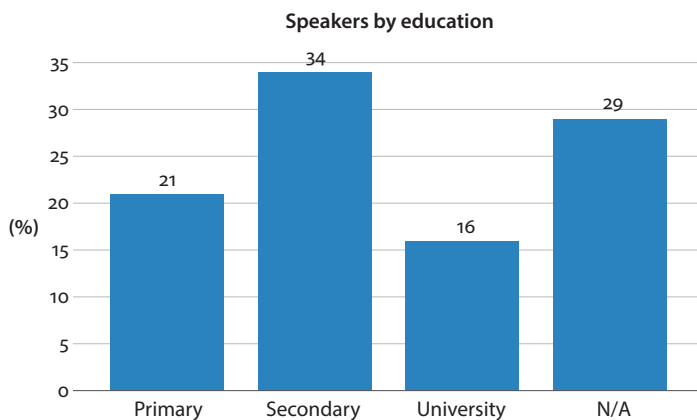
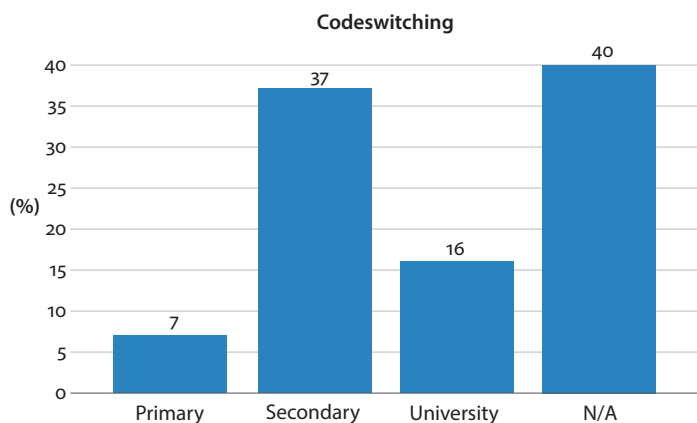


Figure 3. Gender and codeswitching

Finally, Figure 4 shows a breakdown of proportions of codeswitching according to level of education. Proportions are as expected for speakers with secondary education and university education: both categories of speakers codeswitch equally frequently. This can be explained by the fact that both groups have extended exposure to European languages (note that the majority of codeswitches are into English or French). In contrast, speakers educated to primary level of education codeswitch considerably less frequently.



a.



b.

Figure 4. Education and codeswitching

5. Conclusion

The CPE corpus represents the first attempt to provide a systematic database of a largely uncoded variety emerging from a highly complex contact situation. The corpus is aimed at providing a resource for linguistic description and comparison. It allows linguists to identify and describe recurring grammatical patterns, as well as the phonology of the language (given the availability of sound files deposited with the text files). It also allows comparison of CPE with other pidgin/creole languages, other Cameroonian and West African languages, and other varieties of post-colonial English. Furthermore, the corpus provides an exceptional resource for the study of general/theoretical linguistics, creolistics, typology, language contact and change, sociolinguistics, and discourse analysis.

Researchers and teachers can exploit the corpus and design their own investigations into this variety. The case studies provided here offer a snapshot of the potential uses of the CPE corpus. These case studies are illustrative of the richness of the material and the information (lexical, grammatical and sociolinguistic) available from the CPE corpus. Many of the results made available in the case studies require a social explanation, and we believe its discussion in an African context (among those with first-hand knowledge of the language as well as the social context) would be extremely profitable.

The CPE corpus significantly expands the database illustrating the spectrum of specific and general variation found in pidgin/creole varieties and African Englishes, which has been identified as an exciting new research avenue (Peter & Wolf 2007; Hackert 2010). Funding is being sought for the compilation of a larger 1,000,000-word corpus of spoken CPE, which would allow more robust linguistic generalisations to emerge. More substantial funding would also allow us to address the limitations identified by the pilot project, with a view to increasing representativeness, comparability and balance of the corpus.

References

- Anchimbe, Eric A. 2013. *Language Policy and Identity Construction: The Dynamics of Cameroon's Multilingualism* [Impact: Studies in Language and Society 32]. Amsterdam: John Benjamins. <https://doi.org/10.1075/impact.32>
- Anthony, Laurence. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo: Waseda University. <<http://www.laurenceanthony.net/>>
- Ayafor, Miriam. 2008. Cameroon Pidgin English (Kamtok): Morphology and syntax. In *Varieties of English, Vol. 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 428–450. Berlin: Mouton de Gruyter.
- Ayafor, Miriam. 2014. Cameroon Pidgin English orthography. Paper presented at the 14th International Colloquium of Creole Studies, Aix-en-Provence, France.

- Ayafor, Miriam & Green, Melanie. 2014. *A Spoken Corpus of Cameroon Pidgin English: Pre-Pilot Study*. Ms, University of Sussex.
- Ayafor, Miriam & Green, Melanie. 2017. *Cameroon Pidgin English* [London Oriental and African Language Library 20]. Amsterdam: John Benjamins. <https://doi.org/10.1075/loall.20>
- Bellama, David, Nkwele, Solomon & Yudom, Joseph. 1983. *An Introduction to Cameroon Pidgin*. Yaoundé: Peace Corps Cameroon.
- Brato, Thorsten. 2018. 'Outdoorings' the Historical Corpus of English in Ghana. *English Today* 134: 25–34. <https://doi.org/10.1017/S0266078417000517>
- Bruyn, Adrienne. 2009. Grammaticalization in Creoles: Ordinary and not-so-ordinary cases. *Studies in Language* 33: 312–337. <https://doi.org/10.1075/sl.33.2.04bru>
- Devlin, Thomas, FitzGerald, Sarah & Green, Melanie. In preparation. *Investigating tone in a spoken corpus of Cameroon Pidgin English*. Ms, University of Sussex.
- Dryer, Matthew S. 2013. Coding of nominal plurality. In *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmath (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://wals.info/chapter/33>> (5 October 2017).
- de Féral, Carole. 1989. *Pidgin-English du Cameroun. Description Linguistique et Sociolinguistique*. Paris: Peeters/Selaf.
- FitzGerald, Sarah. 2017. Investigating Grammaticalisation Paths of Multifunctional *meik* in Cameroon Pidgin English. MA dissertation, SOAS, University of London.
- Green, Melanie, Ayafor, Miriam & Ozón, Gabriel. 2016. *A Spoken Corpus of Cameroon Pidgin English: Pilot Study*. British Academy/Leverhulme funded project. <<http://ota.ox.ac.uk/desc/2563>>
- Green, Melanie & Ozón, Gabriel. 2018a. Information structure in a spoken corpus of Cameroon Pidgin English. In *Information Structure in Lesser-described Languages: Studies in Prosody and Syntax* [Studies in Language Companion Series 199], Evangelia Adamou, Katharina Haude & Martine Vanhove (eds), 329–355. Amsterdam: John Benjamins. <https://doi.org/10.1075/slcs.199.12gre>
- Green, Melanie & Ozón, Gabriel. 2018b. Valency and Transitivity in Contact: Evidence from Cameroon Pidgin English. *Journal of Language Contact*.
- Hackert, Stephanie. 2010. ICE Bahamas: Why and how? *ICAME Journal* 34: 41–53.
- Haspelmath, Martin. 2013. Ditransitive constructions: The verb 'give'. In *The World Atlas of Language Structures Online*, Matthew S. Dryer & Martin Haspelmath (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://wals.info/chapter/105>> (5 October 17).
- Heine, Bernd & Kuteva, Tania. 2005. *Language Contact and Grammatical Change*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511614132>
- Johansson, Victoria. 2008. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics* 53: 61–79.
- Koenig, Edna L., Chia, Emmanuel & Povey, John (eds). 1983. *A Sociolinguistic Profile of Urban Centres in Cameroon*. Los Angeles CA: Crossroads Press.
- Maurer, Philippe & the APiCS Consortium. 2013a. Instrumental expressions. In *Atlas of Pidgin and Creole Language Structures Online*, Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://apics-online.info/parameters/69>> (5 October 2017).
- Maurer, Philippe & the APiCS Consortium. 2013b. 'Give' serial verb constructions. In *Atlas of Pidgin and Creole Language Structures Online*, Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://apics-online.info/parameters/86>> (5 October 2017).

- McEnery, Tony & Gabrielatos, Costas. 2006. English corpus linguistics. In *The Handbook of English Linguistics*, Bas Aarts & April McMahon (eds). Malden MA: Blackwell.
- Menang, Thaddeus. 2004. Cameroon Pidgin English (Kamtok): Phonology. In *A Handbook of Varieties of English*, Vol.1, Edgar W. Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 902–917. Berlin: Mouton de Gruyter.
- Michaelis, Susanne Maria, Maurer, Philippe, Haspelmath, Martin & Huber, Magnus (eds). 2013. *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://apics-online.info/parameters/86>> (5 October 2017).
- Michaelis, Susanne. 2014. Loan valency patterns in Creoles: Evidence from the Atlas of Pidgin and Creole Language Structures (APiCS). Paper presented at the 47th Annual Meeting of the Societas Linguistica Europaea, Adam Mickiewicz University, Poznań, Poland.
- Nelson, Gerald. 1996. The design of the corpus. In *Comparing English Worldwide. The International Corpus of English*, Sidney Greenbaum (ed.), 27–35. Oxford: Clarendon Press.
- Nelson, Gerald. 2002. *Markup Manual for Spoken Texts*. <<http://ice-corpora.net/ice/spoken.doc>>
- Nelson, Gerald & Ozón, Gabriel. 2017. World Englishes and corpus linguistics. In *World Englishes: Re-thinking Paradigms*, Ee Ling Low & Anne Pakir (eds), 149–164. London: Routledge.
- Nkemleke, Daniel & Mbangwana, Paul. 2007. *Manual of Information to Accompany the Corpus of Cameroonian English (CCE)*. Chemnitz: Chemnitz University of Technology, Department of English.
- Ozón, Gabriel, Green, Melanie, Ayafor, Miriam & FitzGerald, Sarah. 2017. Building a spoken corpus of Cameroon Pidgin English: Methodological challenges. *World Englishes* 36(3): 427–447. <https://doi.org/10.1111/weng.12280>
- Peter, Lothar & Wolf, Hans-Georg. 2007. A comparison of the varieties of West African Pidgin English. *World Englishes* 6(1): 3–21. <https://doi.org/10.1111/j.1467-971X.2007.00485.x>
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. <<ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf>>
- Schröder, Anne. 2003. *Status, Functions and Prospects of Pidgin English: An Empirical Approach to Language Dynamics in Cameroon*. Tübingen: Gunter Narr.
- Schröder, Anne. 2013. Cameroon pidgin English structure dataset. In *Atlas of Pidgin and Creole Language Structures Online*, Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds). Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://apics-online.info/contributions/18>> (5 October 17).
- Simo Bobda, Augustin & Wolf, Hans-Georg. 2003. Pidgin English in Cameroon in the New Millennium. In *Studies in African Varieties of English*, Peter Lucko, Peter Lothar & Hans-Georg Wolf (eds), 101–117. Frankfurt: Peter Lang.
- Simons, Gary F. & Fennig, Charles D. (eds). 2017. *Ethnologue: Language of the World*, 20th edn. Dallas TX: SIL International. <<http://www.ethnologue.com>>
- Taylor, Stephanie, Wetherell, Margaret, & Yates, Simeon J. 2001. *Discourse as Data: A Guide for Analysis*. London: Sage/Open University Press.
- Tiomajou, David. 1993. Designing the Corpus of Cameroon English. *ICAME Journal* 17: 119–124.
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.6>
- Todd, Loreto. 1982. *Cameroon*. Heidelberg: Julius Groos. <https://doi.org/10.1075/veaw.t1>
- Wolf, Hans-Georg. 2001. *English in Cameroon*. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110849059>

Practical corpus linguistics

Designing and exploiting a written corpus for research with special reference to Cameroon English

Daniel Nkemleke

University of Yaoundé I

One major contribution of the corpus-based approach to language study is the facility with which linguists can access large amounts of data and search for regularities across text types and varieties of English on the basis of observed frequencies. Such regularities of features, if based on well-designed and/or specialised corpora, can be useful for classroom activities and material design. This chapter describes the basic steps involved in corpus design and corpus exploitation. Section 1 defines the corpus and briefly states the relevance of investigating English second language (ESL) varieties via a corpus. Section 2 describes key steps involved in simple corpus compilation and how students can compile their own corpora for research. Section 3 presents results of two studies on the frequency and use of modals in the corpus of Cameroonian English. These results are intended to illustrate the point that the corpus approach is indispensable if certain types of linguistic information are investigated in a particular variety.

1. Introduction

1.1 Aim

Corpus linguistics has gained notoriety in the field of language study in the last 50 years since Randolph Quirk's launching of the Survey of English Usage Project (Quirk 1961). The subsequent impact generated by first generation corpora such as the LOB and Brown corpora, the ICE (International Corpus of English)¹ project

1. The ICE project had been launched in 1988 (Greenbaum (ed.) 1988), and views as its mission, provision of parallel corpora that sample the English used in the participating countries: Australia, Cameroon, Canada, East Africa (Kenya, Tanzania), Fiji, Ghana, Great Britain, Ireland, Hong Kong, India, New Zealand, Nigeria, Philippines, Singapore, South Africa and the United States of America.

(Greenbaum 1988, 1994), and the relatively recent mega corpora (e.g. Cobuild Bank of English, British National Corpus) have lent credence to this approach to linguistic description. Although there are a number of past and on-going corpus linguistic projects in Africa (e.g. Schmied 1989 for East Africa; de Klerk 2002 for South Africa; Nkemleke 2008 for Cameroon; Akere 1996; cf. Esimaje 2016, 2017 for Nigeria),² corpus linguistics is still relatively unpopular in many linguistic departments in Cameroon, and probably elsewhere in English second language (ESL) communities, and many tertiary teachers here have apprehensions doing corpus linguistics. This timidity seems to have inoculated students in most universities in Cameroon against corpus linguistic work.³ In view of the amount of linguistic information that a computer corpus contains (cf. Biber et al. 1999); and the value of such information for empirical research and language learning and teaching, corpus linguistics is an important methodology for modern linguistic studies – one which students ought to be encouraged to embrace. Although many online open access materials are available on the subject, the need for textbooks on corpus linguistics such as this present volume; with a diversity of research orientations (see Part II), and addressing what undergraduate students need to know about the subject (see Part I) cannot be overemphasized. It is in this connection that this chapter, intends to make a contribution. Section 1 defines the corpus and briefly states the relevance of investigating ESL English varieties via corpora. Section 2 describes key steps involved in simple corpus compilation and how teachers/students can compile their own corpora for research. In the third section some results on the frequency and use of modals in the corpus of Cameroonian English (CCE) are presented. These results are intended to illustrate the point that the corpus approach is indispensable if certain types of linguistic information have to be investigated in a particular variety.

1.2 Why is corpus linguistics relevant for ESL varieties?

The contribution of corpus linguistics to the study of English in general has been widely illustrated, not least to the study of English in Africa (cf. Schmied 1990) as can be seen in its direct application to dictionary making, especially in English Native Language (ENL) varieties. For ESL varieties, corpus linguistics is even more

2. South Africa is the exception here because there have been many corpus projects going on there. I am also aware that some work has been done in Ghana by linguists from Germany.

3. In my own working experience in the university since the compilation of the first Cameroonian corpus (see Tiomajou 1993; Nkemleke 2008), very few colleagues and students have expressed interest in corpus-based linguistics.

relevant because of the potential it has to introduce a systematic approach to the study of these varieties, since a computer corpus contains a large amount of linguistic information which can be processed with speed, bringing out detailed and subtle aspects of usage within seconds. Such systematic investigation of patterns across varieties and even text types, if based on comparable corpora (cf. ICE initiative), can significantly contribute to the search for regularity across varieties of African Englishes, in general (e.g. Gut & Unuabonah this volume), and specificities within national varieties such as Cameroon, Nigeria (e.g. Iyabo this volume) etc. Furthermore, the value of corpus frequencies in language education cannot be overemphasised. Frequency measures of vocabulary and even such simple words as conjunctives (see Iyabo this volume) may be very important for vocabulary grading and selection, with application in such areas as syllabus design, material writing etc.

2. Compilation of the CCE

2.1 Initial considerations

The decision to build the CCE was motivated by the need to have a representative database on which to undertake research on Cameroon English (CamE). The main objective of the project was to set up a million-word corpus of Cameroonian mainstream educated English, entirely based on written usage. It was intended to be useful in a range of ways, including the following: (a) provide a textual basis for a quantitative study of CamE; (b) provide a database for the description of the main features in the variety of English written in Cameroon; (c) provide a source of authentic material for English language teaching in Cameroon, and (d) serve as a database for comparative studies on CamE in relation to other varieties of English in Africa and beyond. Above all, the case for a database of CamE was even more compelling especially in the context of a paucity of electronic databases in Africa⁴ at the time.

2.2 Design parameters

The decision to include texts in the CCE was guided by what researchers at the time considered “representative” of CamE. Accordingly, the following design parameters were adopted for the database: (a) the corpus should contain a broad range of the local written usage of CamE; (b) the corpus should include fiction and non-fiction,

4. Apart from South Africa, the only known project in Africa at the time was ICE-EA (Schmied 1989).

as well as popular, scholarly and literary texts; (c) the texts should be those written by Anglophones (part of Cameroon uses French) who have had at least a secondary school certificate, and (d) the texts should be those written between 1990 and 1994. At the beginning of the project, it was felt that the corpus should be as original as possible, not copying any existing corpus models, but reflecting the particular socio-linguistic and educational realities of Cameroon. However, the research team gradually came to the conclusion that adopting a more conventional format would be suitable for future comparative research. The team decided on a corpus that would contain text categories broadly inspired by the text categories in the Brown and LOB corpora, but not identical to these (see Table 1).

Table 1. Number of texts and words per text category in the CCE

Text categories		Number of texts	Number of tokens
A	Official Press (OP)	183	250,000
B	Students' Essay (SE)	116	130,760
C	Miscellaneous (MI)	32	122,569
D	Novels and Short Stories (NS)	21	104,458
E	Private Press (PP)	50	81,860
F	Government Memoranda (GM)	14	74,692
G	Private Letters (PL)	222	69,347
H	Tourism (TR)	8	30,188
I	Religion (RE)	14	15,839
J	Official Letters (OL)	44	10,632
K	Advertisement (AD)	10	4,547
Total		714	1, 000,451

A comment on the nature of texts in the corpus is in order here. Renouf (1987) states, “it is in the nature of texts that a primary function is accompanied by a secondary one”: a narrative text will for example, contain description. In the non-fiction texts (i.e. all categories except D), surveys and procedures are the commonest types. Here, the writer, through a process of description and comment, seeks to inform the reader about an aspect of life. Examples are texts on the environment, and the National Social Insurance Fund in the MI component. In the procedural texts, the authors give advice on how to proceed in a given activity: they do not exhort the reader to do a particular thing, but only provide relevant facts, and suggest what is intended to be an appropriate course of action should the desire be there. Again, most MI texts are in this category. The majority of texts in PP and RE are also procedural. For example, in the latter we have titles such as: ‘How to Grow in Christian Life’, ‘How to Gain Eternal Life’, and ‘The Way of Christian Discipline’. Narrative texts are also many in the corpus. In these texts the writers present a sequence of events with the communicative intention of entertaining or edifying, gaining

sympathy, exorcising guilt, and so on. The sequence is traditionally chronological, but it varies for stylistic effect. There are invariably elements of description in the text, an underlying position held by the author towards the subject, which may or may not be made explicit. Most texts of the NS fall in this domain. There are also many of such texts in the PP, OP and PL categories. Closely related to narrative texts are biographies. In these texts an account is given of the life of an individual whom the author thinks is sufficiently interesting to draw the reader's attention. For example, 'Bernard Fonlon: The torch and the throne' in MI and many other biographies of persons appointed to positions of responsibility found in OP and PP components of the corpus. In most cases the account will not be exhaustive, but will focus on salient areas of the subject's life.

Another major discourse function of texts in the CCE is argumentation. The author puts forward his case in relation to a particular issue. He would provide relevant information in terms of a descriptive and/or narrative account, in order to support his position. Most often the reader tends to be addressed more directly, and is exhorted to do or become something which the author argues is desirable. The following example explains this: "all should heed opposition call for nation-wide non-stop strike"; "Vote for CPDM or Chaos" and other texts of a persuasive nature in OP and PP components. The texts in AD naturally fall into this category of description. In all these texts some background information about a particular issue is given, and various points of view held in relation to it are presented. The author may, or may not, hold one of them, but the presentation of arguments is balanced. Texts in the SE component also fall in this category because most of the examination questions whose answers constitute texts in SE demand balancing arguments, giving justification, and substantiating. Finally, there is a small group of texts providing information, making requests and giving orders. These include circular letters in the OL component, especially those from government ministries to subordinates. Lastly, there are texts of informal communication between friends, lovers, and family members in the PL category.

One type of structural markup that is necessary for a corpus is the inclusion of biographical information on all text files. This is descriptive information about the texts often inserted in a file folder, and which gives complete information on each text of the corpus such as date, writer, origin, page range etc. The text of the CCE have the following biographical information in a file header, placed in a separate folder: origin of texts (O), author of the text (A), gender of the writer (G), pages from which the text is taken from (P),⁵ date on which the text was written (D), title of the text (T), word file number (W).

5. Sections of texts were selected in most cases during the original compilation. For example, a novel or a newspaper was often not copied entirely.

3. Some findings on the use of modals in CamE⁶

This section reports some findings on the use of modal verbs in CamE, namely the modals that express obligation and necessity (3.1), as well as the overall frequency and common senses of the modals as they occur in the entire database (3.2). The results illustrate the relevance of the corpus approach in tracking subtle aspects of meaning and usage of these verbs in CamE hitherto unavailable. But before these results are given, it is necessary to describe the procedure used to extract the data. Since my study was comparing figures for LOB corpus of British English (BrE) published in Coates (1983), it was necessary to align my figures to these; and so, a sample of 200 for each modal was selected.⁷ This sample was manually extracted after a concordance output for each modal had been produced using AntConc.⁸ The task consisted in selecting each 2nd to the 6th instance of the modal depending on the overall frequency, until a 200 mark was reached. For example, for TR with the occurrence of 12 modals, concordance line 1 and 7 were selected; for SE with the occurrence of 15 modals, concordance line 1, 7 and 15 were selected. The lone figure for AD was taken. I chose to ignore 'ought' and 'need' because of their very low frequencies.

3.1 'Must', 'should', 'ought' and 'need' in CamE

The following modal verbs are used to express obligation and necessity: 'must', 'should', 'ought', and 'need'. They have often been noted to display a significant amount of regional variation (as noted by Trudgill & Hannah 1982) and register variation (Coates 1983; Collins 1988) in Standard English. Such subtle differences across varieties of English are now easy to investigate because of the availability of sampled corpora across English varieties. This section explores the nature and extent of such variation in the CCE, comparing British English (BrE) figures as reported in Coates (1983).⁹ Since Coates selects a sample of 200 instances of each modal for her analysis, Table 2 presents a similar set of samples selected from the CCE (sample figures are in the brackets) for 'must' and 'should', only. The figures

6. These results are taken from my unpublished PhD thesis (Nkemleke 2003)

7. Coates (1983) uses a sample of 200–236 for each modal (Table 6).

8. <www.antlab.sci.waseda.ac.jp/software/antconc3.2.4w.exe> (30 April 2017).

9. The primary source of information on the modals in BrE used in Coates (1983) study is a corpus of over a million and a half words (1,000,000 from the LOB corpus of written BrE and 545,000 from the spoken and non-printed written sections of the Survey of English Usage corpus).

for ‘ought’ and ‘need’ are raw frequencies, since they are less frequently expressed in the corpus.¹⁰

Table 2. Distribution of ‘must’, ‘should’, ‘ought’ and ‘need’ in the CCE

Text/Forms	OP	SE	MI	RE	NS	PP	GM	PL	TR	OL	AD	Total
must	121 (26)	15 (3)	522 (104)	108 (21)	56 (11)	8 (2)	79 (19)	57 (11)	12 (2)	– (–)	1 (1)	979 (200)
should	152 (30)	29 (7)	27 (57)	209 (42)	36 (8)	41 (8)	78 (15)	131 (26)	10 (2)	23 (5)	– (–)	985 (200)
ought	13	4	15	15	2	–	0	1	–	1	–	54
need	1	1	2			1	1	3	–	–	–	9

As Table 2 shows, ‘must’ and ‘should’ have the highest frequencies especially in the MI, RE, GM and PL text categories. The only modal with the highest frequency in a single text category is ‘must’ (in MI). This stems from its over-generalized use in the CCE (Examples (1) to (3)). The high figure for ‘should’ can also be attributed to the same phenomenon where, the strong/subjective cases are prevalent (see Examples (4) to (6)).

- (1) ...first year in secondary school, you must do your assignments, revision exercise every evening (PL 565)
- (2) ...and employment opportunities. You must find this out yourself. Draw up a plan of study... (MI 623)
- (3) ..needs to carry out restitution. You must act at once or face the consequences (RE 14)
- (4) ... what Cameroonians want and I say we should all gives our support for a leader who will truly lead us out of this... (PL56)
- (5) ... qualities are essential. The disciple maker should also give the young disciple a scheme of work without which he...
- (6) Jangali tax, it should be an obligation for the citizens to pay... (OP 48)

These particular uses of ‘must’ and ‘should’ in Coates’ (1983) study are found mostly in the spoken component of her data. The figures for ‘ought’ and ‘need’ are relatively small. For the former, it might be the case as Svartvik and Wright (1977) claim that it is obsolescent. It is not very clear why ‘need’ has the lowest figure in the corpus. However, speculatively one can attribute this to the nature of exposure to the

10. The low frequency figures for ‘ought’ and ‘need’ relative to ‘must’ and ‘should’ raises further research questions. The BrE figures reported in Coates (1983: 23) are: ‘ought’: 105, ‘need’: 71.

language in Cameroonian schools where learners are more likely to come across 'need' as a noun first, and not as a verb.¹¹ One might also ask why the newspaper component of the CCE (i.e. OP and PP) which have up to a total of 331,860 words (cf. Table 1 above) should have some of the lowest frequency of the modals. A possible explanation is that journalists, committed as they are to factual, non-speculative mode of discourse, have less occasion to produce assertions modalized in terms of obligation (at least of a subjective type) and necessity than those engaged in the production of other text categories represented in the corpus (cf. Collins 1991 for a similar situation in Australian English). These conclusions are possible only because of the availability of quantitative data. The frequencies of 'must', 'should', 'ought' and 'need' are relatively low in the CCE. With very low frequencies for 'ought' and 'need', only the forms 'must' and 'should' will be discussed at great length ahead. But before this is done, I should like to distinguish between the notions of 'root' and 'epistemic' in relation to the meaning of obligation and necessity expressed by the modals in question.

Epistemic modality is concerned with the speaker's assumption or assessment of possibilities and in most cases, it indicates the speaker's confidence (or lack of confidence) in the truth of the proposition expressed. The modals relating to assumption are 'must', 'should', and 'ought'; those relating to an assessment of possibilities are 'may', 'might', 'could' and 'will'. To explain or illustrate epistemic modality, we shall consider the example of epistemic 'must' in: *Paul must be in Kumba by now*. This might have the force (in context) of 'I assume, taking into account what time he left home, the time now, and the state of public transport, that Paul is in Kumba'. This interpretation makes explicit the fact that epistemic 'must' involves the speaker in logical inference, even though the facts which lead to the inference are often not stated. Epistemic 'must' can be seen as one way of expressing the logical statement (if P) then Q, that is, the sentence above could be rewritten: 'If Paul left home at 8 o'clock, and if the buses are running, and if it is 9 o'clock, then by now Paul is in Kumba'. A logician's, as opposed to a linguist's view of epistemic modality would lead to a paraphrase such as: 'In the light of what is known, it is necessarily the case that Paul is in Kumba'. Note that the speaker is not mentioned; that is, logicians are primarily concerned with objective epistemic modality, which is clearly related to alethic modality (the modality of the necessary truth of propositions). While objective epistemic modality does occur in natural language (a good example from the CCE (PL) component is: the fact is that if you are going to send the letter by hand, it must be delayed), it is not very common, and this has been recognized by linguists whose definitions have emphasized the subjectivity of epistemic modality. That is, the example above would most commonly be interpreted as meaning that the writer

11. Proof is that of the 419 citations for lexeme 'need' in the CCE, 410 (97.9%) are nouns and only 9 (2.1%) are used as a modal verb (Nkemleke 2003).

was confident about what he/she was saying. In other words, a linguist’s definition might be ‘I’m sure/I confidently assume that Paul is now in Kumba’.

Although often more difficult to characterize than epistemic modality, ‘root’ modality refers to the logic of obligation and permission (cf. Von Wright 1951), and typical root modals such as ‘must’ and ‘may’ cover a range of meaning, of which ‘obligation’ and ‘permission’ represent only the core. In certain contexts, where the authority structure is well-defined, such as when a mother speaks to a child, or a teacher to a pupil, a ‘root’¹² interpretation is more appropriate, and in other contexts, such as formal prose, where the participants are less clearly defined, such an interpretation would be less appropriate. But there is equally a clear basic meaning which is common to both-something like ‘it is necessary’. Thus ‘You must get out of the bath now’ (mother to child) means ‘it is necessary for you to get out of the bath and I am going to see that you do’, while ‘clay pots...must have some protection from severe weather’ means ‘it is necessary for clay pots to have some protection...’ with the speaker/writer’s involvement being less prominent (both examples are taken from Coates 1983: 34–35). Gradience is therefore an essential feature of root modality. While the epistemic modals vary only in terms of subjectivity (with objective cases occurring rarely), the root modals vary both in terms of subjectivity and in terms of strong-weak continuum. Despite the range of meaning covered by the root modals, their essential unity is confirmed by the syntactic patterns associated with them, which distinguish them from their epistemic counterparts. For example, the features “animate subject”, “agentive verb” and “passive voice” are all linked with root meaning. Stress and intonation patterns also distinguish root and epistemic meaning.

‘must’

Table 3 presents the figures for ‘must’ in the databases. Root ‘must’ covers a range of meaning in the CCE. Although some linguists have found it comfortable to discuss these meanings in terms of a single (albeit fuzzy) category (see Coates 1983: 32–36; Collins 1991), I will do so in terms of two sub-categories: root and epistemic. The indeterminate category fits neither the root nor the epistemic category. The following examples illustrate the sub-categories of ‘must’ in the CCE and LOB corpus.

Table 3. Frequency of the root and epistemic meanings of ‘must’ in CamE and BrE

Must	CCE	BrE
Root	123 (61%)	153 (64.8%)
Epistemic	41 (20.5%)	74 (31%)
Indeterminate	36 (18%)	9 (3.8%)
Total	200 (100%)	236 (100%)

12. Palmer (1979) uses the term ‘deontic’ or ‘discourse-oriented’.

Root

Instruction:

- (7) ... anything to Massa Kurt. You *must* come and see me. I am your Lawful father.
(NS 14)

Exhortation and Intention:

- (8) ... the specific cause of event or situation, then one *must* accept that it is the nature of creation that it should ... (m118)
- (9) In the process of taking pictures of these 'gems'. I *must* add too that warts or no wart (PN Yaoundé) continues to... (m103)

Necessity:

- (10) The one who renounces all that he has *must* ask the Lord to supply his needs and not his luxuries. (RE 57)

Epistemic

Necessity:

- (11) I received your letter which you sent through Male 2.¹³ I know you *must* be very worried for what is going on between us. (PL 288)

Example (7) occurs 17 times in the CCE. These examples share one important characteristic, namely that of the writer's involvement in the statement. Palmer (1974) calls this discourse orientation and Lyons (1977) refers to it in terms of subjectivity. Here the meaning of 'must' comes close to that of an imperative, paraphrasable as 'I order you to X'. These cases are typical of the 'core' referred to as 'strong obligation' by some linguists (see for example Coates 1983: 32). They correspond to the native speaker's psychological stereotype of root 'must' (Wells 1979). In other words, they are essentially performative and Lyons states that such examples are rare in the language (cf. Wells 1979) and this is statistically confirmed in Coates (1983) where they constitute only 1/14 of all root examples in the LOB corpus. A few other examples occur in the spoken component of her data (i.e. the Survey). In all respects where it occurs, authority structure (e.g. the home: mother to child or school: teacher to pupil) is evident. According to Coates (1983), apart from a few rare contexts like the law courts, performative root 'must' is rare in BrE because people are either seen as equals or are treated as such, since to do otherwise

13. Proper names (male and female) used in private letters were replaced by 'Male 1', 'Male 2', 'Female 1' etc.

would be impolite and often counterproductive. Its generalized use in CamE is therefore idiosyncratic. It may well suggest a general absence of the politeness phenomenon in the writing style of most people in Cameroon.

Examples (8) and (9) are those of self-exhortation and intention respectively. These two categories account for up to 25 cases of root 'must' in the corpus. With Example (8), the writer urges himself or herself and another person(s) to do something. All such examples were either in the first person impersonal 'we' or the third person 'one'. Lyons (1977) refers to them as pseudo-exhortations. In BrE such examples are typical of lectures, sermons and other forms of oratory (Coates 1983). Example (9) is used holophrastically and it is odd in that the writer is seen to be doing what he/she is in the act of urging himself/herself to do (cf. Huddleston 1984). This specific use of 'must' in BrE occurs mostly in what Coates (1983) classifies as 'private spoken language'.

Up to 32% of the cases of 'must' are used in the CCE to express necessity; of this total, 24 (75%) express root meanings and 8 (25%) express epistemic meanings. In the former (i.e. Example (10)) what is implied is that something is a necessity in an unavoidable sense and must be the case – paraphrasable as 'it is imperative.'/ 'it is important' and in the latter (i.e. Example (11)) that something is logically the case. In this sense, the 'necessity' is logical, expressing the writer's confidence in the truth of what he/she is writing, based on a logical process of deduction from facts known to them (which may or may not be specified). An interesting development in the CCE is that of the strong association of epistemic 'must' with harmonic combinations (Example (12)) and (Example (13)) and linguistic hedges (Example (14)) and (Example (15)).

- (12) ...the clear reason, surely by now you must have or is about to defend your thesis (PL 289)
- (13) Something certainly must have happened, he suspected. As he opens the door to the next ... (NS 211)
- (14) To beat her up. I suppose she must have been rude in her response. (PL 212)

Lyons (1977: 807) introduces the term 'modally harmonic' to describe those combinations of modal auxiliary and another modal word where both modal forms express the same degree of modality. Two harmonic forms therefore are said to be mutually reinforcing (cf. Halliday 1970: 331). As many as 19 different harmonic combinations are found to occur with epistemic 'must' in CamE. In (12) and (13) the writer's confidence is overtly expressed by the harmonic words 'surely' and 'certainly' and the reason for this confidence is found not in the statement but in some shared previous knowledge and/or circumstances known to the writer. Most examples in the corpus with harmonic words or phrases are of this type. In (14)

and (15) however, the process of reasoning is less clear, and the writers' introduce their statements with hedges ('I suppose', 'I imagined'). Coates (1983: 41) states that hedges are frequently found as markers of epistemic 'must' particularly in speech.

Epistemic 'must' also exhibits a strong tendency to occur with some syntactic co-occurrence patterns. For example, it occurs 8 times with the progressive aspect, 17 times with the perfective aspect, 21 times with inanimate subjects and 17 times with stative verbs. It occurs rarely with existential subjects (only 1 example found). It has been claimed by some writers that epistemic 'must' lacks (or 'normally' lacks) a negative form (e.g. Leech 1971: 72; Coates 1983: 46). Although this claim has been challenged by some linguists (e.g. Collins 1991 on Australian English), it appears to be the case with the Cameroonian corpus. Epistemic 'must not' does not occur in my sample.

'should'

There are four different meanings attributed to 'should' in the CCE (Table 4). It has a root meaning, an epistemic meaning, it sometimes functions as a quasi-subjunctive, and it also supplies a first-person variant for hypothetical 'would'. Out of a 200 sample for 'should', 8 instances exhibited hypothetical meaning. These are not considered in this discussion. Frequencies for the relevant meanings are presented in Table 6 for the CCE and LOB corpus.

Table 4. Frequency of the relevant meanings of 'should' in the CCE and LOB corpus

'Should'	CCE	LOB
Root	132 (68.8%)	117 (56.5%)
Epistemic	09 (04.7%)	28 (13.5%)
Quasi subjunctive	29 (15.0%)	38 (18.4%)
Indeterminate	22 (11.5%)	24 (11.6%)
Total	192 (100%)	207 (100%)

Examples of root 'should' can be assigned to a continuum whose extremes coincide with a subjective/strong core and with an objective/weak periphery. Out of the 132 occurrences of root 'should' in the corpus, 95 cases (72%) are the strong/subjective type. Objective cases are very rare, 5% (i.e. 9 out of 132) of the instances of 'should' in the corpus. In the LOB corpus, on the contrary, majority of the examples with subjective elements are found in spoken language accounting for 71% (i.e. 60 instances) of the total occurrences in the Survey sample (see Coates 1983: 60)

A significant difference between the two corpora with respect to root 'should' is the word range preceding 'why-clauses'. Johns (1991) reports that in the Birmingham corpus, adjectives and nouns (e.g. puzzling, clear, mystery) are frequent in this context. Other elements of context reported in this same study include

“probe into why” and “reason(s) why”. These evaluative elements are absent in the CCE (see Examples (15) to (17)).

- (15) ... he escaped, the fact is clear why should a first-class criminal be left loose?
(op 141)
- (16) ...Interrupted Ondoa. Why should a civil servant himself apply for the decision?
(op 19)
- (17) Why should anybody assume that bringing administration closer to the people
will solve all this...?
(op 65)

Epistemic ‘should’ expresses a tentative or less certain assumption than epistemic ‘must’. It stands in the same relation to epistemic ‘must’ as root ‘should’ to root ‘must’. Whereas the writer/speaker in the former says ‘I think it’s probable...’. In the following examples ‘should’ can be paraphrased as ‘I assume that...’

- (18) It may be since he is your close friend you should have his address. Thanks
very much for the “ndolé”?...¹⁴ (PL 102) (‘I assume that you have the
addresses)
- (19) Philosophise or express your beliefs. In one hour you should be able to talk
about every single point in your work...
(M 1089 (‘I assume that you will be able to ...’)

These examples combine subjectivity with logical assumption. Example (19) has a future time reference (‘... you should be able to talk...’), and it happens to be the only example in the corpus with such a reference. Coates (1983: 65) reports that two thirds of the epistemic examples of ‘should’ in the Survey (24 out of 36) have future time reference and the LOB sample very similar (i.e. 18 out of 28).

In subordinate ‘that-clauses’, ‘should’ is common as a quasi-subjunctive in CamE. Up to 14.5% of the tokens are of this type (i.e. 29 out of 200). Similar figures are registered in the LOB corpus: 18.4% (i.e. 38 out of 207). However, the elements of context differ considerably in the two corpora. In the CCE, 8 verbs and 1 adjective precede ‘that-clauses + should’; in the LOB corpus, 5 verbs (with varying frequencies), 8 nouns and 23 adjectives are found to occur in this context (Coates 1983: 69). No corpus example of this use of ‘should’ is found in Australian English (Coates and Leech 1980).

14. A locale dish of vegetables.

3.2 Frequencies and common senses of the modals

I recapitulate in Table 5 the frequencies of modals in the CCE, starting from the most frequent to the least frequent. The LOB figures are given for comparison (cf. Leech 1980).

Table 5. Absolute frequency distribution of modals in the CCE and LOB corpus

Modal	Raw frequency in the CCE	Modal	Raw frequency in LOB corpus
will	2,294	would	3,002
can	1,341	will	2,804
should	985	can	2,141
must	979	could	1,744
would	969	may	1,323
may	877	should	1,285
could	837	must	1,131
might	564	might	775
shall	336	shall	352
ought	54	ought	105
need	9	need	72

These crude totals yielded other sub-divisions based on the root/epistemic distinction and the degree of variation between CamE and BrE was considered to be significant. The investigation made use of traditional notional categories such as ‘ability’, ‘permission’, ‘possibility’, ‘necessity’, and ‘obligation’, while excluding from the count ‘unclear cases’ (i.e. textual instances which could not be assigned to one category or another because of their indeterminate position on a cline). The analysis indicates that the majority of examples can be accounted for by using these categories. The following list, therefore, indicates which meaning is most common for each modal; a semantic label is given, followed by the percentage that the meaning constitutes in the sample studied.

Table 6. Distribution of common senses of the modals in the CCE and LOB corpus

Modal	Most common senses in the CCE and LOB corpus	CCE %	Sample CCE	LOB %	Sample LOB
will	epistemic ‘prediction’	64	200	53	203
can	possibility	47	200	64	231
should	root ‘obligation’	71	200	51	229
must	root ‘obligation’	64	200	65	236
would	hypothetical ‘prediction’	52	200	58	219
may	epistemic ‘possibility’	63	200	61	236
could	root ‘possibility’ (past)	30	200	29	210
might	hypothetical	76	200	37	221
shall	root ‘obligation/necessity’	72	200	34	225
ought	root ‘obligation’	77	54	85	105

Notable findings in Table 6 can be seen in the discrepancy of percentages of the most common senses of the modals in relation to sample figures. However, at more specific levels, the following findings are outstanding:

- i. The performative use of 'must' in CamE is generalised. In BrE, performative 'must' is restricted to those contexts where authority structure is well established (e.g. the home, the courts).
- ii. The use of 'must' as a 'pseudo-exhortation' is used in a generalised manner in CamE. In BrE, such uses of 'must' are typical of spoken language, especially lectures and other oratories.
- iii. The holophrastic use of 'must' in CamE is highly restricted to letter writing. Out of 20 instances of this use of 'must', 17 are from Private Letters. In BrE, this usage is characteristic of spoken language.
- iv. 'Will' is the most frequently used modal for making requests in CamE. In BrE, 'would' is the most preferred form.
- v. 'Must' and 'should' are the major exponents of root meaning in CamE. In BrE, the figures are lower and in Australian English, the two modals mostly express epistemic meaning (Collins 1991: 164).
- vi. The hypothetical use of 'would', and 'should' is less frequent in CamE. In BrE, this meaning is frequently expressed. The Cameroonian user of English seems to find it difficult to cope with sentence structure involving the hypothetical use of modals. This is also true of Indian English (cf. Katikar 1994). Within the range of hypothetical examples, however, very few are used in other specialised pragmatic ways.
- vii. There are frequent cases of uncertainty in the use of 'will' and 'would' in conditional constructions and indirect speech. The most common slip seems to be a failure to conform to the rules of the sequence of tenses. There are instances of the use of 'will', where 'would' is most appropriate and vice versa. Speculatively, the use of 'would' instead of 'will' may be attributed to the general uncertainty of the exact use of the two modals.

4. Discussion and conclusions

This chapter has introduced basic corpus compilation steps and highlighted the relevance of corpus linguistic work for ESL varieties. Several steps involved in the process of corpus exploitation have been discussed and results of a small in-depth discussion of the modals that express obligation and necessity in the CCE presented. The different facets of meaning associated with the modals investigated have practically been easy to track down simply because of the use of a corpus-based methodology, which enables a large amount of data to be accessed and processed

with speed and accuracy, and occurrences counted, classified and described. The corpus method has expanded the scope of linguistic study, hence facilitating comparison across text types and national varieties (e.g. CCE and LOB corpus figures in this study). Students can compile their own small corpora to investigate patterns of language features of choice or even subset of academic registers such as essay introductions, conclusions of dissertations etc. These research goals are all permissible. Counting frequency information does not only help the researcher classify linguistic features and analyze them. They may also be useful in helping classroom teachers make decisions on which features to lay emphasis on in teaching or in material design. For example, with respect to the modals of obligation and necessity discussed above, both root and epistemic meanings occur; but the root-epistemic proportion varies for each case with respect to the CCE and LOB corpus. On the basis of such differences, teachers can design material that highlights certain meanings that they think require adequate attention.

On the basis of the discussion above, it becomes clear that analyses based on quantitative data (corpus) can make informed contribution to research into ‘nativization’ (cf. Kachru 1990) of English within ESL varieties. For one thing, this kind of statistical information, if made available across varieties, can guide the researcher to confirm quantitatively and statistically, impressions gained from introspection or participant observation about qualitative differences between native and non-native Englishes (cf. Sinclair 1982, 1991; Schmied 1999). Overall, questions of a grammatical nature such as the often-cited hypothesis that the “BE verb + ing” construction has been extended in African Englishes (Platt et al. 1984: 72), which might otherwise had been difficult to answer by collecting ad hoc data, are now possible to confront with a corpus. Such a methodological procedure can help the corpus analyst handle any range of complex data in language studies.

Acknowledgements

The quality of this chapter both in form and content has been improved by the comments and suggestions of three anonymous reviewers to whom I am very thankful.

References

- Akere, Funso. 1996. The Nigerian Corpus of English Project: A survey of Nigerian English usage. In *English in Ghana: Proceedings of the Inaugural Meeting of Ghana English Association, held at the University College of Education, Winneba, June 13–15*, Mary Esther Kropp Dakubu (ed.), 115–125.
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.
- Coates, Jennifer & Leech, Geoffrey. 1980. The meanings of the modals in modern British and American English. *York Papers in Linguistics* 8: 23–24.
- Coates, Jennifer. 1983. *The Semantics of Modal Auxiliaries*. London: Croom Helm.
- Collins, Peter. 1988. The semantics of some modals in contemporary Australian English. *Australian Journal of Linguistics* 8: 237–58. <https://doi.org/10.1080/07268608808599400>
- Collins, Peter. 1991. The modals of obligation and necessity in Australian English. In *English Corpus Linguistics*, Karin Aijmer & Bengt Altenberg (eds), 145–165. London: Longman.
- De Klerk, Vivian. 2002. Toward a corpus of Black South African English. *South African Journal of Linguistics and Applied Language Studies* 20(1): 25–35. <https://doi.org/10.2989/16073610209486296>
- Esimaje, Alexandra. 2016. A corpus-based study of tense use by advanced learners of English in Nigeria. In *Academic Writing Across Disciplines in Africa: From Students to Experts*, Josef Schmied & Daniel A. Nkemleke (eds), 33–146. Göttingen: Cuvillier Verlag.
- Esimaje, Alexandra. 2017. Comparing spelling usage by advanced learners in Nigeria and Cameroon: A Conacell-based study. *Review Syllabus* Vol. 7(1): 161–183.
- Greenbaum, Sydney. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7: 315. <https://doi.org/10.1111/j.1467-971X.1988.tb00241.x>
- Greenbaum, Sydney (ed.). 1994. International Corpus of English Project Newsletter 18. *World Englishes* 7: 315. <https://doi.org/10.1111/j.1467-971X.1988.tb00241.x>
- Haliday, Michael A. K. 1970. Functional diversity in language as seen from a consideration of modality and mood in English. *Foundation of Language* 6: 322–361.
- Huddleston, Rodney. 1984. *Introduction to the Grammar of English*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139165785>
- Kachru, Braj B. 1990. *The Alchemy of English: The Spread, Functions, and Models of Non-native Englishes*. Oxford: Pergamon Press.
- Katkar, Pratibha B. 1994. The Meaning of the Modals in Indian English. PhD dissertation, Shivaji University, Kolhapur.
- Leech, Geoffrey. 1971. *Meaning and the English Verb*. London: Longman.
- Leech, Geoffrey. 1980. The meaning of the modals in modern British and American English. *York Papers in Linguistics* 8: 23–34.
- Lyons, John. 1977. *Semantics*. Cambridge: CUP.
- Nkemleke, Daniel. 2003. A Corpus-based Study of the Modal Verbs in Cameroonian English. PhD dissertation, University of Yaoundé 1.
- Nkemleke, Daniel. 2008. *Manual of Information to Accompany the Corpus of Cameroon English*. Chemnitz: Chemnitz University of Technology, Department of English.
- Palmer, Frank. 1974. *The English Verb*. London: Longman.
- Palmer, Frank. 1979. *Modality and the English Modals*. London: Longman.
- Platt, John, H. Webber & M. Ho. 1984. *The New Englishes*. London: Routledge.

- Quirk, Randolph. 1961. The survey of English usage. *The Linguistic Reporter* 3(2): 3–7.
- Renouf, Antoinette. 1987. Corpus development at Birmingham University. In *Looking Up: An Account of the COBUILD PROJECT in Lexical Computing*, John M. Sinclair (ed.), 1–40. Glasgow: Collins.
- Schmied, Josef. 1989. Conference report: Compiling a corpus of East African English. *ICAME Journal* 13: 75–77.
- Schmied, Josef. 1999. Applying contrastive corpora in modern contrastive grammars: The Chemnitz Internet Grammar of English. In *Out of Corpora: Studies in Honour of Stig Johansson*, Hilde Hasselgard & Signe Oksefjell (eds), 21–30. Amsterdam: Rodopi.
- Schmied, Josef. 1990. Corpus linguistics and non-native varieties of English. *World Englishes* 9(3): 255–268. <https://doi.org/10.1111/j.1467-971X.1990.tb00264.x>
- Sinclair, John. 1982. Reflection on computer corpora in English language research. In *Computer in English Language Research*, Stig Johansson (ed.), 243–252. Bergen: Norwegian Computing Centre for the Humanities.
- Sinclair, John (ed.). 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
- Svartvik, J. & Wright, D. 1977. The use of ought in Teenage English. In *Acceptability in Language*, Sidney Greenbaum (ed.), 179–201. The Hague: Mouton.
- Tiomajou, David. 1993. Designing the Corpus of Cameroon English. *ICAME Journal* 17: 119–124.
- Trudgill, Peter & Hannah, Jean. 1982. *International English: A Guide to Varieties of Standard English*. New York NY: Routledge Chapman and Hall.
- Wells, Gordon. 1979. Learning and using the auxiliary verb in English. In *Language Development*, Victor Lee (ed.). London: Croom Helm.
- Von Wright, Georg H. 1951. *An Essay in Modal Logic*. Amsterdam: North-Holland.

PART II

Corpus-based analysis of African Englishes

Evaluating explanations for past-time reference with unmarked verb forms in African Englishes

Bertus van Rooy
North-West University

Previous research suggests that some varieties of English in Africa do not mark the past tense consistently with an inflectional suffix. Various explanations are offered for this state of affairs, including the option of the historical present tense, non-marking because the context already makes clear that the verb denotes a past event, phonological reduction, and limited English language proficiency. This chapter reports on a corpus analysis of spoken conversation in Nigerian English and Black South African English, which indicates that the non-marking of the past tense occurs in about one in every five contexts where an event in the past is represented. However, no convincing support is found for any of the explanations in the previous research on the topic.

1. Introduction

Unmarked finite verb forms are conventionally interpreted as present tense for non-third person-subjects. This is not so straightforward in African Englishes, though. Mesthrie and Bhatt (2008: 58–61) and others note the past tense is not always marked in Outer Circle varieties, due to various possible reasons, such as phonological reduction, contextual predictability, or transfer (see also Makalela 2013: 101–102; Meierkord 2012: 127–128). Example (1) illustrates the phenomenon, where past time verbs in the past tense form are underlined, while those that are not marked for past tense are printed in bold.

- (1) I said beside you came to France you're supposed to peck me like this
he didn't peck o
he say
he said no no you know I cannot
I say thank God they will not discipline you (ICE-NIG con_55)

Evidence for the extent of the feature is not clear-cut. The feature is identified more strongly with American and Caribbean varieties of English (Schneider 2008: 768; Meerkord 2012: 127), but is noted for Nigerian English (Alo & Mesthrie 2008), East African English (Schmied 2008), and Cape Flats English (McCormick 2008) in the *Handbook of Varieties of English*. Schmied (2008: 452) notes that the non-marking of verb inflection is stigmatised in East African English, and only rarely used by educated speakers. Alo and Mesthrie (2008: 325) also note, with reference to Jowitt (1991: 117) that it is a stigmatised feature in Nigerian English.

In his chapter in the same volume, Mesthrie (2008: 190) points to the claim in De Klerk and Gough (2002: 362) that the use of the present tense (inflected) form to refer to events in the past is a feature of Black South African English. He adds, however, that it is unclear from their presentation whether the form is taken from written or spoken data, and also notes that this usage is unattested in his spoken sociolinguistic interview data. However, as far as Black South African English is concerned, Gough (1996: 62), Van der Walt and Van Rooy (2002: 122), Makalela (2004: 360), and Meerkord (2012: 127–128) all identify the feature in Black South African English. Makalela (2013: 102) reports that it is the third most frequent non-standard feature in his corpus of radio broadcasts.

Apart from questions about the extent to which the feature is attested, the relative strength of the various explanations and the possible interaction between these factors have not yet been established. Mesthrie and Bhatt observe (2008: 59–60) that there are multiple possible explanations, such as the historic present tense, contextual cues like adverbials or adjacent main clauses in the past tense, or phonological factors. They also add that these explanations might interact, but they do not go deeper into the possible nature of such interaction.

This chapter aims to test the explanatory value of the various explanations offered by previous studies, by analysing a sample of spoken conversation data from Black South African and Nigerian English, and evaluating the relative strength of each explanation statistically. The choice of these varieties is motivated by the strength of claims for the likelihood of them showing the phenomenon in question based on previous literature, the fact that they represent two very different regions in Africa with different histories of transmission of English, and the availability of corpora of spontaneous conversation. The other potential candidate would have been the East African region, as represented by the ICE-East Africa corpus, but as Hudson-Ettle and Schmied (1999: 6–7) explain in the corpus manual, spontaneous conversation in English does not take place very often, thus a smaller number of conversations were included in this corpus, and these were collected either in classrooms in Kenya with students discussing topics amongst themselves, or with European (British and German) fieldworkers in conversation with Kenyans. Hence, the data were considered insufficiently similar to the other two corpora to

exclude effects from the nature of the data. The explanations or motivations for the non-marking of the past tense are reviewed first, in order to operationalise them for purposes of data analysis. This is followed by a discussion of the method, and then the presentation and interpretation of results.

2. Possible explanations for the non-marking of the past tense

A number of possible explanations for the occurrence of the feature have been proposed by researchers. Mesthrie and Bhatt (2008: 58–60) offer the most comprehensive list of possible explanations, but their proposals are supplemented by further explanations from other researchers.

1. The use of the present tense to denote past events is explained as the *conversational historic present*, where the present tense is used in conversation when the speaker recounts a story (Mesthrie & Bhatt 2008: 58). Meierkord (2012: 128), in reference to Biber et al. (1999: 455), argues that this is a particularly prevalent choice in conversation narrative, and associates with verbs of reported speech, such as *say*, to increase the dramatic effect of the retelling of a story. Example (2) illustrates an instance where a speaker, PK, tells a listener, PG, a story in chronological order, with occasional back-channels from the listener, which illustrates a case of the historic present, although interspersed with some past tense verbs as well. Present tense forms are in bold, and past tense forms are underlined

- (2) <#DPC021:160:PK><[> so</[></[> at that time i was born i was born on th- on that side¹
 <#DPC021:161:PG> <{><[> mm </[>
 <#DPC021:162:PK> <[> so </[></[> i **grow** up there **grow** up
 <#DPC021:165:PG> mm
 <#DPC021:170:PK> then i **go** to school on that side
 <#DPC021:175:PG> yes
 <#DPC021:180:PK> i was starting on primary school
 <#DPC021:185:PG> yes
 <#DPC021:190:PK> george dickerson
 <#DPC021:195:PG> ja <{><[> ??? </[>

1. The Xhosa English corpus was transcribed using only lower case letters, including for the first person subject pronoun 'I' and proper nouns. This choice by the compilers is respected in the reproduction of examples.

<#DPC021:200:PK> <[> till </[></[> standard five grade five then i was
going to mary waters
<#DPC021:205:PG> <{><[> yes </[>
<#DPC021:210:PK><[> then</[></[> i i **go** to mary waters and i **finish** my
<{><[> there </[> dpc021

2. Mesthrie and Bhatt (2008: 59) note that constraints on the use of tense in the native languages of speakers of World Englishes may be transferred to the choice of tense forms in English as well. Makalela (2004; 2013) offers a very specific instance where this factor potentially plays a role, to which he assigns the general label of “Bantu language logic”, which is transferred to English. He notes that in Sepedi (but the same applies to other South African indigenous languages), a consecutive tense option exists, where the second and subsequent verbs in a complex and compound sentence are used in the present tense form, after the first verb has been marked for the past tense. He also notes, however, that a different subject concord prefix is used (Makalela 2014: 101). He offers the following example from his data (3a), for which he provides the Sepedi equivalent, repeated here as (3b).

(3) Consecutive tense (Makalela 2013: 101)

- a. We played this ... jem and we **dance**. (where time reference is a week ago)
- b. Re ralokile mnino wo ebile ra **bina**.
sc play-PST-FV music this also sc dance-FV²
'We played this music and we **dance**.'

Mesthrie and Bhatt (2008: 59) also note a related possibility, where a subordinate clause uses a present tense verb after the main clause has already used a past tense verb. This appears to be a subset of cases denoted by Makalela's consecutive tense explanation, which includes both subordination and coordination.

3. *Adverbials* or other elements of the context can serve to mark the time of the event (De Klerk 2003: 234; Meierkord 2012: 127–128), which makes the morphological marking redundant. This is illustrated by (4) below, where the first line has a present tense verb after the adverb “before”, which gets a continuation in the next line with the adverbials “then” and “in 80's”.

- (4) before we **have** the grahamstown boxing club and fundisile boxing club then
there was a boxing in grahamstown in 80's (dpc064)

2. The gloss is based in part on the gloss provided by Makelela (2013) himself, but slightly more formalised. SC=subject concord, PST=past tense suffix, FV=final vowel. Makalela (2013: 101) notes that the second subject concord morpheme takes the exceptional final vowel -a, rather than the regular -e, to indicate that this is consecutive tense.

4. *Phonological reduction* provides a very different kind of explanation for the absence of past tense marking, where complex consonant clusters are avoided. This can happen due to the final consonant of the verb stem, or when the next word starts with plosives or perhaps other obstruents, as shown by the following example that Mesthrie and Bhatt (2008: 59) quote from Platt, Weber and Ho (1984: 69), where the action of moving is situated in a past time.

(5) I move to hostel.

5. *The level of education* is the final possible explanation from previous research, which, if supported by the data, would suggest that the non-marking of the past tense is not an emerging feature of a new variety, but rather a transitional learner phenomenon. This observation is made by Schmied (2008: 452) and Alo and Mesthrie (2008: 325) for East African and Nigerian English respectively. Minow (2010: 121–124) finds the same for Black South African English, that the more proficient a user is, the less likely he/she is to omit the past tense morpheme. However, Meierkord (2012: 127–128), who analyses a subset of the Black South African English data analysed by Minow (2010) together with the data of speakers of other varieties of South African English, does not find confirmation for this effect.

3. Method

Previous research identified Black South African English and Nigerian English as among the African varieties where the unmarked verb forms with past time reference are most likely to occur, and also indicates that this phenomenon is strongly linked to conversation. Therefore, an analysis was undertaken of Black South African English conversation, taken from De Klerk's (2006) corpus of Xhosa-English, and the spoken conversation from ICE-Nigeria (Wunder, Voormann & Gut 2010), in order to assess the validity of the various explanations advanced in the literature.

From each corpus, a random selection of conversations was made – 26 from the Xhosa-English corpus and 30 from the ICE-Nigeria corpus.³ These files were opened up in a text browser, and the first two pages on screen were selected and annotated by hand for the formal realisation of tense on verbs, alongside an identification for each verb of whether it denotes an event in the past, the present, or hypothetical space. The analysis was limited to simple, non-aspectual forms of the verb phrase. All progressive and perfect forms were ignored, because their

3. The total wordcount of the sample from the Xhosa-English corpus was 28395 tokens, and from ICE-Nigeria 15239 tokens.

interpretation as present or past time reference introduces further complications that may interfere with the accuracy and clarity of the analysis.⁴

Using contextual cues, each verb was classified for its reference to a past time event or a present tense use, either for events in the present or general truths. In practice, the procedure employed was largely similar to that of Minow (2010: 107), where calendar dates, other temporal adverbials, or expressions that imply anteriority were used to identify events or states that were true of the past. Hypothetical events, situated in the present or past, were identified as a separate category and excluded from the subsequent analysis. A typical present time use in the data was the expression of stance by speakers, which pertained to the moment of speaking, as illustrated the use of the verb “think” by Speaker 1 in Example (6).

- (6) <Speaker 1> ah well I **think** I got into this school two thousand and five
 <Speaker 2> and since then
 <Speaker 1> studied English
 I’ve been here yes
 but
 I just noticed it was Mister Ade I used to see first
 <Speaker 2> all right
 <Speaker 1> I used to see Mister Ade
 and then there was this other guy
 who was always going around in native
 <Speaker 2> okay
 <Speaker 1> and then there was you (ICE-NIG, con32)

The remainder of Example (6) mainly contains verbs that denote events situated in a past time-frame. Speaker 1 indicates early on in the extract that the events surrounding his going to university took place in 2005, and from the rest of the passage, it was deduced that all events recounted related to the same time-frame, and hence all instances of past tense verbs underlined in the example were taken to denote events in the past. The present perfect (‘ve been) was excluded from analysis.

Example (7) similarly refers to events in the past, but some of the verbs, printed in bold, use the form of the present tense, while others use the past tense forms. There is no ambiguity in the passage, it remains clear that the entire set of events being retold took place in the same past time-frame, the year 2009 when Speaker 1 arrived at university, and was struck by the difference in food in Yorubaland. The

4. One of these complications is that the simple past tense and the present perfect are in competition for the expression of events that occurred (or at least started) in the past. This competition has been examined for Nigerian English (Werner & Fuchs 2017) and Black South African English (Van Rooy 2016) elsewhere.

verbs in the present tense form (**bolded**) are therefore the ones that are of particular interest to the phenomenon under investigation in this chapter. Two occurrences of the verb “know” in the present tense are not highlighted in passage (7); they represent instances where the speaker leaves the past time-frame and makes contact with the other speaker in the present, seeking confirmation that the interlocutor still follows the story, and are therefore regarded as standard English usage of the present tense.

- (7) <Speaker 1> one of the things that I experienced from the Yoruba land
for the first thing is about the food
bef- before because you know by the first time we came to
the Moro Campus
<Speaker 2> Mo- Moro Campus
<Speaker 1> where we did our pre-degree
<Speaker 2> yeah yeah
P D
<Speaker 1> we seven eight to
<Speaker 2> that is two thousand
<Speaker 1> two thousand and nine session
<Speaker 2> two thousand and nine se-
yeah
<Speaker 1> so at that time
first thing that we
come across
you know what immediately when you **come** to a place the first
thing that you **look** for
is food (ICE-NIG, con14)

The morphological form of a verb (present/past) with past time reference served as dependent variable for the statistical analysis. Each instance of a morphological past tense and each instance of present tense form with past time reference was then further classified for each of the possible explanations identified from previous literature. These classifications were encoded as separate variables that served as independent variables in the statistical analysis.

The conversational historic present was limited to cases where a past event is clearly retold. Thus, the speaker should have made it clear that s/he recounts past events. However, since the difference between direct and indirect speech is quite fuzzy in spoken conversation, and since backshifting is a stylistic option but not an absolute grammatical requirement (Huddleston 2003: 151–158), cases where speech reporting took place, and the present tense was selected, were excluded from the count. In agreement with Minow (2010: 109), who also excluded reported

speech, the reporting verbs in the main clause that introduced reported speech were included, since they could either denote reporting in the present or reporting what happened in the past. A binary feature – narrative / not – was encoded for each instance where the present tense form was used in a past time-frame. This feature cannot be applied to instances where the past tense form was selected, since such cases are by definition never “historic present”. Therefore, only descriptive results are reported for this feature, and it was not used in the inferential statistical analysis.

Consecutive tense, as proposed by Makalela (2003; 2014), was encoded as cases where a non-initial event in a (narrated) sequence of events is presented in the present or past tense form, i.e. for all instances of past time reference, verbs that were not the first verb in a series of clauses that are semantically connected, were classified as consecutive verbs. Verbs that initiated a sequence, and verbs not part of any sequence, were classified as non-consecutive. A binary feature of Ctense or not was therefore encoded, using the abbreviation that Makalela (2013) uses for the phenomenon of consecutive tense. The suggestion by Mesthrie and Bhatt (2008: 59) that subordinate clause verbs could be in the present tense form if they are under the scope of a main clause verb in the past tense form is also in part captured by this variable. However, no extensive syntactic analysis is undertaken of the spoken data to determine main and subordinate clause status, as this is often ambiguous in spoken language. The feature is simplified to an ordering relationship between cohering clauses. This feature is also useful to account for the possibility that narrative sequences are sufficient cover for the unmarking of past tense. In a sense, therefore, this feature compensates for the historical present analysis that could only be applied partially to the data.

The encoding of context for purposes of analysis was limited to specific adverbials (De Klerk 2003: 234; Meierkord 2012: 127–128) that overtly situate an event in the past. Since a wide range of contextual cues had to be combined to classify the forms as denoting present or past time events, it was deemed vacuous to argue that any verb that could be linked to the past from contextual cues was therefore licenced to be in the present tense form, hence the stricter encoding of adverbials as contextual cues. These adverbials had to be present in the same clause as the verb, or else in an adjacent clause that relates to the target verb as an adverbial clause. The variable is therefore binary: an adverbial is either present or absent within the context of the clause of the target verb.

Phonological reduction was encoded in a number of ways. Since this only applies to regular verbs, where complex consonant clusters are avoided (Mesthrie & Bhatt 2008: 59), irregular verbs were set aside as a separate category. Regular verb stems were marked as *simple* if they ended in a vowel or non-nasal sonorant consonant. The following specific classes, that might increase the consonant cluster complexity in combination with the inflection /d/, were identified: *nasal*, *fricative*,

and *plosive*. Among the plosives, cases where a stem ended in a /t/ or /d/, and where normally vowel insertion takes place, were marked as *identical*.

Education as a variable was only coded for the Xhosa-English corpus of Black South African English, where complete details were available for each speaker from the corpus manual. The data were reduced to three levels: (1) incomplete secondary education, (2) secondary education completed, and (3) post-secondary education completed. It turned out, however, that the vast majority of participants fell into category (2) of having completed their secondary education. The same variable is not specifically encoded in the ICE-Nigeria corpus. The corpus manual provides occupations for participants where available, but for the 168 speakers who contributed to the spoken conversations, detail is provided for only 29 participants, of whom 27 were students. It is thus impossible to construct levels for this variable based on the available information.

Two kinds of statistical analyses were performed. For each corpus separately, each independent variable was related to the two or more levels of the dependent variable in a χ^2 -test, to determine if that variable, on its own, provides any kind of account of the distribution of marked and unmarked past tense cases. The basic null hypothesis for all variables is that the distribution of the inflected past tense and the present tense forms should be the same, whether or not the explanatory force is present or absent in a particular context. The alternative hypothesis, that would argue in favour of each explanation, is that the proportion of present tense forms would be higher in the presence of the explanatory force than in the absence thereof. Thus, support for the alternative hypothesis would be that:

- there are more present tense forms in the contexts defined by consecutive tense than elsewhere,
- more present tense forms in the contexts where adverbials indicate location in the past time-frame than elsewhere,
- more present tense forms in phonological environments that constrain the realisation of the final plosive of the past tense suffix than in simple contexts with stems than end in vowels or sonorant consonants, or
- more present tense forms in the speakers with lower levels of education.

Thereafter, the variables, except *historical present*, were entered into a decision tree model, implemented in *Statistica* version 13.2 (see Statsoft 2005: 26–29), to see if some larger, coherent picture of the data can be obtained through the combination of variables. Decision trees are based on similar algorithms as regression models, and work in a stepwise fashion to select the most informative variable, performs a binary split on the data in terms of that, and from there on, continues with the same process until the most exhaustive classification of the data is obtained. The

extent to which the classification is continued was limited by the condition that classification only continues until at least 100 cases remain to be split. This is done to avoid overfitting of the model into much more detailed subclassifications.⁵

4. Results

The overall results for the frequency of the choice between present and past tense forms in past time-frames are presented in Table 1.

Table 1. Overall raw frequency of present and past tense forms (with percentages in brackets, together with normalised frequencies)

	Past tense form	Present tense form	Normalised omissions per 1,000 words
Nigerian English	345 (80%)	85 (20%)	5.58
Black South African English	410 (78%)	115 (22%)	4.05

Previous studies that undertook comprehensive quantification of the phenomenon of non-marking of the past tense are few. Minow (2010: 111) in an analysis of the Corpus of Black and Coloured South African English in contact finds that Black speakers have an omission rate of 15.2%, from a total of 1381 possible instances of the past tense. Makalela (2013: 99) reports that the phenomenon of omission occurs with a frequency of 0.69 per 1,000 words. The present results suggest in the first place that the frequency of omission in the data from Black South African and Nigerian English is quite similar, and furthermore that a relatively similar rate is reported by Minow on a different corpus.

Makalela’s (2013) number cannot be expressed as an omission rate in the absence of a report of the total number of past tense forms used, but if one extrapolates from the normalised frequency to a corpus that is reported to contain 209,000 words, a number of approximately 144 instances is obtained. However, Table 1 reports the frequency of omission per 1,000 words, while it is possible to convert Minow’s finding to the same scale. Minow (2010) used a corpus of 85,431 words (Minow 2010: 99). Thus, if 247 unmarked verbs denoting events or states in the past occurred in the entire corpus, the frequency normalises to 2.89 per 1000 words,

5. The same data were also submitted to a decision tree classification in *R*, using the *Ctree* function in the *partykit* package. There, the setting is that a split is only made when a statistically significant result can be obtained by the split, which is a much more stringent condition than the one imposed by *Statistica*. In consequence, a much simpler and less informative tree was obtained in *R*, which is compared to the *Statistica* tree in the results section.

considerably more frequent than the number reported by Makalela, and closer to the number obtained for the Xhosa-English corpus of De Klerk.⁶

Makalela (2013) makes the following claim in interpreting his data against the background of previous research:

In the isiXhosa corpus data, there are not cases of consecutive tense. The only tense related cases in the isiXhosa corpus include extensive use of progressive aspect, regularization of verbal systems and overgeneralization of past tense (De Klerk 2003: 477).
(Makalela 2013: 102)

It appears as if he bases the claim that De Klerk's corpus does not contain evidence of consecutive tense on the report of the analysis by De Klerk (2003), rather than on an independent analysis of the corpus itself. When an analysis of the corpus itself is undertaken, as is the case in this chapter, then it emerges that there are such instances. The suggestion that the use of consecutive tense is limited to the Sotho group of languages, which Minow (2010: 106) offers, with reference to Makalela (2004) and Finn (1986), appears to be based on available previous findings, which were limited to examples from the Sotho languages, rather than positive evidence that this phenomenon is not also attested in the Nguni family of languages, of which isiXhosa is a member. As is indicated below, instances representing consecutive tense are found in the Xhosa-English corpus. Differences between the findings cannot be resolved in the absence of publicly available corpora, and thus the issue will have to be left here, while we proceed to the analysis of the specific individual factors that may provide an account of the data.

The first possible explanation for the use of the present tense form in past time contexts is the so-called historical present. According to this explanation, retelling a story from the past may invite the speaker to switch to the present tense to achieve dramatic effect (Biber et al. 1999: 454–455). This is not so much a grammatical

6. Apart from reporting the normalised frequency of 0.69 per 1,000 words in Table 2 (Makalela 2013: 99), he continues in the discussion section to report much larger numbers, where he writes the following: "The participants of the study produced consecutive tense usage with a total normalised frequency of 11,377 out of 16,396 obligatory contexts. This feature had the third highest relative frequency of 0.69 per 1,000 words in the radio corpus." (Makalela 2013: 102) I am unsure how to resolve the difference in the two quantifications within the same article, as a frequency of 11,377 in a corpus of 209,000 words (Makalela 2013: 98) translates to a normalised frequency of 54.44 per 1,000 words.

In deciding how to report Makalela's (2013) findings, I am guided by considerations of probability. The number of 0.69 per 1,000 words is in the same order of magnitude of numbers reported elsewhere, whereas the number of 55.44 per 1,000 words is so large, that it exceeds the total normalised frequency of all past tense verbs reported for the Longman Spoken and Written English corpus by Biber et al. (1999: 456). I therefore suspect that the smaller number is the correct one and the larger number the results of a miscalculation. E-mail inquiries to the author to resolve the difference were left unanswered.

explanation than a stylistic choice, and hence it is impossible to identify the converse, that is cases where a speaker chooses to tell a story simply in the past, with accuracy. No further statistical analysis is undertaken of this option, except to report its overall frequency, which is done in Table 2.

Table 2. Overall raw frequency of present tense forms in contexts where narration of past events takes place

	Not narration	Narration	Total
Nigerian English	50	35	85
Black South African English	75	40	115

In Nigerian English, 40% of the cases of unmarked present tense form in past contexts can be read as historical present, while the corresponding percentage is 35% in Black South African English. However, similar narrative use also occurs in instances where the past tense form is used, which indicates that the stylistic option that native speakers use, is also available to these two groups of non-native speakers. The notion of historical present tense accounts for fewer than half of all instances of the present tense use in past time contexts, so there is clearly a need to look further at other possible explanatory factors.

Closely related to historical present tense is the notion of consecutive tense, as proposed by Makalela (2004, 2013). It is the only explanatory factor for the use of present tense forms with past time reference in his analysis. Table 3 presents the relevant results.

Table 3. Frequency of present and past tense forms (with percentages in brackets) in environments where the consecutive tense (Ctense) conditions are met and where the conditions are not met

	Conditions not met	Ctense conditions met	Total
<i>Nigerian English</i> ($\chi^2 = 1.22, p > 0.05$)			
Inflected past form	282 (82%)	63 (18%)	345
Present form	65 (76%)	20 (24%)	85
Total	347 (81%)	83 (19%)	430
<i>Black South African English</i> ($\chi^2 = 0.63, p > 0.05$)			
Inflected past form	331 (81%)	79 (19%)	410
Present form	89 (77%)	26 (23%)	115
Total	420 (80%)	105 (20%)	525

The possibility that consecutive tense motivates the omission is not supported by the data. The percentages of omission/present tense use are only slightly higher in the case of potential consecutive tense environments than elsewhere. The statistical

evaluation reveals that there is insufficient evidence to reject a null hypothesis here. Thus, while evidence of present tense forms in the contexts of consecutive environments can be found in the data, it cannot be said to explain the selection of the present tense form, as the selection of the past tense form in the same environment is approximately equally likely.

Another explanation, which certainly has good face validity, is that the inflection of the verb can be omitted if there is an adverbial in context that marks the past time-frame of the event or state denoted by the verb. The data from the two corpora show, however, that the opposite is true: the presence of an adverbial is associated with a higher likelihood of overt past tense inflection of the verb. Table 4 shows that the present form is selected mainly where there is no adverbial present. When looking at all present tense forms in Nigerian English, only 6% of them are in environments with an adverbial, while the remaining 94% occur in environments without adverbials. The corresponding percentages for inflected past tense forms indicates that the past form associates more strongly (21%) with adverbials than the present. The picture for Black South African English is similar, with the present tense forms more frequently associated with environments without adverbials (83%) than is the case for inflected past tense forms (71%).

Table 4. Frequency of present and past tense forms (with percentages in brackets) in environments where an adverbial marks the past-time frame and where there is no adverbial

	No adverbial	Adverbial present	Total
<i>Nigerian English</i> ($\chi^2 = 10.64, p < 0.05$)			
Inflected past form	273 (79%)	72 (21%)	345
Present form	81 (94%)	5 (6%)	86
Total	354 (80%)	77 (20%)	431
<i>Black South African English</i> ($\chi^2 = 7.50, p < 0.05$)			
Inflected past form	290 (71%)	120 (29%)	410
Present form	96 (83%)	19 (17%)	115
Total	386 (74%)	139 (26%)	525

If anything, the presence of an adverb increases the change that the verb is also inflected for the past tense. For both corpora, this effect is statistically significant, but then in the opposite direction of the alternative hypothesis, which means that the null hypothesis cannot be rejected. It remains for future research to explore if there isn't rather a priming or reinforcement effect that motivates the selection of the past tense forms in the environment of past time adverbials.

The next possible explanation considered in this study is the phonological shape of the stem ending. Although Mesthrie and Bhatt (2008: 59) mention the

possible effect of the onset of the following syllable on consonant cluster simplification, Van Rooy (2007: 34–35), in one of the few quantitative studies of the extent of consonant cluster simplification, indicates that the previous consonant within the same syllable coda has a much more extensive effect on consonant deletion than the onset of the following syllable. The only exception, which is consistent with the example in Mesthrie and Bhatt (2008: 59), is where the following syllable starts with an identical consonant to the final consonant of the previous syllable. To analyse the simplification of identical consonants, an analysis of phonetic data rather than just an orthographic transcription would be required, which falls outside the scope of the present study. Given the availability of the sound files for ICE-Nigeria, it would be possible in future research to explore this possibility for that corpus. Moreover, identical consonants across word boundaries would be a relatively small proportion of the overall contexts, with the prevalence of consonant cluster formation in the verb itself a more likely occurrence. The overall frequency of the presence or absence of the past tense suffix in different phonological environments is reported in Table 5. The initial phonological analysis was recoded to only three categories: obstruent, identical and sonorant, with irregular forms discarded from the analysis. This was done in the main because of data sparsity. Irregularly inflected verb forms were not considered in the analysis, since they would not be affected in the same way by the phonological constraint.

Table 5. Frequency of present and inflected past Tense forms (with percentages in brackets) in different phonological environments

Form	Obstruent		Identical	Sonorant		Total	Irregular
	Plosive	Fricative		Nasal	Simple		
<i>Nigerian English</i> ($\chi^2 = 4.71, p > 0.05$)							
Inflected past form	12	15	17	5	16	65	280
Present form	1	3	10	1	5	20	65
Total	13	18	27	6	21	85	345
<i>Black South African English</i> ($\chi^2 = 6.45, p < 0.05$)							
Inflected past form	9	14	23	14	20	80	330
Present form	8	6	7		5	26	89
Total	17	20	30	14	25		419

The χ^2 -analysis indicates no statistically significant difference for Nigerian English, but a small, statistically significant, difference in the expected direction for Black South African English, where obstruents are deleted slightly more frequently and sonorants slightly less frequently than the expected or overall average values. The effect is especially obvious in the case of plosives in Black South African English, which is consistent with the finding of Van Rooy (2007: 35) that final plosive clusters

are often simplified in this variety of English. Thus, a phonological account offers a partial explanation of a small number of cases in Black South African English, but this phonological account then has no grammatical implication, as speakers may well have intended to produce a past tense form, and in their minds, did so, with the phonological realisation not fully doing justice to their intent.

The final analysis, which was only performed on the Black South African English data due to the unavailability of corresponding information in the ICE-Nigeria metadata for the spoken conversations, was to determine if level of education, as a measure of proficiency in English, plays a role. The results are presented in Table 6.

Table 6. Distribution of inflected past tense forms and present tense forms by educational level for Black South African English (with column percentages in brackets)

Form	Incomplete secondary	Secondary completed	Post-secondary completed	Total
Inflected past form	30 (83%)	350 (77%)	29 (85%)	409
Present form	6 (17%)	104 (23%)	5 (15%)	115
Total	36	454	34	524

The results indicate that the vast majority of participants whose data were analysed, completed secondary education, which makes inferences about any pattern dubious, and does not offer enough statistical power to refute the null hypothesis or the alternative hypothesis. At the same time, the few participants with a lower educational level than completed secondary education had a higher proportion of inflected past tense forms than those who did complete their secondary education, so there isn't even tentative evidence in support of the alternative hypothesis.

Summarising across the various factors, it emerges that no explanation that is offered in the available literature provides a very compelling account for the omission of the inflected past tense in either Nigerian or Black South African English. The only strong statistically significant finding is a finding in the opposite direction than expected on the basis of previous literature, namely that the presence of an adverbial indicating the past time-frame of the event decreases the likelihood that the past tense inflection would be omitted. A very small phonological effect of obstruent cluster deletion also emerges from the analysis of the data from Black South African English only.

In view of the analysis of the individual explanations, a deeper analysis in terms of the possible interactions of the various explanatory factors seems unlikely to yield any further insight. This suspicion is in part confirmed by the decision tree that is obtained with the stricter selection criteria imposed by the algorithm implemented in *R*. This analysis indicates that only a single variable makes any significant contribution to the modelling of the data, which is the presence or absence

of an adverbial. Since this variable is in any case the most important variable of the decision tree extracted in *Statistica*, the discussion therefore is integrated with the discussion that follows the decision tree in Figure 1.

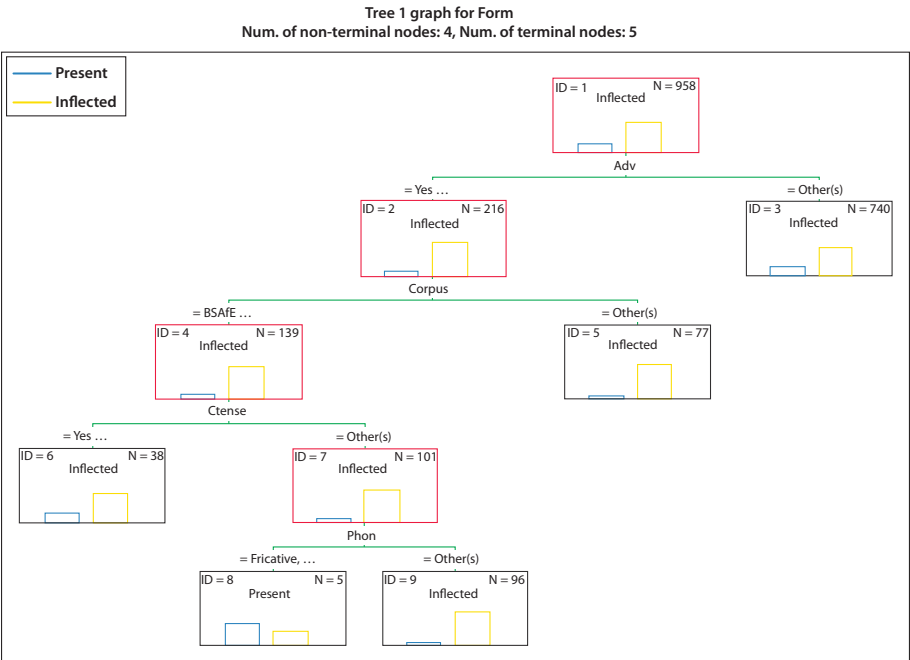


Figure 1. Decision tree of all data in *Statistica*

In the decision tree, three of the predictor variables were entered: The presence of an adverbial, Ctense and the phonological shape of the stem-ending, together with corpus, to see if there isn't perhaps an explanation that holds for the one corpus but not the other. Summarising across all the insight that we obtain from the tree, two main points emerge: the hypothesis that the use of adverbs makes the marking of the past tense superfluous is directly refuted, and other factors play a very marginal role in sometimes increasing the likelihood of non-marking of the past tense, but only in Black South African English.

The first split in the decision tree is on the variable adverbial. For cases where the verb is under the immediate scope of an adverb, 89% of all cases have a marked verb phrase, whereas when there isn't an adverb in the environment of the verb, marking takes place in only 76% of all cases. Thus, the adverb does not substitute for tense marking, but quite the contrary, the adverb and tense marking go together more often, and it is when there isn't an adverb that the tense marking is present somewhat less often.

On the left side of the decision tree, there is a further split for corpus. Black South African English allows some refinement, but no further factors contribute insight into the Nigerian English data. For Nigerian English, cases where an adverb is present shows an even higher past tense marking rate at 94%. For Black South African English, two factors appear to play a minor role when the adverb is present. The first split is on consecutive tense, as proposed by Makalela. In the 38 relevant cases, 9 (or 24%) show non-marking of the past tense, but this number is simply similar to the overall proportion of non-marking where no adverb is present, so it does not really provide compelling support for the strength of this factor, and is rather an artefact of the statistical method. The other factor that is identified by the decision tree is phonological. The final split shows that final plosives or fricatives increase the probability of non-marking of the past tense in the remainder of Black South African English cases, where there is an adverb but no consecutive tense. However, this finding again has to be taken with much caution, as it applies to exactly five cases, which means that we really ought to discard this finding altogether. This is consistent with the only statistically significant difference from the analysis of individual variables, where the phonological account for Black South African English omission is supported, but offers an explanation for a very small number of cases only.

5. Conclusion

The analysis of spoken data from Nigerian and Black South African English in this chapter does not provide support for the explanations provided in most previous research on the non-marking of the past tense. While there are instances that can be classified as conversational historical present (Mesthrie & Bhatt 2008: 58–59; Meierkord 2012: 128), this accounts for less than half of all observed instances of non-marking. Furthermore, there are narrative portions in the conversation that do not make use of the present tense option, so it would be opportunistic to “explain” data that fit the “explanation”, while ignoring data that do not fit the explanation. More specific further explanations also fall short of providing an account. The likelihood of present tense use in the context of consecutive tense (Makalela 2004: 360–361; Makalela 2013: 101–102), which also relates to the subordinate clause use (Mesthrie & Bhatt 2008: 59), is not higher than in the contexts not covered by this syntactic environment. In the presence of adverbials that specify the time, the past tense is more likely to be used than in contexts not so specified, which is the exact opposite from what one would expect from the explanatory account provided in previous research. Finally, while phonological factors show no statistically significant pattern in Nigerian English, a very small

number of instances can be accounted for as phonological simplification processes in Black South African English. However, this is not an account of the past tense use as such, but simply consistent with a general pattern of consonant cluster simplification in syllable codas, as Van Rooy (2007) shows for all codas, irrespective of whether they occur across a morpheme boundary or within a single morphological unit. No unexpected deeper insight emerges from the further statistical exploration by means of decision trees.

The main difference between this study and most previous research is the way in which corpus linguistics is employed as a method. While a number of previous studies made use of corpora, the corpora mainly served as material for exemplification of particular potential explanations, without testing the validity of the same explanations on the contexts where their past tense form was realised overtly, which is a necessary control mechanism for the validity of any explanation: only if the frequency of past tense marking is lower in an environment that is hypothesised to favour non-marking than in other environments can the inference be drawn that there is support for the explanation associated with that environment. Some researchers did use the corpora to provide a quantitative measure of the overall extent of the process, but did not extend the quantification to the comparison of overt and omitted past tense morphemes in the same environments.

One exception to the methodological limitations noted above is the study of Minow (2010) on Black South African English. She considers a range of factors that account for variation in past tense marking, such as proficiency, native language family, and finds support for a proficiency effect. Thus, the least proficient users of English are most likely to omit past tense marking, which confirms the report by Alo and Mesthrie (2008) and Schmied (2008) for Nigerian and East African English, that it is a learner phenomenon and not a strong candidate for a variety feature. However, even Minow (2010: 121–124) chooses to focus only on instances of omission of the past tense marking when considering the explanatory value of adverbials and the historic present, and hence does not control for the extent to which the omission rate is similar in contexts where these explanatory factors are absent. The limitations of the present corpora, where information about proficiency/education level is partly incomplete, and where the speakers are generally of similar backgrounds makes it unwise to reject Minow's account. However, the present study offers no corroborating evidence for her account, and it must remain an unresolved question for future research with corpora that represent a wider range of proficiencies than an ICE-corpus what is usually limited to participants who have at least completed secondary education.

In sum, this chapter contributes two important insights to the study of English in Africa from a corpus linguistic perspective. On a methodological level, it shows the value of not only collecting and analysing corpora, as a number of studies

reviewed in this paper have done, but also the need for corpus linguistic methods grounded in thorough quantification of variables across target and control contexts. Through this method, in the second place, it was possible to show that the non-marking of the past tense should not be accepted as a typical feature of either of the varieties of African English examined here, since it is not very frequent, and it does not show very clear linguistic motivation. Relying on the corpus findings of Minow (2010) and the suggestions by Alo and Mesthrie (2008) and Schmied (2008), it is best regarded as a feature of less proficient users of English, which is likely to decrease or disappear altogether with increased proficiency.

Acknowledgements

The author acknowledges the support of the National Research Foundation of South Africa (Grant Number 109369). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research is that of the author, and that the NRF accepts no liability whatsoever in this regard. An earlier version of this chapter was presented as a paper at the annual conference of the *International Association for World Englishes* in Syracuse, USA, 2 July 2017. Haidee Kruger commented on an earlier version of this manuscript. Their feedback, together with the comments of two anonymous reviewers, contributed to the final manuscript, but all remaining shortcomings are for my own account.

References

- Alo, M. A. & Mesthrie, Rajend. 2008. Nigerian English: morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 323–339. Berlin: Mouton de Gruyter.
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- de Klerk, Vivian. 2003. Xhosa English as an institutionalised variety of English: In search of evidence. *English World-Wide* 24(2): 221–243. <https://doi.org/10.1075/eww.24.2.05dek>
- de Klerk, Vivian. 2006. *Corpus Linguistics and World Englishes: An Analysis of Xhosa-English*. London: Continuum.
- de Klerk, Vivian & Gough, David. 2002. Black South African English. In *Language in South Africa*, Rajend Mesthrie (ed.), 356–378. Cambridge: CUP.
<https://doi.org/10.1017/CBO9780511486692.019>
- Gough, David. 1996. Black English in South Africa. In *Focus on South Africa*, [Varieties of English around the World G15], Vivian de Klerk (ed.), 53–77. Amsterdam: John Benjamins.
<https://doi.org/10.1075/veaw.g15.06gou>
- Finn, Stephen M. 1986. The abuse of English by students at Vista University. *English Usage in Southern Africa* 17: 1–8.

- Huddleston, Rodney. 2003. The verb. In *The Cambridge Grammar of the English Language*, Rodney Huddleston & Gregory K. Pullum (eds), 71–212. Cambridge: CUP.
- Hudson-Ettle, Diana M. & Schmied, Josef. 1999. *Manual to Accompany The East African Component of The International Corpus of English*. Chemnitz: Chemnitz University of Technology. <<https://www.tu-chemnitz.de/phil/english/sections/linguist/real/independent/ICE-EA/>> (13 November 2017).
- Jowitt, David. 1991. *Nigerian English usage*. Ikeja: Longman Nigeria.
- Makalela, Leketi. 2004. Making sense of BSAE for linguistic democracy in South Africa. *World Englishes* 23: 355–366. <https://doi.org/10.1111/j.0883-2919.2004.00363.x>
- Makalela, Leketi. 2013. Black South African English on the radio. *World Englishes* 32(1): 93–107. <https://doi.org/10.1111/weng.12007>
- McCormick, Kay. 2008. Cape Flats English: Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 521–534. Berlin: Mouton de Gruyter.
- Meierkord, Christiane. 2012. *Interactions Across Englishes: Linguistic Choices in Local and International Contexts*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139026703>
- Meshtrie, Rajend. 2008. Black South African English: Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 488–500. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110208429>
- Mesthrie, Rajend & Bhatt, Rakesh. 2008. *World Englishes*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511791321>
- Minow, Verena. 2010. *Variation in the Grammar of Black South African English*. Frankfurt: Peter Lang.
- Platt, John, Heidi Weber & Mian L. Ho. 1984. *The New Englishes*. London: Routledge and Kegan Paul.
- Schmied, Josef. 2008. East African English (Kenya, Uganda, Tanzania): Morphology and syntax. In *Varieties of English, 4: Africa, South and Southeast Asia*, Rajend Mesthrie (ed.), 451–471. Berlin: Mouton de Gruyter.
- Schneider, Edgar W. 2008. Synopsis: Morphological and syntactic variation in the Americas and the Caribbean. In *Varieties of English, Vol. 2: The Americas and the Caribbean*, Edgar W. Schneider (ed.), 763–776. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110208405>
- Statsoft. 2005. *Statistica Data Miner*. Tulsa: Statsoft.
- van der Walt, Johann L. & van Rooy, Bertus. 2002. Towards a norm in South African Englishes. *World Englishes* 21: 113–128. <https://doi.org/10.1111/1467-971X.00235>
- van Rooy, Bertus. 2007. Consonant clusters and resyllabification in Black South African English. *Language Matters* 38(1): 26–45. <https://doi.org/10.1080/10228190701640017>
- van Rooy, Bertus. 2016. Present perfect and past tense in Black South African English. In *Re-assessing the Present Perfect in English: Corpus Studies and Beyond*, Valentin Werner, Cristina Suárez-Gómez & Elena Seoane (eds), 150–168. Berlin: Mouton de Gruyter.
- Werner, Valentin & Fuchs, Robert. 2017. The present perfect in Nigerian English. *English Language and Linguistics* 21(1): 129–153. <https://doi.org/10.1017/S1360674316000137>
- Wunder, Eva-Maria, Voormann, Holger & Gut, Ulrike. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.

The use of stance markers in West African Englishes

Ulrike Gut and Foluke Unuabonah

University of Münster / Redeemer's University

This chapter investigates the extent of similarity in the use of stance markers in two national varieties of West African English, Nigerian English and Ghanaian English, and compares them to British English. The frequency and stylistic variability of four semantic groups of stance markers were examined in ICE-Nigeria and ICE-Ghana and compared with ICE-Great Britain. The results are mixed: the two West African varieties show an overall lower frequency of stance markers compared to British English but the speakers of the two West African English varieties do not demonstrate lower stylistic variability in the use of stance markers across different text types. Notwithstanding, there are systematic differences in stance marker usage between the two West African English varieties.

1. Introduction

As a consequence of colonialism, West Africa is generally divided along linguistic lines into two major groups: Anglophone and Francophone West Africa (Omoniyi 2008). The Anglophone West African countries include Nigeria, Ghana, Gambia, Sierra Leone and Liberia, and are countries that had contacts with the English language through trade between the 15th and 19th century and through colonisation from the 19th century to the mid-20th century (see Omoniyi 2008; Gut 2017). However, while Ghana and Nigeria are former British administrative colonies, Sierra Leone, Gambia and Liberia mainly served as resettlement areas for freed slaves from Northern America, United Kingdom, the Caribbean islands and slave ships (Gut 2017). Today, these countries share close socio-economic and cultural ties. They belong to the Economic Community of West African States (ECOWAS), which provides free mobility of persons within the region. Thus, there are many citizens of one member state living in other member states; these help to form different social and business relations as well as intermarriages (see Chiluba 2013). Moreover, the five Anglophone West African countries Nigeria, Ghana, Gambia,

Liberia and Sierra Leone are member states of the West Africa Examination Council (WAEC), which holds examinations and awards certificates to students at the end of their secondary school education (see Keteku, Adebayo & Meyers 2008).

In all of the Anglophone West African countries, English is used in a multilingual setting alongside various, sometimes hundreds of indigenous first languages and forms of English pidgins (Dolphyne 1995; Akindele & Adebite 1999; Gut 2017) and enjoys the status of a national or official language. English functions as the language of government, education, law, media, business and commerce, of prestige and social advancement, but with the exception of Liberia where the use of English is much more widespread, it is mainly spoken by the educated minority in all of these countries. Again, with the exception of the Liberian settlers and their descendants, English is a second language (ESL) for most West African speakers and is acquired through formal education, although the numbers of first language speakers, especially in large West African cities, are on the increase (see Jowitt 1991 for Nigerian English). Accordingly, a lectal continuum with different forms of English that are used by different speaker groups exists in the Anglophone West African countries. In this regard, Udofot (2003) identifies three distinct sub-varieties of English in Nigeria that are correlated with the speakers' educational background. Likewise, Conteh-Morgan (1997) identifies three sub-varieties of English in Sierra Leone and Singler (1997) describes three distinct varieties of English in Liberia. The emergence of local standard varieties of English and their documentation differ extensively across the Anglophone West African countries (Gut 2012). Hardly any systematic descriptions exist of the English(es) spoken in Gambia, Liberia and Sierra Leone, while most linguistic features of Nigerian and Ghanaian English have been described and dictionaries (e.g. Dako 2003; Adebite et al. 2014), but no grammars have as yet been published for these varieties of English.

To the best of our knowledge, corpus-based studies of educated varieties of West African Englishes have only just become possible with the publication of ICE Nigeria (Wunder et al. 2010) and the ongoing compilation of ICE Ghana, which represent the variety of university-educated English speakers in these countries. These resources will not only open up new avenues for studying the linguistic properties of those West African Englishes with a novel large-scale empirical method but can also contribute to resolving the question of whether a pan-West African variety of English is emerging in the region. While some scholars have argued for the existence of West African English (WAE) (e.g. Wolf & Igboanusi 2003: 70; Omoniyi 2008), large-scale empirical investigations so far have found very few linguistic features that are indeed peculiar to this area (see e.g. Gut 2017). For instance, Brato and Huber (2012) compared the phonologies and morphosyntactic structures of various African Englishes and found a high similarity between Nigerian English (NigE) and Cameroon English (CamE) and some similarity between those two varieties and

Ghanaian English (GhE). By contrast, Liberian Settler English has distinct phonological and morphosyntactic properties. Compared to other African Englishes, however, only a few phonological and no morphosyntactic features exist that set the West African varieties of English apart. It is thus the overall aim of this study to investigate further the extent of similarity between the West African Englishes NigE and GhE. It will do so by focussing on a linguistic feature hitherto unexplored for these African Englishes: the use of stance markers which are discourse-pragmatic features. As it has been demonstrated by previous scholars (e.g. Schilling-Estes 2006; Schneider 2012; Pichler 2016), a variety is not only identified through lexical, phonological and grammatical features, but also, through discourse-pragmatic preferences.

Stance markers play a vital role in everyday communication where they are used to express personal attitudes, emotions, and judgements about the epistemic status of a proposition (Gray & Biber 2015). Their use has been well studied in native varieties of English (e.g. Biber et al. 1999; Conrad & Biber 2000; Kärkkäinen 2003, 2010; Biber 2006; Englebretson 2007; Aull & Lancaster 2014; Gray & Biber 2015). Previous research has shown that stance markers in English are used more often in spoken than in written text types, both in terms of overall frequency and range of different markers (Biber et al. 1999). Studies of English as a Second Language (ESL) and English as a Foreign Language (EFL) varieties have so far focused mostly on either only written language (e.g. Kärkkäinen 1992; Hyland & Milton 1997; McEnery & Kifle 2002) or only spoken language (Baumgarten & House 2010; Zhang & Sabet 2014; Gablasova & Brezina 2015; Gablasova et al. 2015). Most of these studies have analysed either a few selected stance markers or studied stance in selected text types such as academic discourse and conversations. For example, Gablasova et al. (2015) found that EFL and ESL speakers overuse individual markers while having a smaller repertoire of stance markers, which however expands with increasing proficiency in academic discourse. The study further indicated that there is a lower frequency of stance markers in monologic tasks than in dialogic tasks. Furthermore, the EFL and ESL speakers favour stance verbs and adverbs more than stance adjectives.

The only comparative study of the use of stance markers by native English and ESL/EFL speakers was carried out by Biber and Staples (2014). Based on an analysis of the conversations, interviews and lectures contained in the Hong Kong Corpus of Spoken English, they found that the ESL speakers employ more certainty adverbials, especially in the interviews, and used fewer emphatics in the conversations and lectures compared to the English native speakers. In addition, the ESL speakers show distinct preference patterns for individual stance markers, overusing *actually*, *of course*, *maybe*, and *usually*.

Based on these findings, it is expected in this study that speakers of NigE and GhE, like the EFL and ESL speakers analysed by Gablasova et al. (2015) and Biber

and Staples (2014), will show an overall lower frequency of stance markers compared to British English (BrEng), as well as a restricted range of stance markers and an overuse of some stance markers. Moreover, as found for these non-native speakers, we expect lower stylistic variation of stance markers in NigE and GhE compared to BrEng across different spoken and written text categories. Thus, this study will test the following hypotheses:

- i. speakers of NigE and GhE show an overall lower frequency of stance markers compared to BrEng
- ii. speakers of NigE and GhE employ a restricted inventory of stance markers compared to BrEng
- iii. speakers of NigE and GhE overuse some stance markers compared to BrE speakers
- iv. there is less stylistic variation in the use of stance markers across different types of spoken and written language in NigE and GhE compared to BrE

As a second goal, the study will test differences in the use of stance markers between speakers of NigE and GhE and thus contribute to resolving the question of whether there is a pan-West African variety of English.

2. Stance markers within a variational pragmatic framework

Stance markers are used as both subjective self-expressions, which allow speakers to position themselves towards their propositions, and as intersubjective devices that position speakers towards their interlocutors (Biber et al. 1999; Hunston & Thompson 2000; Bucholtz & Hall 2005; Kärkkäinen 2003, 2006; Bygate 2009; Gablasova et al. 2015). Biber et al. (1999) differentiate between attitudinal, epistemic and style of speaking stance. Epistemic stance deals with the speaker's cognitive understanding of certainty and uncertainty (Kärkkäinen 2010; Zhang & Sabet 2014), attitudinal stance involves the expression of personal emotions and evaluations (Gray & Biber 2015), while style of speaking stance indicates personal comments on the mode of expressing messages. A more functional categorisation is found in Biber and Staples (2014) where stance is divided into stance markers of certainty, uncertainty, attitude and emphasis. Stance is most commonly expressed by first person pronouns and verbal expressions such as *I bet*, *I think*, and *I guess*, adverbs such as *maybe* and *certainly*, and adjectives such as *fortunate* and *surprised*. Other expressions that indicate stance include nouns such as *fact* and *notion*, and modal verbs such as *may* and *will*.

Stance has been examined from different theoretical perspectives. These include grammatical (Hunston & Sinclair 2000), sociolinguistic (Jaffe 2009; Johnstone 2009) and discourse-semantic (Martin & White 2005; White 2011) approaches.

In this study, we examine stance markers from a variational pragmatic approach (Barron & Schneider 2009), where stance markers are seen as subjective and inter-subjective devices which speakers use to position themselves towards their propositions and interlocutors, and which vary based on text type, activity, speaker roles and regional variety (Biber & Staples 2014; Gablasova & Brezina 2015; Gablasova et al. 2015). Variational pragmatics focuses on the systematic investigation of intralingual variation based on the macro-social factors of region, gender, social class, age and ethnicity (Barron & Schneider 2009; Schneider 2012; Barron 2014; Barron et al. 2015). Other factors include discourse types and genre, register, community of practice, speech, writing, and levels of formality (Barron & Schneider 2009). It is an approach at the intersection between pragmatics and dialectology and is based on the principles of empiricity, contrastivity, and comparability (Barron et al. 2015). This implies that variational pragmatic analysis is empirical and must involve the comparison of features between at least two different varieties of a language in order to arrive at variety-specific or variety-preferential linguistic features. The analysis must also involve comparable data sets (Barron et al. 2015). These characteristics make this theory suitable for the analysis of stance markers in West African Englishes: the present study focuses on the systematic investigation of intralingual variation in the use of stance markers, based on the macro-social factors of region and discourse type in order to arrive at variety-specific or variety-preferential linguistic features of West African English. This study meets the requirements of the variational pragmatic framework: it is empirical, employing a corpus-based analysis of stance markers and involving the contrasting of features among three different varieties of English. The data sets are comparable since the three ICE subcorpora used in this study are based on the same design.

3. Data and method

The data were drawn from the Nigerian, Ghanaian,¹ and the British subcorpora of the ICE (International Corpus of English; Greenbaum 1991; <<http://ice-corpora.net/ice/>>) project and included the text categories of private conversations, broadcast interviews and discussions, classroom lessons, broadcast talks and non-broadcast talks, student essays and exams, academic writing and popular writing. Despite the common design of the ICE subcorpora, it cannot be taken for granted that the same type of language is contained in the individual text categories. Classroom lessons, for example, might include quite different types of language in the three countries, as will be discussed below. Table 1 shows the number of word tokens for each text

1. We are extremely grateful to Prof. Magnus Huber for making available to us a pre-publication version of ICE-Ghana.

category in the three subcorpora. By today's standards, these token numbers might be considered very small, however, no larger comparable databases have been collected yet for these text types in these varieties of English.

Table 1. Number of words of the analysed text categories in the three ICE subcorpora

Text type	file name ICE-Gha and ICE-GB	file name ICE-Nig	ICE-Nig	ICE-Gha	ICE-GB
Conversations	S1A 01–90	conv	180,789	213,412	160,747
Broadcast interviews	S1B 41–50	bint	20,356	242,10	20,055
Broadcast discussions	S1B 21–40	bdis	40,290	52,305	41,638
Classroom lessons	S1B 01–20	les	41,394	27,461	35,512
Broadcast talks	S2B 21–30	btal	40,138	15,547	38,832
Non-broadcast talks	S2B 41–50	nbtal	20,156	20,260	15,380
Student essays	W1A 01–20	ess	19,997	21,518	21,933
Student exams	W1A 21–40	ex	20,211	21,194	21,678
Academic writing	W2A 01–80	AHum, ASoc, ATec, ANat	80,045	83,900	85,586
Popular writing	W2B 01–80	PHum, PSoc, PTec, PNat	80,144	84,326	86,285
TOTAL			543,520	479,807	527,646

We searched the corpora with AntConc (freely available at <<http://www.laurenceanthony.net/software/antconc/>>; see Chapter 1.2) for markers which have been shown to exhibit fairly stable meanings (Gablasova et al. 2015). The markers are stated below, and following Biber and Staples (2014), they have been divided into four major semantic and functional groups:

a. Stance markers of certainty

actually, certainly, certain, clear, clearly, definitely, evident that, evidently, for sure, no doubt, obvious that, obviously, predictably, surely, undoubtedly, without doubt, impossible, not possible, I am/I'm certain, I am/I'm confident, I am/I'm convinced, I am/I'm sure, I am/I'm very sure, I know, I bet, I don't doubt

b. Stance markers of doubt/possibility

apparently, doubtful, likely, maybe, perhaps, possible, possibly, probable, probably, unlikely, I am not confident, I am not convinced, I am not sure, I'm not certain, appear(s), seem(s), I assume, I believe, I can't believe, I don't believe, I doubt, I can gather, I gather, I guess, I can't guess, I don't know, I presume, I reckon, I suppose, I suspect, I think, I can think

c. Stance markers of attitude/perspective

arguably, basically, fortunately, unfortunately, frankly, generally, hopefully, ideally, importantly, incredibly, literally, luckily, necessarily, normally, personally, seriously, significantly, surprisingly, theoretically, usually, fortunate, unfortunate, important

to/that, surprised, surprising, essential to/that, crucial to/that, interesting to/that, logical to, necessary to/that, wise to, unwise to, nice to/that, unluckily

d. Emphatic stance markers

absolutely, completely, crucially, exactly, extremely, particularly, really, totally

Manual searches of the data led to the removal of repetitions as in Example (1). Instances where the markers did not have stance meanings were also excluded (see Examples (2) and (3)).

- (1) meanwhile *maybe maybe maybe* when they went to the farm too he didn't weed
uh he didn't weed (ICE-Gha S1A-014.txt)
- (2) I used to wonder how colourful the university can look if the Yoruba lecturer
also *appears* in his own native wear ditto the Idoma and all others
(ICE-Nig nbta1_05.txt)
- (3) yes I'm a Hausa man now they are very much *surprised* with the name they heard
from me they ha- okay I say just forget about this now (ICE-Nig con_14.txt)

4. Results

First, the rate of occurrence of stance markers in the three ICE subcorpora as a whole is presented, followed by the findings on their frequency in the different text types. Table 2 shows the frequency per 1,000 words and the total number of stance markers that occur in ICE-Nig, ICE-Gha and ICE-GB in each of the four semantic groups (see Appendix for the individual stance markers).

Table 2. Frequency per 1,000 words of stance markers in ICE-Nig, ICE-Gha and ICE-GB (absolute number in brackets)

Stance marker types	ICE-Nig	ICE-Gha	ICE-GB
Stance markers of certainty	1.6 (876)	1.3 (715)	2.6 (1,350)
Stance markers of doubt/possibility	2.8 (1,514)	3.4 (1,928)	5.96 (3,141)
Stance markers of attitude/perspective	1.0 (564)	1.3 (748)	1.2 (652)
Emphatic stance markers	1.4 (765)	1.6 (897)	2.55 (1,344)
Total	6.8 (3,719)	7.6 (4,288)	12.3 (6,487)

Table 2 shows that ICE-GB has the highest rate of occurrence of stance markers with a frequency of 12.3 per 1,000 words. This is followed by ICE-Gha and ICE-Nig with 7.6 and 6.8 per 1,000 words respectively. Whilst the frequency of markers of certainty, doubt/possibility and emphatic markers is roughly twice as high in British English compared to the two West African Englishes, there is no difference across the three varieties in terms of frequency of markers of attitude and perspective.

Table 2 further shows that the four groups of markers have different relative frequencies across the three varieties: stance markers of doubt/possibility have the highest frequency among the four semantic groups in each of the ICE corpora. In ICE-GB and ICE-Nig, markers of doubt/possibility are followed by stance markers of certainty while markers of doubt/possibility are followed by emphatic stance markers in ICE-Gha. The opposite trend occurs in the markers that occupy the third position: while emphatic stance markers rank third in ICE-GB and ICE-Nig, stance markers of certainty rank third in ICE-Gha. In the three ICE corpora, stance markers of attitude/perspective have the least rate of occurrence.

Table 3 illustrates the total number of different markers that occur in each of the four semantic groups in the three corpora. It shows that only in the category of stance markers of doubt/possibility do BrE speakers use a larger inventory of different markers compared to speakers of the two West African varieties of English. Markers that occur only in ICE-GB are *I presume*, *I reckon*, *I'm/I am not convinced*, *I'm/I am not confident* and *doubtful*. *I'm/I am not sure*, conversely, is a marker that occurs in both West African varieties of English but not in ICE-GB. Likewise, in the category of stance markers of certainty, the marker *evident that* occurs only in the two West African corpora (albeit very rarely) but not in ICE-GB. Given the relatively small size of the subcorpora as discussed above, these findings cannot be over-interpreted though and they cannot be taken to imply that some markers are absent from a particular variety of English altogether.

Table 3. Total number of different stance markers occurring in the three corpora in each category

Stance marker types	ICE-Nig	ICE-Gha	ICE-GB
Stance markers of certainty	23	21	21
Stance markers of doubt/possibility	22	22	30
Stance markers of attitude/perspective	31	35	35
Emphatic stance markers	7	7	8
Total	83	84	94

Table 4 lists the most frequently used markers in each category for the three varieties of English. It shows that while in all semantic categories the most frequently used marker is shared by all three varieties, distinct differences in patterns of use exist: *certainly* and *obviously* are markers of certainty that are used frequently only in ICE-GB. ICE-Gh and ICE-Nig share a high use of *definitely*, while *I'm/I am sure* occurs frequently only in ICE-Nig. Likewise, in the category of markers of doubt/possibility, *maybe* occurs frequently only in the West African varieties of English. *Probably* is frequent in all three corpora while *seem(s)* is frequent in GhE and BrE only.

Table 4. Four most frequently occurring markers for each category in the three corpora (occurrence per thousand words in brackets)

Stance marker types	ICE-Nig	ICE-Gha	ICE-GB
Stance markers of certainty	actually (0.62)	actually (0.56)	actually (1.3)
	I know (0.46)	I know (0.23)	certainly (0.3)
	definitely (0.14)	definitely (0.1)	obviously (0.27)
	I'm/I am sure (0.06)	clearly (0.07)	I know (0.1)
Stance markers of doubt/possibility	I think (0.97)	I think (1.4)	I think (2.1)
	maybe (0.6)	maybe (0.8)	seem(s), I don't
	I don't know (0.4)	I don't know (0.22)	know (0.58)
	probably (0.16)	seem(s), probably (0.2)	probably (0.5)
Stance markers of attitude/perspective	usually (0.23)	usually (0.3)	usually (0.25)
	generally (0.12)	normally (0.17)	basically (0.13)
	normally (0.1)	generally (0.11)	generally (0.11)
	unfortunately (0.09)	unfortunately (0.07)	necessarily, normally (0.08)
Emphatic stance markers	really (0.97)	really (1.3)	really (1.6)
	exactly (0.15)	completely (0.08)	particularly (0.28)
	particularly (0.13)	exactly, particularly (0.06)	absolutely (0.17)
	completely, absolutely (0.04)		completely (0.13)

Different preference patterns can also be observed in the category of markers of attitude and perspective, where only BrE speakers use *basically* and *necessarily* frequently. The two West African varieties of English share a high use of *unfortunately*. This is also true for the emphatic stance marker *exactly*, which occurs frequently in ICE-Nig and ICE-Gh only. Conversely, *absolutely* occurs frequently in ICE-GB and ICE-Nig only. There are only few stance markers, namely *normally* and *I know*, that have a higher frequency in ICE-Gh and ICE-Nig (and *usually* in ICE-Gh) than in ICE-GB, while *generally* has the same rate of occurrence in the three ICE corpora. All other stance markers occur with a higher frequency in ICE-GB than in the other two corpora. Extracts, showing examples of stance markers, from each of the semantic groups, which occur most frequently in the three corpora are presented in Examples (4) to (15).

Actually

- (4) we ended up in the second world war yes you can *actually* use that competitive revaluation to attack economies that you do not like I think
(ICE-Nig bdis_01.txt)
- (5) <\$B> M A G R A E you know <\$A> yea I *actually* don't know much about him
<\$B> Magrae <\$A> but I read something that he wrote (ICE-Gha S2B-043.txt)
- (6) I presume this is the first <ICE-GB:S1A-002 #122:2:B> Well I'm *actually* involved in an integrated youth group uhm called Unity (ICE-GB S1A-002)

I think

- (7) though I fear that would take all our time here today but *I think* it's a matter of style perhaps another friend wouldn't have written as much as was (ICE-Nig btr_03.txt)
- (8) He doesn't need an introduction and he warned me to be brief but *I think* that this is my opportunity and uh if I say it what i like it (ICE-Gha S2B-045.txt)
- (9) <ICE-GB:S1A-023 #260:1:B> *I think* it's probably academic <ICE-GB:S1A-023 #261:1:B> Mm yes I can't see it now (ICE-GB:S1A-023)

Usually

- (10) even when they have similar solubilities in a particular solvent. *Usually* a glass column about 30 cm long and 2 cm in diameter, which narrows (ICE-Nig ATec_07.txt)
- (11) so you want the protein in a folding competent confirmation and *usually* it is the <heat shock?> proteins that help this to go on (ICE-Gha S2A-022.txt)
- (12) uhm I *usually* skip breakfast and <,> uhm travel on cups of coffee <,> or tea <,> <ICE-GB:S1A-059 #49:1:A> Mhm (ICE-GB:S1A-059)

Really

- (13) they would have looked checked and seen that the real voters *really* should not be more than this number so this additions they put are not *really* voters (ICE-Nig bdis_17.txt)
- (14) you very much uh Mister Forson for this uh interesting summary and I must say I *really* enjoyed reading your thesis gave me a lot of insight (ICE-Gha S2B-044.txt)
- (15) Having said that <,> uhm <,> I can *really* only say how it was for me when I came to work (ICE-GB:S1A-001)

Table 5 presents the frequency of the different word classes of stance markers that occur in the three corpora. It shows that in all three varieties, stance adverbials have a higher frequency than verbs, which in turn have a higher frequency than adjectives. While the frequency of adverbial and verbal stance markers is roughly twice as high in BrE compared to the two West African English varieties, the frequencies of adjectives in the three corpora are nearly the same.

Table 5. Frequency (per thousand words) of the different word classes of stance markers in the three corpora

Word class	ICE-Nig	ICE-Gha	ICE-GB
Adverbials	4.08	4.72	7.2
Verbs	2.2	2.41	4.13
Adjectives	0.56	0.65	0.69

Use of stance markers in different text types

Table 6 shows the frequency of stance markers across the different text types. For all three varieties of English, conversations have the highest rate of occurrence of stance markers. Furthermore, free speech in broadcast interviews and discussions as well as class lessons has a higher frequency of stance markers than scripted speech (broadcast and non-broadcast talks) and written language in all three corpora. Differences between BrE on the one hand and the two West African Englishes on the other exist in terms of the range of stylistic variation, which is much higher in the former (from 19.9 per thousand words in the conversations to 4.7 in the broadcast talks) than the latter (11.4 and 11.7 respectively in the conversations to 3 and 3.1 respectively in the student essays and exams). Moreover, a striking difference between BrE and the two West African varieties of English lies within the different types of free speech: while in BrE stance markers are used with nearly equal rate in conversations, broadcast discussions and classroom lessons, in GhE and NigE broadcast speech has fewer stance markers than conversations, and classroom lessons show an even lower rate. Another striking difference is that student essays and examinations have the lowest frequency of stance markers in the two West African English varieties, while broadcast talks have the lowest frequency of stance markers in BrE.

Table 6. Frequency (per 1,000 words) of stance markers in the different ICE text types

Text categories	ICE-Nig	ICE-Gha	ICE-GB
conversations	11.4	11.7	19.9
broadcast interviews	9.4	9.1	13.2
broadcast discussions	9.1	8.9	19
classroom lessons	5.8	7.5	19.6
broadcast talks	3.8	3.3	4.7
popular writing	3.4	3.5	6.1
academic writings	3.1	3.7	6.4
non-broadcast talks	3.2	4.8	6.3
Student essays + examinations	3	3.1	6.7

Stylistic differences in the use of stance markers across all three varieties of English become apparent in Figures 1a to c, which show the frequency of the different semantic categories of stance markers across the different analysed text types in the three subcorpora. Only in BrE do markers of doubt have the highest rate in all text types (Figure 1a). In GhE, they are surpassed by attitude markers in student essays and exams as well as in academic and popular writing (Figure 1b). In NigE, markers of doubt have a lower rate of occurrence than attitude markers in academic and popular writing and a lower rate than both certainty and emphatic markers in broadcast interviews (Figure 1c).

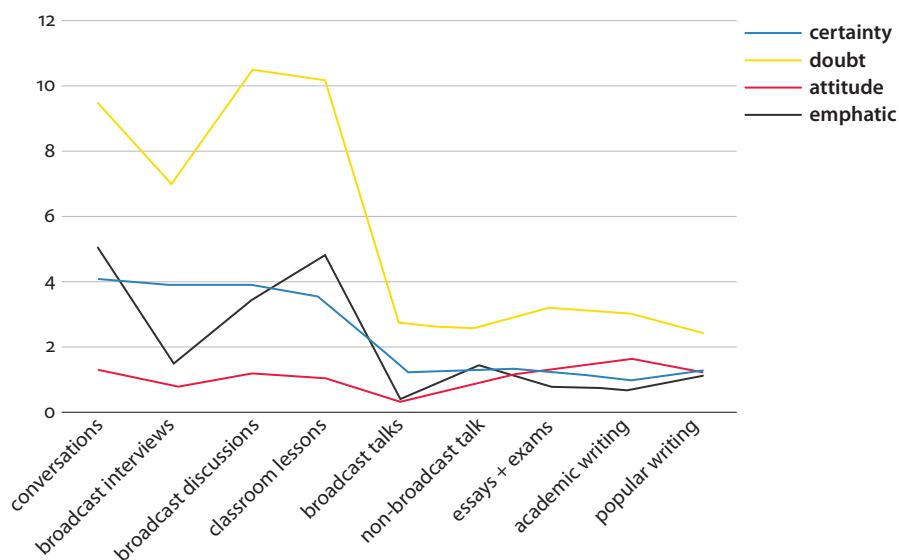


Figure 1a. Frequency of markers of certainty, doubt, attitude and emphatic markers across the analysed text categories in ICE-GB

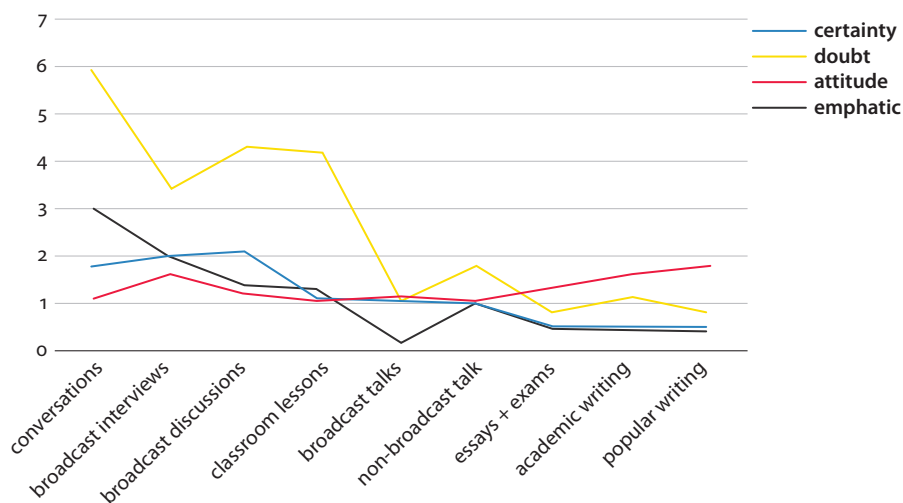


Figure 1b. Frequency of markers of certainty, doubt, attitude and emphatic markers across the analysed text categories in ICE-Gha

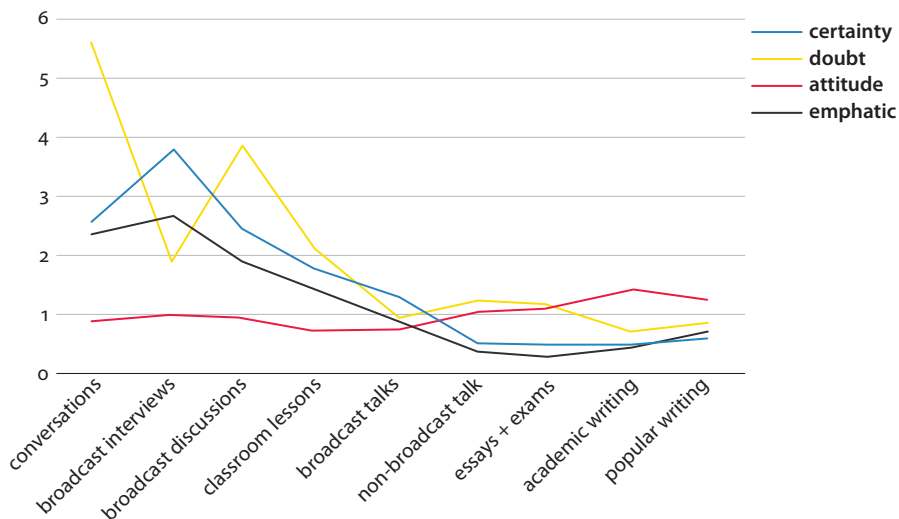


Figure 1c. Frequency of markers of certainty, doubt, attitude and emphatic markers across the analysed text categories in ICE-Nig

Moreover, in BrE, markers of certainty show a distinct drop in frequency between free speech on the one hand and scripted speech and written language on the other. For the two West African English varieties, only the spoken text categories conversations and broadcast speech have a higher rate of certainty markers but classroom lessons group together with scripted speech and written language. By the same token, emphatic markers have a high frequency in BrE in conversations, broadcast discussions and class lessons, while in GhE and NigE they occur frequently in conversations and broadcast interviews only. The three varieties of English also share some usage patterns: markers of attitude are used more often in text types of scripted speech or written language than in free speech and these markers of attitude have a very low frequency in broadcast interviews across the three ICE-corpora. Moreover, one of the features that GhE shares with BrE is that markers of doubt are used more frequently than other stance marker types in broadcast interviews while a similar pattern of usage is found in BrE and GhE conversations: markers of doubt occur more frequently than markers of emphasis which are then followed by markers of certainty and attitude.

5. Summary and discussion

The objective of this study was to discover patterns of stance marker use in Nigerian, Ghanaian and British English. Our first prediction that speakers of NigE and GhE show an overall lower frequency of stance markers compared to BrEng can be mostly confirmed. When taken together, stance markers are used nearly twice as frequently in BrE than in the two West African varieties of English; however, when looking at the individual semantic types of stance markers, it becomes apparent that only markers of certainty, doubt/possibility and emphatic markers are used more frequently in BrE while stance markers of attitude and perspective have the same rate of occurrence in all three varieties.

Our second hypothesis, that speakers of NigE and GhE employ a restricted inventory of stance markers compared to BrEng, can only be partly confirmed. On the whole, slightly more different stance markers occur in ICE-GB compared to the other two corpora – although many of those have an absolute frequency of below five, which makes comparisons very difficult – but when analysing the different semantic categories, it was found that the inventory of stance markers was only smaller for the two West African varieties in terms of stance markers of doubt and possibility. In the other three categories, no differences were found. The West African English speakers thus do not show the same usage patterns as the L2 speakers analysed by Gablasova et al. (2015). Furthermore, they use more adverbials than verbs in line with their BrE counterparts. This is slightly different from the findings in Gablasova et al. (2015), where the non-native English speakers used more verbs than adverbials.

Hypothesis three predicted that speakers of NigE and GhE overuse some stance markers. In absolute terms, this is only true for the stance marker *usually* in GhE as well as for *normally* and *I know* in both West African varieties of English. All other markers that occur with a frequency of at least $n = 10$ (to avoid misinterpretations due to low numbers) are more frequent in ICE-GB. The West African English speakers thus do not show overuse patterns like the Hong Kong English speakers that were analysed by Biber and Staples (2014). Relatively speaking, however, some preference patterns become evident: the two West African varieties of English have a relatively higher use of *unfortunately* and *exactly* compared to ICE-GB, while they disprefer, relatively speaking, the use of *basically* and *necessarily*. Furthermore, *absolutely* is dispreferred, relatively speaking, in GhE. The preference for some markers such as *maybe*, *I'm/ I am sure* and *I'm/I am not sure* corroborates the findings of Gablasova et al. (2015), in which the L2 speakers had a high preference for these markers in signalling certainty and uncertainty. Also, the preference for some markers such as *maybe*, *actually* and *usually* has also been indicated for Hong Kong ESL speakers who have a preference for these markers (Biber & Staples 2014). The tendency of West African English speakers to use these

markers frequently may be as a result of their multifunctional roles in the different text types. For example, Gablasova et al. (2015) opine that *maybe* functions as a hedging device in face-threatening situations such as questions, suggestions and comments. These functions have been found also for the two varieties of West African English, where *maybe* performs other secondary functions apart from indicating uncertainty. These are illustrated in Examples (16) to (22).

Maybe serving as a hedging device in suggestions and directives:

- (16) so *maybe* I will suggest that you cross-check from the deans before they leave the authentic list of members of their faculties (ICE-Nig btr_10.txt)
- (17) I I I don't want to go into that Uhm <\$A><ICE-GH:S1A-003#39:1:A> Yeah but *maybe* you want to say something about your research <ICE-GH:S1A-003#40:1:A> That would give us something to talk about <\$B><(ICE-Gha S1A-003.txt)

Maybe functioning as a face-saving device:

- (18) before I continue it is quite possible that *maybe* a few of you don't have the minutes they're producing more kindly bear with us more are coming it's my (ICE-Nig btr_02.txt)
- (19) <\$B> That's what <? Magdalena?> *Maybe* when I have credit I'll just call her to say hi <\$A> You'll call her (ICE-Gha S1A-010-D.txt)
- (20) <\$C> yeah I prefer it that way <\$B> that is your style right <\$C> I prefer it that way <\$B> uhm ok *maybe* that will make you happy <\$C> (ICE-Gha S1A-039.txt)

Maybe indicating repair/correction:

- (21) they represent the interest of a group of people *maybe* communities for the purpose of protecting their interest in any development project (ICE-Nig les_01.txt)
- (22) <\$A> yesterday was a major class <\$B> yes <\$A> anybody anybody or *maybe* somebody in a green green green kind of what shade green anybody (ICE-Gha S1A-011.txt)

The preference for *unfortunately* (which deals with negative assessment) after the close synonyms *usually*, *generally* and *normally* (which deal with normality; see Martin & White 2005) may indicate a strong cultural need on the part of the speakers of the West African Englishes to express their stance in the face of an unwanted outcome. In other words, the speakers or writers have to explicitly indicate that they share or sympathise with their interlocutors or the society when there is an unpleasant situation as indicated in Examples (23) and (24).

- (23) Mallam Ribadu erm was invited to come and help the commission with certain information and the stuff like that but *unfortunately* that effort was erm sort of being politicised and all the stuff like that and erm eventually he he directed
(ICE-Nig bint_01.txt)
- (24) can't you just hook me up with a nice girl <\$D> i know a number of nice girls <\$A> oh really <\$D> *unfortunately* they are all they are all in level four hundred
(ICE-Gha S1A-022.txt)

Unfortunately may also be used by the speaker to explicitly relate his or her stance towards an unpleasant incidence in order to invite the sympathy of the interlocutor as cited in Examples (25) and (26).

- (25) thinking that I had a new garment so I don't need to use the old garment again there is a new wineskin but *unfortunately* I lost the P C and I couldn't tell my father for months
(ICE-Nig con_10.txt)
- (26) the Togolese team were not victims they volunteered for the disaster it sounds callous I know to say that but *unfortunately* it is the sad truth four time boxing
(ICE Gha S2B-028.txt)

Unfortunately may be used by the speaker to express an apology as shown in Examples (27) and (28).

- (27) see you're unlucky why that you won't get this from AGGN oh hmm next time we'll make it from mid ANON *unfortunately* I don't celebrate birthday so I don't agree I don't say that to hard men like you I only say to
(ICE-Nig con_16.txt)
- (28) ah did you go to church last Sunday <\$A> ah *unfortunately* I couldn't <\$B> what what happened <\$A> Uh we are undertaking a project at a friend place
(ICE-Gha S1A-008.txt)

The fourth hypothesis, which predicted that there is less stylistic variation in the use of stance markers across different types of spoken and written language in NigE and GhE compared to BrE, is not supported by our data. Although the absolute difference between frequency rates of stance marker usage is greater in BrE than in the two West African varieties of English, all three Englishes show distinct stylistic variation across the different text types. What they have in common is a clear drop of stance marker usage between the different types of free speech on the one hand and scripted speech and written language on the other. This is especially apparent in the markers of attitude. Differences between the three varieties exist in terms of usage of markers of doubt, which form the largest group of stance markers in all text types in BrE but not in GhE and NigE. The low rate of occurrence of markers of doubt in academic and popular writing in GhE and NigE may suggest that West African English speakers are reluctant to express doubt in writing, and in

particular for NigE speakers, to express doubt in interviews where speakers need to convince their addressees of the speakers' positions on issues. Hence, stance markers of certainty are used more frequently than other types of stance markers in broadcast interviews which can be compared to Hong Kong ESL speakers, who also use more certainty adverbials in interviews (Biber & Staples 2014). Also, the lower rate of occurrence of stance markers in broadcast interviews, discussion and classroom lessons when compared to conversations, which are all different types of free speech, in the two West African English varieties may indicate that NigE and GhE speakers prefer to explicitly express their stance in private than in public speech. One type of speech that stands out as very different in terms of stance marker usage in the two West African varieties of English are class lessons that show far fewer emphatic and certainty markers than class lessons in BrE. This also corroborates some of the findings in Biber and Staples (2014), where Hong Kong ESL speakers employed fewer emphatics in lectures, which in a way are similar to classroom lessons. On the other hand, these findings raise the question of whether the individual text types represent the same registers across the different ICE sub-corpora. Future research on grammatical and lexical features of these text types might confirm that language use in these text types differs systematically across the varieties of English and thus reflect register differences.

Concerning the second goal of the study, the results have contributed mixed results to the question of whether there is a pan-West African variety of English. Some commonalities between the two varieties that set them apart from BrE indeed exist in terms of preference patterns and the use of stance markers across different text types. However, systematic differences in stance marker usage between the two West African varieties of English were also found. The present study has limited itself to two West African English varieties and these may not be adequate to answer the question of whether there is indeed a pan-West African variety of English. More comparative investigations need to be conducted, not only in these two varieties, but also in other varieties such as Gambian English, Liberian English and Sierra Leonean English. Such comparative investigations on stance marker usage may cover the collocational patterns of these markers and their discourse-pragmatic functions in a possible pan-West African variety of English. Furthermore, the study of stylistic variability in this paper has been restricted to selected text types since ICE-Gh is still under construction. More findings from other text types such as broadcast news, cross-examination, parliamentary debates, editorial and social letters will further help to answer the question of the existence of such a variety of English. As it has been noted elsewhere, studies on linguistic variation have tended to focus on phonological, lexical and grammatical variation with less attention on discourse-pragmatic features (Schilling-Estes 2006; Schneider 2012; Pichler 2016), which may be as a result of the unfixed nature of discourse-pragmatic variables. However, as it has been earlier mentioned, discourse-pragmatic features such as

stance markers and discourse markers can be used to identify varieties of a language, as they are dependent on context and social meaning (Barron & Schneider 2009; Aijmer 2013; Pichler 2016). Hence, they assist in distinguishing one variety of a language from another. This may explain the recent increase in the number of studies on discourse-pragmatic variation in, for example, English (see Aijmer 2013; Biber & Staples 2014; Zhang & Sabet 2014). Notwithstanding, more comparative lexical and grammatical studies are equally required in order to establish the possible existence of WAE as a variety in its own right.

6. Conclusion

In conclusion, we have identified the fact that while the study of stance markers is prevalent for native varieties of English, there is very limited comprehensive work on the study of stance in ESL varieties, and in particular, West African Englishes. Hence in this study, we have investigated the level of similarity in the use of stance markers in two national varieties of West African English: NigE and GhE and compared the results with a native variety of English: BrE. The analysis presents a mixed bag of results: the two varieties show an overall lower frequency of stance markers compared to BrE but only make use of a restricted range of markers of doubt when compared to BrE. The speakers of the two West African English varieties do not overuse any stance markers in particular; however, they show distinct preference patterns for certain stance markers compared to BrE. Moreover, the speakers of the two West African English varieties do not demonstrate lower stylistic variability in the use of stance markers across different text types of spoken and written English. Notwithstanding, systematic differences in stance marker usage between the two West African English varieties were also found.

As recognised in previous work (Romero-Trillo 2008; Rühlemann 2010; Rühlemann & Aijmer 2015), there are challenges in the investigation of discourse-pragmatic features such as stance markers, using corpus methods due to the fact that these markers have multifunctional roles which are dependent on the context of situation. So far, ways of tackling these challenges have either focused on a form-to-function or a function-to-form investigation (Rühlemann & Aijmer 2015). In a form-to-function examination, the researcher pre-selects items (which have been found to have discourse-pragmatic functions) to be searched in the corpora and examines their functions using concordance lines. This, then, rules out other forms in which stance may be marked through other means than the pre-selected items. In a function-to-form investigation, the researcher identifies a function and investigates its forms in the corpora. This also has its limitations as functions cannot be retrieved but only identified through surface forms, i.e. in the case of speech acts (Rühlemann & Aijmer 2015). A method that can help address the stated challenges

is the pragmatic annotation of corpora (see Weisser 2015) which has been applied to a small number of corpora (Rühlemann & O'Donnell 2012; Danescu-Niculescu et al. 2013). This method is not widely used yet because the automatic assignment of tags often leads to imprecision, which leads to a manual tagging of items that can be time and resource consuming. Hence, further work on current semi-automatic pragmatic annotation is required in order to allow for a more efficient and precise way of annotating pragmatic phenomena (Rühlemann & Aijmer 2015; Weisser 2015). Such automatic pragmatic annotation can then be applied to corpora of African Englishes in order to identify variations in the use of discourse-pragmatic phenomena such as stance markers in African varieties of English.

Acknowledgements

This work is supported by the Alexander von Humboldt Foundation.

References

- Adegbite, Adewale, Udofot, Inyang & Ayoola, Kehinde. 2014. *A Dictionary of Nigerian English*. Ile-Ife: Obafemi Awolowo University.
- Aijmer, Karin. 2013. *Understanding Pragmatic Markers: A Variational Pragmatic Approach*. Edinburgh: EUP.
- Akindele, Femi & Adegbite, Wale. 1999. *The Sociology and Politics of English in Nigeria: An Introduction*. Ile-Ife: Obafemi Awolowo University Press.
- Aull, Laura L. & Lancaster, Zak. 2014. Linguistic markers of stance in early and advanced academic writings: A corpus-based comparison. *Written Communication* 32(2):151–183. <https://doi.org/10.1177/0741088314527055>
- Barron, Anne. 2014. Variational pragmatics. In *The Encyclopedia of Applied Linguistics*, Carol A. Chapelle (ed.), 1–7. Malden MA: Blackwell. <https://doi.org/10.1002/9781405198431.wbeal1429>
- Barron, Anne, Pandarova, Irina & Muderack, Karoline. 2015. Tag questions across Irish English and British English: A corpus analysis of form and function. *Multilingua* 34(4): 495–525.
- Barron, Anne & Schneider, Klaus P. 2009. Variational pragmatics: Studying the impact of social factors on language use in interaction. *Intercultural Pragmatics* 6: 425–442. <https://doi.org/10.1515/IPRG.2009.023>
- Baumgarten, Nicole & House, Juliane. 2010. 'I think' and 'I don't know' in English as lingua franca and native English discourse. *Journal of Pragmatics* 42(5): 1184–1200. <https://doi.org/10.1016/j.pragma.2009.09.018>
- Biber, Douglas. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5: 97–116. <https://doi.org/10.1016/j.jeap.2006.05.001>
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, Finegan, Edward & Quirk, Randolph. 1999. *Longman Grammar of Spoken and Written English*. Cambridge MA: The MIT Press.

- Biber, Douglas & Staples, Shelley. 2014. Variation in the realisation of stance adverbials. In *Spoken Corpora and Linguistic Studies* [Studies in Corpus Linguistics 61], Tommaso Raso & Heliana Mello (eds), 271–294. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.61.10bib>
- Brato, Thorsten & Huber, Magnus. 2012. English in Africa. In *Areal Features of the Anglophone World*, Raymond Hickey (ed.), 161–185. Berlin: Mouton de Gruyter.
- Bucholtz, Mary & Hall, Kira. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies* 7(4–5): 585–614. <https://doi.org/10.1177/1461445605054407>
- Bygate, Martin. 2009. Teaching and testing speaking. In *The Handbook of Language Teaching*, Michael H. Long & Catherine J. Doughty (eds), 412–440. Chichester: Wiley-Blackwell. <https://doi.org/10.1002/9781444315783.ch23>
- Chiluwa, Innocent. 2013. West African English in digital discourse. *Covenant Journal of Language Studies* 1(1): 42–61.
- Conrad, Susan & Biber, Douglas. 2000. Adverbial marking of stance in speech and writing. In *Evaluation in Text: Authorial stance and the Construction of Discourse*, Susan Hunston & Geoffrey Thompson (eds), 56–73. Oxford: OUP.
- Conteh-Morgan, Miriam. 1997. English in Sierra Leone: A description of the language as used in this West African Country and a consideration of its status there. *English Today* 13: 52–56. <https://doi.org/10.1017/S026607840000986X>
- Dako, Kari. 2003. *Ghanaianisms – A Glossary*. Accra: Ghana University Press.
- Danescu-Niculescu-Mizil, Cristian, Sudhof, Moritz, Jurafsky, Dan, Leskovec, Jure & Potts, Christopher. 2013. *A Computational Approach to Politeness with Application to Social Factors*. <<https://nlp.stanford.edu/pubs/politeness.pdf>> (16 October 2017).
- Dolphyne, Florence. 1995. A note on the English language in Ghana. In *New Englishes: A West African Perspective*, Ayo Bamgbose, Ayo Banjo & Andrew Thomas (eds), 27–33. Ibadan: Mosuro.
- Englebretson, Robert. 2007. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction* [Pragmatics & Beyond New Series 164]. Amsterdam: John Benjamins. <https://doi.org/10.1075/pbns.164>
- Gablasova, Dana & Brezina, Vaclav. 2015. Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In *Yearbook of Corpus Linguistics and Pragmatics*, Jesús Romero-Trillo (ed.), 117–136. Dordrecht: Springer.
- Gablasova, Dana, Brezina, Vaclav, McEnery, Tony & Boyd, Elaine. 2015. Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*: 1–26. <https://doi.org/10.1093/applin/amv055>.
- Gray, Bethany & Biber, Douglas. 2015. Stance markers. In *Corpus Pragmatics: A Handbook*, Karin Aijmer & Christoph Rühlemann (eds), 219–248. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139057493.012>
- Greenbaum, Sidney. 1991. ICE: The International Corpus of English. *English Today* 7(4): 3–7. <https://doi.org/10.1017/S0266078400005836>
- Gut, Ulrike. 2012. Standards of English in West Africa. In *Standards of English: Codified Varieties around the World*, Raymond Hickey (ed.), 213–228. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139023832.012>
- Gut, Ulrike. 2017. English in West Africa. In *The Oxford Handbook of World Englishes*, Markku Filppula, Juhani Klemola & Devyani Sharma (eds), 491–507. Oxford: OUP.
- Hunston, Susan & Sinclair, John. 2000. A local grammar of evaluation. In *Evaluation in Text: Authorial stance and the Construction of Discourse*, Susan Hunston & Geoffrey Thompson (eds), 74–101. Oxford: OUP.
- Hunston, Susan & Thompson, Geoffrey. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: OUP.

- Hyland, Ken & Milton, John. 1997. Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6(2): 183–205.
[https://doi.org/10.1016/S1060-3743\(97\)90033-3](https://doi.org/10.1016/S1060-3743(97)90033-3)
- Jaffe, Alexandra. 2009. Introduction: The sociolinguistics of stance. In *Stance: Sociolinguistic Perspectives*, Alexandra Jaffe (ed.), 3–28. Oxford: OUP.
<https://doi.org/10.1093/acprof:oso/9780195331646.001.0001>
- Johnstone, Barbara. 2009. Stance, style and the linguistic individual. In *Stance: Sociolinguistic Perspectives*, Alexandra Jaffe (ed.), 29–53. Oxford: OUP.
<https://doi.org/10.1093/acprof:oso/9780195331646.003.0002>
- Jowitt, David. 1991. *Nigerian English Usage: An Introduction*. Lagos: Longman.
- Kärkkäinen, Elise. 1992. Modality as a strategy in interaction: Epistemic modality in the language of native and non-native speakers of English. *Pragmatics and Language Learning* 3: 197–216.
- Kärkkäinen, Elise. 2003. *Epistemic Stance in English Conversation: A Description of its Interactional Functions, with a Focus on I Think* [Pragmatics & Beyond New Series 115]. Amsterdam: John Benjamins. <https://doi.org/10.1075/pbns.115>
- Kärkkäinen, Elise. 2006. Stance taking in conversation: From subjectivity to intersubjectivity. *Text & Talk* 26(6): 699–731. <https://doi.org/10.1515/TEXT.2006.029>
- Kärkkäinen, Elise. 2010. Position and scope of epistemic phrases in planned and unplanned American speech. In *New Approaches to Hedging*, Gunther Kaltenböck, Wiltrud Mihatsch & Stefan Schneider (eds), 203–236. Bingley: Emerald.
https://doi.org/10.1163/9789004253247_011
- Keteku, Nancy W., Adebayo, Folashade & Meyers, James P. 2008. Anglophone West Africa: Education, Credentials, Recruitment & Resources. Paper presented at NAFSA Conference, Washington DC. <<https://shelbycearley.files.wordpress.com/2010/06/anglophone-west-africa-education-credentials-recruitment-resources.pdf>> (8 January 2017).
- Martin, James R. & White, Peter R. R. 2005. *The Language of Evaluation: Appraisal in English*. New York NY: Palgrave Macmillan. <https://doi.org/10.1057/9780230511910>
- McEnery, Tony & Kifle, Nazareth A. 2002. Epistemic modality in argumentative essays of second-language writers. In *Academic Discourse*, John Flowerdew (ed.), 182–195. Harlow: Longman.
- Omoniyi, Tope. 2008. West African Englishes. In *Handbook of World Englishes*, Braj Kachru, Yumuna Kachru & Cecil Nelson (eds), 172–187. London: Wiley.
- Pichler, Heike. 2016. Introduction: Discourse-pragmatic variation and change. In *Discourse-Pragmatic Variation and Change: Insights from English*, Heike Pichler (ed.), 1–18. Cambridge: CUP. <https://doi.org/10.1017/CBO9781107295476>
- Romero-Trillo, Jesus. 2008. Introduction: Pragmatics and corpus linguistics. A mutualistic entente. In *Pragmatics and Corpus Linguistics. A Mutualistic Entente*, Jesus Romero-Trillo (ed.), 1–10. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110199024>
- Rühlemann, Christoph. 2010. What can a corpus tell us about pragmatics? In *Routledge Handbook of Corpus Linguistics*, Anne O'Keeffe & Michael McCarthy (eds), 288–301. London: Routledge.
- Rühlemann, Christoph & Aijmer, Karin. 2015. Corpus pragmatics: Laying the foundations. In *Corpus Pragmatics: A Handbook*, Karin Aijmer & Christoph Rühlemann (eds), 1–28. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139057493>
- Rühlemann, Christoph & O'Donnell, Brook. 2012. Introducing a corpus of conversational stories: Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory* 8(2): 313–350. <https://doi.org/10.1515/cllt-2012-0015>

- Schilling-Estes, Natalie. 2006. Dialect variation. In *An Introduction to Language and Linguistics*, Ralph W. Fasold & Jeff Connor-Linton (eds), 311–342. Cambridge: CUP.
- Schneider, Klaus P. 2012. Appropriate behaviour across varieties of English. *Journal of Pragmatics* 44(9): 1022–1037. <https://doi.org/10.1016/j.pragma.2011.09.015>
- Singler, John. 1997. The configuration of Liberia's Englishes. *World Englishes* 16: 205–231. <https://doi.org/10.1111/1467-971X.00060>
- Udofot, Inyang. 2003. Stress and rhythm in the Nigerian accent of English. *English World Wide* 24: 201–220. <https://doi.org/10.1075/eww.24.2.04udo>
- Weisser, Martin. 2015. Speech act annotation. In *Corpus Pragmatics: A Handbook*, Karin Aijmer & Christoph Rühlemann (eds), 84–116. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139057493.005>
- White, Peter R. R. 2011. Appraisal. In *Discursive Pragmatics* [Handbook of Pragmatics Highlights 8], Jan Zienkowski, Jan-Ola Östman & Jeff Verschueren (eds), 14–36. Amsterdam: John Benjamins. <https://doi.org/10.1075/hoph.8.02whi>
- Wolf, Hans-Georg & Igboanusi, Herbert. 2003. A preliminary comparison of some lexical items in Nigerian English and Cameroon English. In *Studies in African Varieties of English*, Peter Lucko, Lothar Peter and Hans-Georg Wolf (eds), 69–81. Bern: Peter Lang.
- Wunder, Eva-Maria, Voormann, Holger & Gut, Ulrike. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.
- Zhang, Grace Q. & Sabet, Peyman G. 2014. Elastic 'I think': Stretching over L1 and L2. *Applied Linguistics* 37(3): 334–353. <https://doi.org/10.1093/applin/amu020>.

Appendix

Total number of different stance markers of certainty in ICE-Nig, ICE-Gha and ICE-GB

Stance markers of certainty	ICE-Nig	ICE-Gha	ICE-GB
actually	0.62 (338)	0.56 (317)	1.3 (668)
certainly	0.04 (21)	0.06 (33)	0.3 (182)
clearly	0.04 (23)	0.07 (38)	0.2 (83)
definitely	0.14 (76)	0.1 (58)	0.1 (53)
evidently	0.004 (2)	0.002 (1)	0.006 (3)
evident that	0.004 (2)	0.01 (3)	0
for sure	0.006 (3)	0.01(8)	0.009 (5)
no doubt	0.04 (23)	0.003 (2)	0.06 (30)
obviously	0.04 (19)	0.06 (32)	0.28 (145)
obvious that	0.01(6)	0.01(6)	0.006 (3)
predictably	0	0	0.004 (2)
surely	0.03 (15)	0.02 (11)	0.07 (37)
undoubtedly	0.006 (3)	0.01(6)	0.03 (18)
without doubt	0.004 (2)	0	0.006 (3)
impossible	0.04 (23)	0.02 (11)	0.06 (31)
not impossible	0.01 (8)	0.01 (8)	0.01 (6)
I am/I'm certain	0	0.002 (1)	0
I am/I'm confident	0.009 (5)	0.002 (1)	0.02 (1)

Stance markers of certainty	ICE-Nig	ICE-Gha	ICE-GB
I am/I'm convinced	0.002 (1)	0	0
I am/I'm sure	0.06 (32)	0.004 (2)	0.01 (5)
I am/I'm very sure	0.009 (5)	0.05 (26)	0
I know	0.46 (251)	0.24 (134)	0.1 (56)
I (can) bet	0.006 (3)	0.008 (4)	0.02 (8)
I don't doubt	0	0	0

Total number of different stance markers of doubt/possibility in ICE-Nig, ICE-Gha and ICE-GB

Stance markers of doubt/possibility	ICE-Nig	ICE-Gha	ICE-GB
appear(s)	0.07 (36)	0.07 (41)	0.2 (127)
apparently	0.02 (11)	0.02 (12)	0.6 (309)
doubtful	0	0	0.004 (2)
likely	0.05 (29)	0.1 (63)	0.2 (114)
maybe	0.59 (322)	0.8 (461)	0.2 (116)
perhaps	0.08 (45)	0.37 (193)	0.4 (193)
possible	0.09 (51)	0.07 (40)	0.3 (171)
possibly	0.02 (10)	0.007 (4)	0.13 (68)
probable	0.002 (1)	0	0.008 (4)
probably	0.16 (85)	0.2 (111)	0.51 (270)
seem(s)	0.12 (67)	0.2 (113)	0.59 (309)
unlikely	0.004 (2)	0.004 (2)	0.06 (34)
I assume	0.002 (1)	0.002 (1)	0.01(6)
I believe	0.14 (74)	0.1 (55)	0.1 (51)
I can't believe	0.002 (1)	0.02 (10)	0.009 (5)
I can gather	0	0	0.006 (3)
I don't believe	0.006 (3)	0.004 (2)	0.01(7)
I am/I'm not certain	0	0.002(1)	0
I am/I'm not confident	0	0	0.002 (1)
I am/I'm not convinced	0	0	0.004 (2)
I am/I'm not sure	0.03 (14)	0.03 (15)	0
I don't know	0.4 (218)	0.23 (128)	0.58 (305)
I doubt	0.006 (3)	0.006 (3)	0.02 (9)
I gather	0.002 (1)	0	0.004 (2)
I can gather	0	0	0.006 (3)
I can't guess	0	0	0.002 (1)
I guess	0.02 (12)	0.04 (22)	0.004 (19)
I presume	0	0	0.02 (9)
I reckon	0	0	0.002 (9)
I suppose	0.002 (1)	0.002 (1)	0.25 (133)
I suspect	0	0.004 (2)	0.02 (8)
I think	1.0 (527)	1.4 (796)	2.1 (1112)
I can think	0	0	0.002 (1)

Total number of different stance markers of attitude/perspective in ICE-Nig, ICE-Gha and ICE-GB

Stance markers of attitude/perspective	ICE-Nig	ICE-Gha	ICE-GB
arguably	0	0.004 (2)	0.009 (5)
basically	0.06 (32)	0.08 (46)	0.14 (72)
crucial to/that	0	0.006(3)	0.002 (1)
essential to/that	0.01(6)	0.009(5)	0.01(6)
fortunate	0.002 (1)	0.004 (2)	0.009 (5)
fortunately	0.006 (3)	0.01 (7)	0.02 (9)
frankly	0.01 (6)	0.006 (3)	0.02 (8)
generally	0.12 (66)	0.11 (63)	0.12 (61)
hopefully	0.004 (2)	0.02 (13)	0.03 (14)
ideally	0.002 (1)	0.002 (1)	0.02 (8)
importantly	0.01 (8)	0.01 (7)	0.02 (10)
important to/that	0.06 (35)	0.08 (46)	0.07 (36)
incredibly	0	0.002 (1)	0.02 (12)
interesting to/that	0.02 (10)	0.03 (17)	0.02 (12)
literally	0.009 (5)	0.009 (5)	0.02 (13)
logical to	0	0	0.004 (2)
luckily	0.002 (1)	0.005 (3)	0.004 (2)
necessarily	0.02 (12)	0.06 (31)	0.09 (45)
necessary to/that	0.02 (13)	0.03 (19)	0.06 (31)
nice to/that	0.006 (3)	0.07 (4)	0.02 (12)
normally	0.1 (57)	0.17 (96)	0.09 (45)
personally	0.04 (21)	0.02 (14)	0.02 (10)
seriously	0.05 (29)	0.09 (49)	0.03 (14)
significantly	0.05 (28)	0.002 (1)	0.02 (10)
surprised	0.02 (12)	0.05 (30)	0.03 (18)
surprising	0.009 (5)	0.02 (14)	0.03 (14)
surprisingly	0.009 (5)	0.005 (3)	0.01 (7)
theoretically	0	0.004 (2)	0.004 (2)
unfortunately	0.1 (53)	0.07 (42)	0.06 (3)
unfortunate	0.05 (26)	0.03 (19)	0.01 (7)
unluckily	0	0	0
usually	0.24 (123)	0.3 (169)	0.24 (129)
unwise to	0	0	0
wise to	0.002 (1)	0.002 (1)	0.004 (2)

Total number of different emphatic stance markers in ICE-Nig, ICE-Gha and ICE-GB

Emphatic stance markers	ICE-Nig	ICE-Gha	ICE-GB
absolutely	0.04 (21)	0.007 (4)	0.17(92)
completely	0.04 (22)	0.09 (48)	0.13(71)
crucially	0	0	0.006 (3)
exactly	0.15 (84)	0.06 (36)	0.17 (92)
extremely	0.03 (17)	0.03 (16)	0.09 (49)
particularly	0.13 (72)	0.06 (34)	0.28 (147)
really	0.98 (532)	1.3 (729)	1.6 (846)
totally	0.03 (17)	0.05 (30)	0.08 (44)

Namibian English on the web

Lexical and morphosyntactic features in a Corpus of Namibian Online Newspapers (CNamON)

Alexander Kautzsch
University of Regensburg

The present chapter introduces the Corpus of Namibian Online Newspapers (CNamON), which encompasses the contents of seventeen news sources as available on the Internet from May to June 2016. These sources add up to roughly 44 million words of text. The corpus was compiled to facilitate taking a systematic account of the lexical and structural properties of Namibian English, which have not been investigated to date. This chapter focusses on the technical details of the compilation process of CNamON as well as its set-up and its usefulness for linguistic research. To demonstrate the corpus' potential for analyses on the linguistic levels of lexis and (morpho-)syntax, both a qualitative stock-taking and exemplary quantitative analyses of structural characteristics of English in Namibia are provided.

1. Introduction

The role English plays in Namibia is unique as it was declared the only official language of the country at independence in 1990, although there were hardly any speakers of English as a first language. As a consequence of this political decision by the SWAPO (South West African People's Organization) government, English has been used in all public domains (government, administration, education, the media) for more than 25 years now. Several studies have been published on education and language policy, discussing, among other topics, the implications of the English-only decision (e.g. Frydman 2011; Harlech-Jones 1995, 2006; Pütz 1995; Töttemeyer 2010), but the overall body of research on English in Namibia is still in its infancy. Besides a pilot study on language attitudes and use (Buschfeld & Kautzsch 2014), which also provides a tentative list of potential linguistic features, the areas of linguistic description dealt with so far include codeswitching (Stell 2014a, 2014b), pronunciation (Kautzsch & Schröder 2016; Zähres 2016; Kautzsch

et al. in prep.), and pragmatics (Schneider & Schröder in prep.). Interestingly, the systematic identification and investigation of the linguistic properties of English in Namibia from a World Englishes and variationist perspective, that is considering lexical, morpho-syntactic and syntactic properties as indicators of structural nativization, is largely missing, as is an empirical examination of the assertion that Namibian English (Name) is merely an offshoot of South African English (SAfrE).

This is where the contribution at hand comes in. To obtain a first account of the lexical and structural properties of English as used in Namibia, I have compiled the Corpus of Namibian Online News (CNamON), comprising about 44 million words of written English as published on the websites of seventeen Namibian newspapers and magazines. After a brief overview of Namibia (Section 2), the corpus will be introduced (Section 3). Then, in order to demonstrate the potential of CNamON as a research tool, I will provide some qualitative stock-taking of Namibianisms on the levels of lexis (Section 4.1), phrase structure and syntax (4.2), as well as discourse (4.3), and three exemplary quantitative accounts of low-frequency features (4.4), for which a corpus of this size is very beneficial. The summary and discussion (Section 5) will address the issues of potential nativization of English in Namibia as well as the alleged similarity to South African English.

2. Namibia: An overview

Besides a quick overview of the facts and figures of Namibia, this section provides a brief history of Namibia, as well as an account of the linguistic ecology of this highly multi-ethnic nation.

2.1 Facts and figures

With a size of 832,690 square kilometres and a population of about two million (Wallace 2011: 2), Namibia is one of the least populated countries of the world (United Nations Department of Economic and Social Affairs 2011: 60). As regards the make-up of the population, the majority of 87.5% belong to one of several indigenous ethnicities (Ovambo 50%, Kavango 9%, Herero 7%, Damara 7%, Nama 5%, Caprivian 4%, Bushmen 3%, Basters 2%, Tswana 0.5%); another 6% is white (mostly Afrikaners, also called Boers, and some Germans), while 6.5% are of mixed ethnicity (CIA 2017; Buschfeld & Kautzsch 2014: 122).

Since independence in 1990, Namibia's official language has been English. In addition, sixteen Bantu languages (e.g., Oshiwambo, Kavango, or Otjiherero), about nine Khoisan languages (e.g., Nama/Damara or San), and two Indo-European

languages (Afrikaans and German) are spoken in the country (cf. Böhm 2003: 525; Buschfeld & Kautzsch 2014; Simons & Fennig 2017; Namibia Statistics Agency 2011: 68; NIDS 2006: 42).

2.2 A brief history of Namibia

The history of the territory now called Namibia has seen four larger phases. Up to the late 19th century, in the precolonial era, the region was settled by Khoisan (San, Nama, Damara) and Bantu (Ovambo, Herero, Kavango, Caprivian) peoples. In the course of the 19th century, the Oorlams and the Basters, two ethnically mixed, Afrikaans speaking communities from South Africa, started settling in different locations (for details, cf. Tonchi et al. 2012: 37–38, 322–323).

The second phase in the history of Namibia comprises the years 1884 to 1915, when it was a German colony called “German South-West Africa”. With the end of the First World War, that is the beginning of the third phase, the country, renamed to South-West Africa, was handed over to South Africa as a mandate and remained under South African administration / occupation until independence in 1990. During this time, its name was again changed to Namibia in 1968. The years of South-African rule saw a long and violent liberation struggle, led by the SWAPO and its figurehead and first president to be, Sam Nujoma, which finally resulted in independence in 1990. From then onwards, Namibia has been a democracy with SWAPO being the dominant party ever since. From a linguistic point of view, 1990 introduced a turning point in the linguistic ecology of Namibia, with English being made the only official language. The reason for this decision is centred around the idea to have an ethnically neutral language for intranational communication as a unifying force as well as a to provide Namibians with a language that facilitates international communication (for details cf. Böhm 2003: 518–9; Buschfeld & Kautzsch 2014; Klein 1984: 15; Maho 1998: 8–9, Wallace 2011: 115–6; Kautzsch *fc.*).

2.3 Linguistic ecology

Due to the presence of a wide variety of ethnic groups, multilingualism is the norm in Namibia rather than the exception, just like in the rest of Southern Africa (cf. Kamwangamalu 2013: 792). Table 1 shows the 2011 census results for the main languages spoken per household, with Oshiwambo languages (48.9%) being most widely used by far, followed at a distance by Nama/ Damara (11.3%), Afrikaans (10.4%) as well as Otjiherero (8.6%) and Kavango (8.5%) languages.

Table 1. Languages spoken in Namibia (per household) (Namibia Statistics Agency 2011: 68; adapted)

Main language spoken	Number of households	Percentage
Oshiwambo languages	227,103	48.9
Nama/Damara	52,450	11.3
Afrikaans	48,238	10.4
Otjiherero languages	40,000	8.6
Kavango languages	39,566	8.5
Caprivi languages	22,484	4.8
English	15,912	3.4
Other African languages	5,795	1.3
German	4,359	0.9
San languages	3,745	0.8
Other European languages	3,306	0.7
Setswana	1,328	0.3
Asian languages	461	0.1
Don't know	92	0.0
Total	464,839	100.0

English, the *de jure* official language of the country, is used as the main language by only 3.4 percent of all households. These figures have been supported in Buschfeld & Kautzsch's (2014) pilot study on language attitudes and use in Namibia, where they found that indigenous languages are clearly preferred in the family domain. However, English and Afrikaans are widely used for interethnic communication in all other domains (daily life, friends, education, business, etc.), with English being preferred by the younger generation, while the older generation still favors Afrikaans (cf. Buschfeld & Kautzsch 2014). As regards the shape and structure of the English used in Namibia, it is not hard to predict ethnic differences on various levels of linguistic description resulting from multiple contact of English with the local languages (cf. Buschfeld & Kautzsch 2014 for a tentative list of features on different linguistic levels; cf. Kautzsch and Schröder 2016; Kautzsch et al. in prep. for vowels; Schneider & Schröder in prep. for pragmatics).

As regards the treatment of English as spoken in Namibia within the World Englishes paradigm, it seems timely to start investigating whether NamE is merely an offshoot of SAfrE, as has been claimed or rather implied by, for example, Trudgill & Hannah (2002: 28). Taking into consideration the findings about attitudes and use as well as the special multi-ethnic and multi-lingual ecology of Namibia, it would not come as a surprise if Namibianisms on all linguistic levels were identified, which constitutes one of the aims of the present investigation.

3. A Corpus of Namibian Online Newspapers (CNamON)

This section presents details of my Corpus of Namibian Online Newspapers, CNamON. It provides a survey of the print media landscape in Namibia (3.1), a survey of the content and size of CNamON (3.2), some information about the technical side of the compilation process (3.3) as well as an account of the corpus's benefits and limitations (3.4).

3.1 Namibian newspapers and magazines

A comprehensive account of the Namibian print media is hard to obtain. Table 2 collects all information available about the founding years, circulation figures and ownerships from a variety of sources (i.e. Rothe 2010; Larsen 2007; Riddle 1989; Friedrich-Ebert-Stiftung 2015; Wikipedia's "List of newspapers in Namibia" and its German version "Liste namibischer Zeitungen und Zeitschriften", as well as the media's own websites as referenced in the footnotes for Table 2).

The oldest English language newspaper is the *Namib Times*, founded in 1958 and owned by Namib Times Publishing, is currently published twice a week only regionally for Walvis Bay and Swakopmund (Rothe 2010: 34). The oldest nationwide English language newspapers are the weekly *Windhoek Observer* (founded in 1978; owned by Paragon Media) and the now daily *Namibian* (1985; Free Press of Namibia Trust) (Rothe 2010: 28, 36; Friedrich-Ebert-Stiftung 2015: 28). The traditional Afrikaans daily *Die Republikein* (1977; Namibia Media Holdings [NMH]) "started including English language articles a few years ago" (Rothe 2010: 21; according to Riddle 1989: 48 it had been published in "Afrikaans/English/German" already before 1989), and also the oldest newspaper of Namibia, the German daily *Allgemeine Zeitung* (1916; NHM), provides some content in English.

Since independence, several other English-medium print media have been established: the now daily national government-owned *New Era* (1991; New Era Publications Corporation), the weekly *Namibia Economist* (1991; privately owned), which was turned into an online-only magazine in December 2016, and the four weekly nationwide tabloids *Confidente* (2011; Max Media Conglomerate), *Informante* (2006; Trustco Group), the *Namibian Sun* (2007; NMH), and *The Villager* (2011; Omalaeti Media Group). In addition, there are three weekly community papers circulated locally only: *Erongo* (2014; NMH) for the Erongo region, *EwilyaNooli* (2015; NMH) for northern Namibia, and *Windhoek Express* (2012; NMH) for the Khomas region around Windhoek. Finally, two monthly magazines, *Insight* (2004; privately owned) and *Prime Focus* (2008; Omalaeti Media Group), as well as *Oshili Nashi Popiwe*, an online-only news provider (2014; unknown), complete the picture.

Table 2. Namibian newspapers and magazines

Name	Type	Circulation (est.)	Founded	Owned by
<i>Allgemeine Zeitung</i>	Daily	4,000 to 5,000 per day	1916	Namibia Media Holdings (NMH)
<i>Namibian Sun</i>	Daily	20,000 per day	2007	NMH
<i>New Era</i>	Daily	40,000 per day	1991	Government-owned New Era Publications Corporation
<i>Republikein</i>	Daily	20,000 per day	1977	NMH
<i>The Namibian</i>	Daily	40,000 to 60,000 (Fri)	1985	Free Press of Namibia Trust
<i>Confidante</i>	Weekly	8,000 per week	2011	Max Media Conglomerate
<i>Erongo</i>	Weekly	?	2014	NMH (community paper for Erongo)
<i>EwilyaNooli</i>	Weekly	11,000 per month	2015	NMH (community paper for northern Namibia)
<i>Informante</i>	Weekly	65,000 per week	2006 (?)	Trustco Group
<i>Namib Times</i>	Weekly	14,000 per week	1958	Namib Times Publishing
<i>The Villager</i>	Weekly	10,000 per week ^a	2011	Omalaeti Media Group
<i>Windhoek Express</i>	Weekly	?	2012 (?)	NMH (community paper for Khomas region)
<i>Windhoek Observer</i>	Weekly	12,000 to 15,000	1978	Paragon Media
<i>Namibia Economist</i>	Weekly / Online ^b	7,000 per week (until Dec 2016)	1991	Private (Daniel Steinmann) ^c
<i>Insight</i>	Magazine	3,000 per month	2004	Private (TangeniAmupadhi)
<i>Prime Focus</i>	Magazine	5,000 per month ^d	2008	Omalaeti Media Group
<i>Oshili Nashi Popiwe</i>	Online	?	2014 (?)	unknown

(Friedrich-Ebert-Stiftung 2015; Rothe 2010; Riddle 1989; Larsen 2007)

a. <<http://www.omalaeti.com/files/files/TV%20profile.pdf>>

b. Exclusively online since December 2016 <<https://economist.com.na/20832/headlines/economist-goes-fully-digital/>>

c. <<https://economist.com.na/20832/headlines/economist-goes-fully-digital/>>

d. <<http://www.omalaeti.com/files/files/Prime%20Focus%20magazine.pdf>>

3.2 Content and size

In addition to the online-only sources, all news providers mentioned above also run websites where their print content is represented in digital format. It is these online data which were collected and integrated to build CNamON, which, taken together, contains the online content of seventeen news services, totalling about 44 million words (cf. Table 3).

The number of archived articles as well as the time span covered vary greatly by source. As documented in Table 3, some websites, e.g., *Allgemeine Zeitung*, *Informante*, *Insight* or *New Era*, provide articles older than ten years, while others, of course often mirroring their founding years, only cover shorter periods. As far as size is concerned, some sources only provide a small number of articles, while other archives are massive. As a consequence, the respective sub-corpora range from about 73,000 words for the *Namib Times* to more than 11 million words for the *Namibian Sun*.

Table 3. Namibian news sources online

Name	URL	Time span	Size (words)
<i>Allgemeine Zeitung</i>	< http://www.az.com.na/ >	2004–2016	199,689
<i>Confidente</i>	< http://www.confidente.com.na/ >	2012–2016	2,474,804
<i>Erongo</i>	< http://www.erongo.com.na/ >	2014–2016	325,906
<i>EwilyaNooli</i>	< http://www.ewi.com.na/ >	2015–2016	99,162
<i>Informante</i>	< http://www.informante.web.na/ >	2006–2016	6,218,562
<i>Insight</i>	< http://www.insight.com.na/ >	2006–2016	2,062,312
<i>Namib Times</i>	< http://www.namibtimes.net/ >	2015–2016	73,149
<i>Namibia Economist</i>	< http://www.economist.com.na/ >	2011–2016	1,828,232
<i>Namibian Sun</i>	< http://namibiansun.com/ >	2010–2016	11,867,697
<i>New Era</i>	< http://www.newera.com.na/ >	2005–2016	8,293,654
<i>Oshili24</i>	< http://www.oshili24.com.na/ >	2014–2016	83,144
<i>Prime Focus</i>	< http://www.primefocusmag.com/ >	2010–2016	1,136,041
<i>Republikein</i>	< http://www.republikein.com.na/ >	2009–2016	459,314
<i>The Namibian</i>	< http://www.namibian.com.na/ >	2013–2016	430,318
<i>The Villager</i>	< http://www.thevillager.com.na/ >	2011–2016	5,373,599
<i>Windhoek Express</i>	< http://www.we.com.na/ >	2014–2016	473,714
<i>Windhoek Observer</i>	< http://www.observer.com.na/ >	2012–2016	3,419,597
			44,818,894

3.3 Compilation

The articles comprising CNamON were collected between May and July 2016. Using the software HTTrack (Roche 2017), the websites of the above-mentioned news services were downloaded for offline viewing. Following the selection of relevant *.htm(l) files, that is identifying the ones containing news content and dismissing duplicates and overview pages, from the downloaded folder structure, plain text files were created by means of a beta version of BootCaT¹ (Baroni &

1. I am most grateful to Eros Zanchetta (Laboratory of Terminological Research, University of Bologna) for providing me with this version of BootCaT.

Bernardini 2004), a bootstrapping application, which removes HTML code from HTML files. The plain text files obtained had to undergo several stages of post-processing. First, all files were recoded to UTF-8, the text format recommended for AntConc (Anthony 2014), a common application used for corpus searches. Second, since many text files contained a variety of “odd” characters or character combinations, e.g. Ã«, âœœ, or â€”, which had resulted from recoding from office applications to XML or HTML, these had to be replaced (cf. TexTextin 2011 for an overview of correspondences between “odd” and intended characters or character combinations, e.g. Ã«→ ë; âœœ→ “; â€”→ –, etc.). Next, the plain text files had to be split up into smaller chunks of about 5MB maximum to prevent AntConc from crashing. Finally, the files were part-of-speech tagged using TagAnt (Anthony 2015), with the resulting tagged files again being adjusted in terms of size. All in all, the untagged and tagged versions of CNamON consist of 106 and 123 text files, respectively, which can conveniently be loaded into and searched with concordancing software.

3.4 Benefits and limitations

Having presented the details of the compilation, content, and size of CNamON, I will now turn to a brief evaluation of its benefits and limitations for an investigation of Name.

Beginning with the benefits, the texts are (excluding a small proportion of uncorrected contributions by readers, that is opinions, classifieds, SMSs, readers’ comments) produced and edited by highly proficient users, which implies that structural peculiarities on the level of lexis, morphology, morpho-syntax, syntax, and discourse cannot easily be dismissed as learner mistakes (cf. Gries & Bernaisch 2016: 7 on the same argument with reference to a South Asian newspaper corpus).

The above-mentioned contributions by readers can also be seen as an asset, delivering a sub-corpus or sub-corpora representing written registers close to speech. Moreover, CNamON contains articles from both tabloid and quality newspapers, although at times the line between them is difficult to draw in the Namibian context. Yet, the mere fact that the corpus comprises a range of registers also makes it a valid tool for analyzing stylistic variation in news language.

One of the big advantages of CNamON is its size of 44 million words, which provides an excellent starting point for identifying features of English as used in Namibia, even low frequency features. Being a collection of news texts, CNamON complements the massive NOW corpus (*News on the Web*; currently 4.82 billion words, growing daily; Davis 2013b), which includes news from twenty countries (e.g. UK, US, South Africa, Kenya, Tanzania, Ghana, Nigeria, etc.) but not from

Namibia, so that a comparison of CNameON and NOW can help shed light on the relationship between NameE and other Englishes around the world in general and African Englishes in particular. Contrasting the use of English in Namibia and in South Africa will be especially fruitful for examining the alleged similarity of NameE and SAfrE.

One of the limitations of CNameON, which it shares with many large corpora like NOW (Davies 2013b) or GloWbE (Davis 2013a), is that no demographic information about the authors of the texts is available, which prevents analyses according to certain sociolinguistic parameters such as age, educational status or ethnicity, although the latter might be inferred from the authors' names. Moreover, the fact that CNameON comprises mostly edited written language is likely to result in an absence or underrepresentation of colloquial features more likely to be found in speech. Finally, it needs to be mentioned that newspaper language is not a single genre, but rather contains different text types (e.g., opinion, reports, interviews, classifieds, etc.), which requires careful consideration during any analyses.

4. Lexical and syntactic features of NameE in CNameON

This section now turns to CNameON as a research tool. At the heart of the present investigation lies the attempt to identify potential features of NameE and document their existence and usage in CNameON.² I will take a decidedly qualitative approach in 4.1 to 4.3 as a first step, considering peculiarities of lexis, word formation, morphology, phrase structure, syntax, and discourse as found in Namibian news services online. In 4.4, three quantitative accounts of selected features will be presented to demonstrate the power of CNameON in analyzing low frequency features.

It is important to recall that new features in emerging Englishes will always originate from two language acquisition processes. The one will be cross-linguistic influence, also known as native-language or L1 transfer, which will give the English used in Namibia its particular Namibian flavor. The other will be universal features of language learning and usage, which are responsible for similarities across Englishes (cf. de Klerk and Gough 2002: 363; McCormick 2002: 228).

2. My quest for Namibianisms was partly inspired by personal observations in the country, by conversations with Namibians, by Wikipedia's entry on "Namlish" (<https://en.wikipedia.org/wiki/Namlish>), some lay people's reports on the usage of English in Namibia (e.g. <<http://www.namibian.org/travel/misc/namlish.html>>), as well as by two masters theses dealing with errors found in the writings of Namibian school children (Mungungu 2010; Nghikembua 2015).

My objective here, from a World Englishes perspective, is to answer the research question of whether the English language as used in online news services in Namibia provides evidence of “the emergence of locally distinctive linguistic forms and structures” (Schneider 2007: 71), in other words, of structural nativization. Although structural nativization in Schneider’s sense merely refers to territories with a British (or American) colonial history, recent theorizing has shown that non-postcolonial Englishes, like in the Netherlands (Edwards 2016) or, for that matter, Namibia (Buschfeld & Kautzsch 2014, 2017; Kautzsch *fc.*) are also capable of entering a nativization phase, especially when widely used intra-nationally. For the current investigation, this translates into the assumption that, as soon as certain linguistic phenomena occur regularly in published written English, they are deeply entrenched in the grammar and lexis of the area. The following sections will now turn to a documentation of such nativized or nativizing features of NamE.

4.1 Lexis and word-formation

The most obvious influence of local languages on emerging Englishes in any region is the usage of loan words. In a multi-lingual setting like Namibia, loan words can originate from a wide range of source languages. In this section, the focus will be on borrowings from German, the language of the first colonizers, Afrikaans, the lingua franca used widely during the time of South African rule, and Oshiwambo, the language spoken by the largest ethnic group.

4.1.1 *German loanwords*

The influence of German on the NamE lexicon seems to be restricted to proper names, that is first names and place names. Table 4 exemplarily lists German first names that were found in C_{NamON} in combination with indigenous last names and gives the numbers of individuals who are named that way. The most popular German first name in the list is *Andreas* (159 individuals), followed by *Erwin* (39), *Reinhold* (29), etc. Obviously, this does not mirror the popularity of German first names with indigenous Namibians, but rather which individuals are mentioned in news services. What is striking about the names as such is that in Germany, according to my own impression, most if not all of them have the connotation that their bearers belong to the group of people aged about 45–80. This is different in Namibia, where these names seem to be generation independent: the two *Elfriedas* listed in the table, for example, are 19 and 44, respectively (*Nambian Sun*, 2014-04-27; *New Era* 2013-12-04) while Hage *Gottfried* Geingob, Namibia’s current president, is 75.

Table 4. German first names in combination with indigenous last names

Name	Individuals in CNameON	Examples
Andreas	159	Andreas A. Kashima, Andreas Chauks !Aebeb
Berthold	10	Berthold Mbinda (RnB singer), Berthold Karumendu
Elfrieda	6	Elfrieda Hausiku, Elfrieda Kenamuni
Emil	8	Emil Garoeb, Emil Kanduuombe
Erwin	39	Erwin Afrikaner, Erwin Kaujende (football players)
Franz	7	Franz Uirab, Franz Tumuna
Gerhard	26	Gerhard Ipingeliputa, Gerhard Shipena
Gottlieb	23	Gottlieb Ndokosho (professional fighter), Gottlieb Amanyanga
Gottfried	14	Hage Gottfried Geingob (president), Gottfried Ashipala,
Reinhold	29	Reinhold Amugatanu, Reinhold Geikhoebob
Renate	7	Renathe [sic!] Kandongo

In addition to first names, German place names are also found, which is exemplified in an announcement of one of Namibia's mobile phone providers published in *Confidente* (cf. Example (1)). Given Namibia's German colonial and settlement history, the fact that these names are still in use does not come as a surprise, even if some places have been renamed, e.g. the constituency of *Lüderitz* to *!Nami#Nüs*.

- (1) The following are the various sites TN Mobile has upgraded in all regions of the country during the past one year: 53 sites in Khomas Region, including the entire Windhoek, Kapps Farm, Hosea Kutako International Airport, Omeya Gold Estate and **Elissenheim**, 23 sites in Otjozondjupa Region, including Otjiwarongo, Grootfontein, Okahandja, Otavi, Okakarara, Midgard, **Bergquell**, Osire, Okorusu Mine, Klein Waterberg and Ohorongo Cement, [...] 16 sites in Hardap Region, including Rehoboth, **Mariental**, **Maltahohe**, Aranos, Stampriet, Kalkrand and Gibeon, [...] 11 sites in //Karas Region, including **Aussenkehr**, Rosh Pinah, Skorpion Zink Mine, Aus, **Grunau**, Karasburg, **Luderitz** and Oranjemund, [...] 12 sites in Omaheke Region, including Gobabis, **Steinhausen**, Summer Down, Buitepos, Aminius, Epukiro and Omaere, [...].
(*Confidente*, 2014-08-29)

4.1.2 Afrikaans loanwords

As mentioned above, Namibia never belonged to South Africa but used to be an administrative region of that country for more than seventy years. Since with South African occupation also Afrikaans entered Namibia and developed into a means of inter-ethnic communication, many Afrikaans words are used in NamE. Among them are objects or events of daily life (e.g. *bakkie*, *braai*), food (*biltong*, *wors*), *lekker*, an adjective with a very general meaning, or a form of address (*oom*).

- *bakkie*: ‘pick-up truck’ [659]³
 - (2) He was seated at the back of a bakkie with three other relatives.
(*Confidente*, 2013-08-29)
 - (3) ... who collected all the items that had fallen off our bakkie.
(*Erongo*, 2015-01-29)
- *braai*: ‘barbeque’, used as a noun or verb [635]
 - (4) A kitchen facility, including a sink and basic braai amenities, are provided at each site.
(*Allgemeine Zeitung*, 2008-11-06)
 - (5) For us, summertime means relaxing with great jams while braaing.
(*The Villager*, 2012-10-28)
- *biltong*: (Afr. *bil* ‘buttock’ + *tong* ‘tongue’) ‘dried meat’ [141]
 - (6) Namibia’s finest biltong is manufactured by Closwa, Piet’s Biltong and Butchery on Wheels in town.
(*Windhoek Express*, 2015-10-22)
- *wors*: ‘sausage’ [110]
 - (7) The plant has a sophisticated turbo smoker, mincing machine and meat crusher to produce boerewors, Russian and Vienna sausages, biltong, droewors, chilli bites and assorted meats.
(*Namibian Sun*, 2014-12-28)
- *lekker*: ‘good, tasty’ [95]
 - (8) Lady May enjoys big, lekker and heavy meals when she visits restaurants or attends parties.
(*Informante*, 2011-10-13)
 - (9) Grapevine had a super lekker weekend...
(*Confidente*, 2013-09-12)
- *klap*: ‘punch’, used as a noun or verb [11]
 - (10) The third guy was about to do the same when he stopped and gave me a moerse [huge] klap in my face before getting out.
(*Namibian Sun*, 2013-07-12)
 - (11) Being ‘klapped’ over the head with a blunt metal instrument would be preferable to reading this drivel.
(*Insight Namibia*, 2009-10-05)
- *oom*: ‘uncle’ [98] form of address for elderly people, with reference to boers
 - (12) A fitting posthumous tribute to Harold George Van Reenen fondly referred to as Oom Harry, a man who [...]
(*Erongo*, 2015-09-24)
 - (13) Sometime ago, I stopped at a service station in Omaruru to fill up my car, when an oom – probably in his 80s – pulled up right opposite me in a battered bakkie.
(*Namibian Sun*, 2014-10-31)

3. Figures in brackets give the frequency of the respective words in CNamON.

4.1.3 Oshiwambo loanwords

The third major source language for loanwords in NamE is Oshiwambo, the language spoken by the ethnic majority. Similar to Afrikaans, these words also refer to food (*kapana*, *marula*), drink (*tombo*), objects of daily life (*lapa*, *cuca shop*), and forms of address (*meme*, *tate*). In addition, we find *makakunyas*, a derogatory term for Owambos fighting for South Africa during the liberation struggle.

- *meme* / *tate*: forms of address for elderly women / men [*meme*: 472; *tate*: 511]
 - (14) Other key people that I was introduced to included the late Cde Jacob Shifula and his wife, the late meme Maria, and the Senior Headman of Elyambala, Tate Kayupa Mwatotele. (New Era, 2015-07-28)
- *kapana*: barbecue meat [497]
 - (15) She said if Lukas attends school again she will be able to sell kapana to generate an income. (New Era, 2015-08-07)
 - (16) We meet a lot of different people on a daily basis, some who have not even tasted kapana yet but always keep coming back for more. (Confidente, 2013-10-03)
- *marula*: fruit of the marula tree; *omagongo*: drink made from marula fruit [*marula*: 244; *omagongo*: 75, including “Omagongo Festival”]
 - (17) The alleged victim and wife of Staff Sergeant at the base [...] said that the incident happened two weeks ago when she was ordered to take omagongo (an Oshiwambo traditional brew extracted from marula fruits) to Masipa’s room in the base. (Confidente, 2013-02-28)
- *tombo*: ‘privately brewed alcoholic drink’ [231]
 - (18) My own research led me to believe that the people were abusing a locally brewed drink, tombo, a potent and dangerous drink. (Confidente, 2014-09-25)
 - (19) Every day Sheehama makes 25 litres of tombo, which she sells only to people from her street. (Windhoek Observer, 2014-09-26)
- *cuca shop*: ‘unlicensed shop selling alcohol’ [168]
 - (20) Back at the cuca shop, she was alerted by one of the villagers that there was a fire coming from one of the huts in her homestead. (Namibian Sun, 2011-08-29)
 - (21) Nawa said the victim was on her way from a local cuca shop when the suspect approached her and started courting her. (New Era, 2016-04-29)
- *makakunyas*: ‘blood-suckers’; black Owambos in the paramilitary South African Koevoet unit, fighting against their own people [26]

(22) Our most traumatic experience with the Makakunyas.
(*Namibian Sun*, 2014-05-29)

(23) The makakunyas were responsible for the destruction of properties and the killing of innocent Namibian people during the war.
(*New Era*, 2015-03-02)

– *lapa*: ‘roof under which a barbecue takes place’ [22]

(24) In the summer months there is a traditional verandah and an open lapa for lunch and dinner or for simply relaxing with your friends.
(*Namibian Sun*, 2012-07-27)

(25) During the morning, training was given under the campsite’s lapa and large shady trees adjacent to the farm’s dry river bed.
(*Windhoek Express*, 2016-03-03)

4.1.4 *English lexis and word-formation*

Turning to some peculiarities of English lexis and word-formation, NamE has developed new blends like the abbreviation for the Namibian police force, *Nampol*, and the nickname for the local variety, *Namlish*, which is mostly used with a slight derogatory undertone:

– *Nampol*: Namibian Police [2606]

(26) The manager on duty only then reacted to get the acid cleaned up and called Nampol, of which 12 men showed up. (*Allgemeine Zeitung*, 2016-01-29)

(27) With KhomasNampol women soccer team, Katanga won the police championship for two consecutive years, 2015 and 2016.
(*EwilyaNooli*, 2016-07-27)

– *Namlish*: Namibia + English [82]

(28) Drinking holes, mini-markets, small to medium enterprises and even Government buildings can be seen all over the Northern regions exampling Namlish of the highest degree.
(*Informante*, 2009-02-04)

(29) Namlish is the local is lekker chakalaka spice that you bring to a good old British pot roast.
(*Insight Namibia*, 2012-09-13)

Further, two new compounds have been formed to describe two groups of people emerging from Namibia’s struggle for liberation from South African oppression, the *struggle kids* and the *born frees*:

– *Struggle kids*: Namibians who grew up in exile during the liberation struggle [1324]

(30) President Hifikepunye Pohamba was reportedly quoted as having dismissed selective special treatment to the children of the liberation struggle or “struggle kids” as they prefer to regard themselves
(*Confidante*, 2013-03-20)

- (31) Police arrested a few of the ‘struggle kids’ and then later took them back to the Swapo Party headquarters ... (*Windhoek Observer*, 2013-03-07)
- *Born frees*: Namibians born after Independence [128]
- (32) For the younger generation, a day-visit to the Heroes Acre is a must since many of the born-frees are only too ignorant of the sacrifices their parents had to make to bring liberty to Namibia. (*Namibia Economist*, 2015-12-04)
- (33) However, most of the born frees the Windhoek Observer spoke to admitted that the silver jubilee Independence celebrations were the first they had ever attended. (*Windhoek Observer*, 2015-03-27)

Moreover, the word *learner* has acquired a localized meaning referring to pupils in Namibia more regularly than in other varieties, excluding SAfrE.⁴

- *Learners*: pupils [9627]
- (34) The learners were faced with a number of challenges aimed to engage them as a team and to sharpen their problem-solving abilities. (*Republikein*, 2016-06-06)
- (35) However, learners are forced to sleep on the floors as there are no beds or mattresses, and there are also no bathrooms. (*The Namibian*, 2016-05-24)

Another observation can be made about the word *pain* used as a verb. In British or American English the pattern *It pains me so much that...*, as exemplified in (36), may also be used, while examples as shown in (37) to (39), are unlikely and suggest localized usages, which seem to increase the frequency of the verb *to pain* in NamE as opposed to other Englishes.⁵ For a quantification of this, cf. 4.4.3.

- *Pain*→*to pain* [*pained* 30, *pain/pains* 91, *paining* 20] (see below)
- (36) It pains me so much that my children have deserted me, because I carried five pregnancies for nine months each but yet I am suffering. (*New Era*, 2015-08-14)
- (37) You pained me: mother tells daughter’s killer. (*Confidente*, 2013-04-18)

4. A full-fledged quantification of the semantics of the word *learner* across varieties, as suggested by an anonymous reviewer, goes beyond the scope of this contribution. A search in the NOW corpus, however, suggests that, in addition to Namibia and South Africa, the meaning ‘pupil’ can also be found in some other varieties, among them India, Bangladesh, Kenya, Sri Lanka, or the Philippines.

5. I am grateful to an anonymous reviewer for pointing me to the fact that examples like in (38) or (39) may occur in other Englishes and that the overall higher frequency of this verb in Namibia may be a result of the discourse situation. An ongoing empirical analysis on the complementation patterns of *pain*, which goes beyond the one reported on in 4.4.3 (and also beyond the scope of this contribution), however, supports my assertions made above that NamE displays certain peculiarities in the use of this verb.

(38) When I was standing I would *pain*.

(39) I'd like to thank our medic DonathaGawanas who is always there to help when my body is *paining*.

Finally, new prepositional verbs like *return back*, *discuss about*, or *enter into* also occur in NamE. Their occurrence seems to be a universal feature of L2 Englishes (cf. for example Mukherjee 2007; Nesselhauf 2011).

– *return back* instead of *come back* or *return*

(40) Mathew said that, Ennio has returned back to his club Bournemouth in the English championship ... (The Villager, 2014-09-29)

(41) President Geingob returns back to Windhoek on today. (The Villager, 2015-05-07)

– *discuss about* instead of *discuss* or *talk about*

(42) He said he was tired of taking treatment and could not freely discuss about the side effects of taking treatment that he encountered. (The Villager, 2011-08-28)

(43) In my previous articles, we discussed about the inherent components of policy making process (Confidente, 2015-10-15)

– *enter into* instead of *enter* or *go into*

(44) At the eighth Conference of Parties in Madrid, the UNCCD entered into a new phase with the adoption of ... (Namibian Sun, 2013-09-23)

(45) ... the day we signed the allocation forms is not necessarily the day we entered into the houses ... (Informante, 2009-10-22)

4.2 Phrase structure and syntax

On the linguistic levels of phrase structure and syntax, NamE as documented in CNamON also displays an array of localized usages in verb, adjective, adverb, and noun phrases, as well as with discourse markers. Most of my investigation here was inspired by two accounts of errors made by Namibian pupils (cf. Mungungu 2010; Nghikembua 2015), but the fact that these structures occur in published written texts suggests that they ought to be seen as features rather than errors (for a literature review on the distinction between error and feature see, e.g. Groves 2010). Besides, some of the examples given below come from an unedited and thus more speech-like sub-category of CNamON, that is readers' comments and direct quotations; these can, therefore, not always be interpreted as evidence of features fully entrenched in the written register, but at least as evidence of potential candidates for structural nativization that deserve further investigation in spoken data.

4.2.1 Verb phrase

In the verb phrase, the standard English semi-modal *used to* sometimes occurs without the final <d> as *use to*, both in its standard reference to past events, as in (46) and (47), and, more interestingly, also with present reference, as in (48).

- (46) Africa is growing and is no longer the dark continent it use to be.
(*New Era*, 2014-03-26)
- (47) Namibia is not a beutiful [sic!] country as use to be anymore ... this brings me a headache.
(*New Era*, 2015-10-22)
- (48) thy [sic!] usually ask 18-25 years but old people dont[sic!] use to complain.
(*New Era*, 2014-01-15)

Another structure found in CNamON is the usage of the progressive outside its usual contexts. On the one hand, stative verbs are used in the progressive to express an unlimited state, as in (49) to (50), while, on the other, dynamic verbs occur in the progressive to encode habitual, non-temporary actions.⁶

- (49) And one of these is that we are having abundant land.
(*Namibia Economist*, 2013-05-03)
- (50) We looked at what is existing (in terms of policies) so that we can build on and strengthen it.
(*New Era*, 2016-02-25)
- (51) ... and the majority of black people are having air for dinner and shitting in baskets under their beds?
(*The Villager*, 2012-11-25)
- (52) I thank parents who are teaching their children about our culture because culture instils discipline which is lacking in many of our children.
(*New Era*, 2014-08-12)

The last verbal pattern to be mentioned here is *go* + bare infinitive, as in (53) and (54), or *go* + *and* + bare infinitive, as in (55) and (56), which seem to add some dynamism to the action described. Since, surprisingly, it occurs more frequently in CNamON than in other Englishes, a quantitative account of this structure will be presented in 4.4.2 below.

- (53) If he wants a place of his own he must go build it like he said in Katutura because ...
(*Windhoek Observer*, 2015-10-02)
- (54) I also do outreaches at schools were [sic!] I go talk to children and I encourage them to study.
(*The Villager*, 2016-05-23)

6. Many thanks go to an anonymous reviewer for pointing out inconsistencies in an earlier version of this contribution as regards the distinction between stative verbs with unlimited state semantics and dynamic verbs in habitual contexts.

- (55) You should go and apply for child support, which by law ...
(*Namibian Sun*, 2013-04-17)
- (56) We really want to go and vote but it does not change things.
(*New Era*, 2008-03-27)

4.2.2 Adjective phrase

Peculiarities in the adjective phrase in NamE entail, as in many other – mostly L2 – varieties, the double marking of comparatives and superlatives, shown in (57) to (60), as well as the non-standard intensifiers *very much* and *real*, exemplified in (61) to (64), where standard English would prefer *very* and *really*.⁷

- (57) ... meaning that it should also be more cheaper to service a set of residential plots in a specific area, than servicing ...
(*The Villager*, 2013-05-27)
- (58) ... and this is made more easier with the use of earplugs which he carries in his pocket...
(*New Era*, 2006-06-31)
- (59) 'Richard drinks a lot and has an abnormal tendency of becoming jealous for the most simplest things,' said Albertina Noreseb, Richard's mother.
(*The Villager*, 2012-01-23)
- (60) ...lies and innuendoes carried in that piece such as those quoted above which I will tackle in the most strongest terms in this piece. (*New Era*, 2006-01-30)
- (61) The xenophobic wound will be very much difficult to heal if the barbaric and maiming of foreign...
(*Confidente*, 2015-05-21)
- (62) Learning is a skill that is very much similar to sport. (*Prime Focus*, 2011-07)
- (63) He is a real awesome person and I love him very much.
(*Informante*, 2013-06-12)
- (64) we should also emphasise that we want to shame those that had the opportunity to be real good mothers but chose the easy way out (*The Villager*, 2015-05-11)

4.2.3 Adverb phrase

A very unusual structure found in CNamON, and to my knowledge not documented for any other variety of English so far, involves the time adverbs *today*, *yesterday*, and *tomorrow*. As shown in (65) to (70), these are regularly preceded by the preposition *on* in NamE, possibly in analogy with reference to days of the week, like, e.g. *on Monday*, *on Tuesday*, etc.

7. Of course, the use of *real* as an adverb is quite common in informal standard American English

- (65) ... on the 20th of October, the CEO of the Roads Authority announced in Windhoek on today. (*The Villager*, 2014-10-14)
- (66) Looking at the submission of both parties, I suggest we reconvene on today. (*Namibian Sun*, 2013-04-05)
- (67) The tournament started on yesterday and will go on until March 27 at Kuisebmond stadium. (*Confidente*, 2016-03-24)
- (68) Among other precious metals, platinum rose nearly 1 per cent on yesterday as labour strikes continued in top producer South Africa. (*Republikein*, 2014-04-01)
- (69) The visitors arrive at 10am on tomorrow (Friday). (*Informante*, 2009-11-26)
- (70) the FNCC will be closed to public today, but will reopen with regular hours on tomorrow. (*Windhoek Express*, 2016-07-14)

4.2.4 Noun phrase

Finally, three observations of determiner usage within the noun phrase can be made.⁸ The first is the occurrence of singular demonstrative *this* with plural reference, as in (71) and (72); the second involves the numeral *one* being used as a determinative and thus replacing an indefinite article, as in (73) to (75).

- (71) This are the babies of the 90s we talking about. (*New Era*, 2015-10-23)
- (72) This are just some of the interventions we put in place to address significant growth at the coast. (*Confidente*, 2014-10-09)
- (73) Namibia is one country where over 52% of its citizens survive through informal jobs (unemployed); these include shebeens, selling kapana (braai meat) or vending. (*Informante*, 2014-11-26)
- (74) Pssst! was playing around on Youtube the other day and came across a new Kwaito music video by Tulisan – one adorable boy never seen on the Namibian screens. (*Namibian Sun*, 2012-01-13)
- (75) Regardless of the dilapidated state of the gym, Hitman said Kilimanjaro is one boxing club in the north with good administration... (*New Era*, 2015-02-23)

8. It is not my intention here to provide features that occur exclusively in Namibia. The ones mentioned here are, like the adjective phrase features presented above, not restricted to NamE but serve to place NamE among other Englishes spoken in countries with a colonial background. What is interesting about the occurrence of these features in Namibia is that their emergence does not require a colonial past, which demonstrates in turn that the distinction between postcolonial and non-postcolonial Englishes might not reflect linguistic realities. For a discussion of this see Buschfeld & Kautzsch (2017). A more detailed account of the theoretical implications of my findings needs to be given elsewhere.

Third, Examples (76) to (79) demonstrate deviations from the standard distribution of the forms of the indefinite article *a/an*, where in NamE *a* can occur before vowels and *an* before consonants. Since CNamON delivers a comparatively high number of instances, a more detailed account is given in 4.4.1 below.

- (76) ... to see an institution which was supposed to be run on a impartial basis...
(*Confidente*, 2013-06-20)
- (77) ... the Isuzu-branded system also has a audio input terminal catering for ...
(*Informante*, 2014-01-16)
- (78) ... street sex workers, with well-paying clients, charge up to N\$800 per round of an sexual encounter, which normally lasts for less than five minutes.
(*The Villager*, 2013-09-09)
- (79) ... products regulations by South African authorities will have an dramatic impact ...
(*Windhoek Observer*, 2014-01-09)

4.3 Discourse

The last section of my qualitative account focusses on discourse markers. The first is the question tag *ne?*, which is regularly transferred from Afrikaans to English, as shown in (80) and (81).

- (80) Ja, some women just have a way with words, ne? (*Namibian Sun*, 2016-03-11)
- (81) So folks, no more under-dosing if you want your animals to recover from a disease, ne?
(*The Villager*, 2011-10-31)

And also, two of the three “new quotatives” *be like*, as in (82), and *go*, as in (83), have made their way to NamE; the third, *be all*, does not occur in CNamON.

- (82) I had someone come up to me on a night out, and he was like “Hi Sean”, and I said Hi. He seemed to be a bit offended that I did not recognise him. He continued like “I’m so and so from Facebook, you must know me.” I was just like, “Uhm, no sorry, I really don’t.”
(*The Villager*, 2015-11-02)
- (83) And there she was approaching from the far end headed straight at me and in her well versed Afrikaans she goes “Ekskuushoor, hier is geenwerknies, onsstelnieaannie.” I was shocked and enraged.
(*New Era*, 2007-08-31)

4.4 Three quantitative accounts

The present section now turns to quantitative accounts of three features of NamE: the non-standard usage of the indefinite article, the verbal pattern *go* + (*and*) + bare infinitive, and the usage of *pain* as a verb. These accounts are aimed at underlining three assumptions: First, NamE is undergoing structural nativization, that is it is developing features that are widely used and might set it apart from other varieties of English. Second, by identifying differences to SAfrE, it can be shown that NamE is developing into a new distinct variety and cannot be seen as a mere offshoot of SAfrE. Finally, the investigation of low-frequency features here highlights one of the advantages of a large corpus of an emerging variety, that is that it facilitates analyses that would not be possible with smaller corpora.

4.4.1 *Non-standard usage of indefinite article forms*

My first quantification examines the non-standard usage of the two forms of the indefinite article, that is the use of *a* in front of vowels and the use of *an* in front of consonants. Table 5 presents the overall frequencies of the two forms in CNamON and provides the token numbers as well as their proportions for the non-standard usages. Obviously, we are dealing with a comparatively small percentage of the non-standard occurrences (0.04% for *a*, 0.38% for *an*), which is hardly surprising, given the fact that CNamON contains written, potentially edited language. Nevertheless, the presence of close to 900 non-standard cases in online newspapers is indicative of the fact that these cannot simply be dismissed as slips of the pen overlooked by spell-checkers.

An explanation of this deviant usage of the indefinite article forms may be found in two languages deeply entrenched in the Namibian linguistic ecology: Oshiwambo and Afrikaans. The former does not have an article system (Crane et al. 2004: 12), and in the latter, the indefinite article is invariably spelt as '*n*, while the most widespread pronunciation is [ə] (Donaldson 1993: 65). Taking these two observations together, the non-standard usages of *a* and *an* may be a result of mutually reinforced L1 Oshiwambo / L1&L2 Afrikaans transfer.

Table 5. Non-standard usage of indefinite article forms

	Total (N)	Non-standard (N)	Non-standard (%)
<i>a</i>	825,992	327	0.04
<i>an</i>	142,825	549	0.38

4.4.2 *go* + (and) + *bare infinitive*

In the second quantitative account, I compare the frequencies of the structures *go* + bare infinitive and *go* + *and* + bare infinitive in CNameON and in the Englishes represented in the NOW corpus (Davies 2013b).

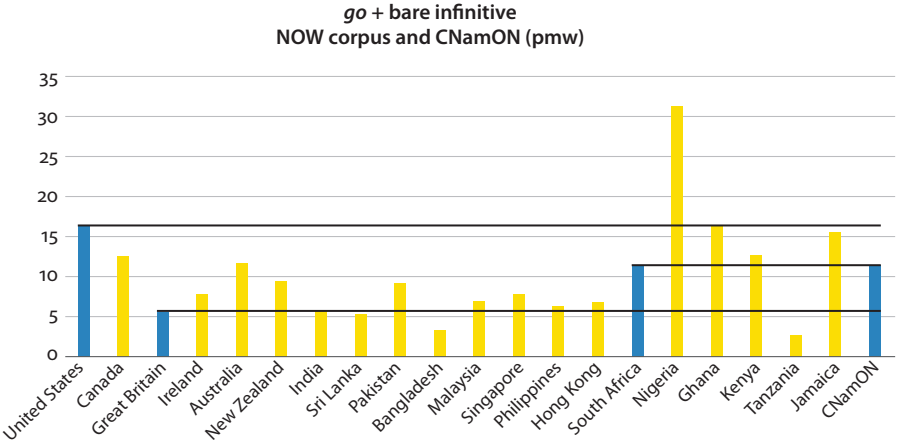


Figure 1. *go* + bare infinitive in NOW and CNameON

Figure 1 suggests that *go* + bare infinitive is not particularly rare in Englishes across the globe, occurring between three (Bangladesh) and thirty-one times (Nigeria) per million words (pmw). In Namibia (the column on the far right, CNameON), this pattern has the same frequency as in South Africa. American English (AmE) and British English (BrE) serve as standard reference varieties here, showing that this structure is less popular in Britain and more popular in the US than it is in Namibia and South Africa.

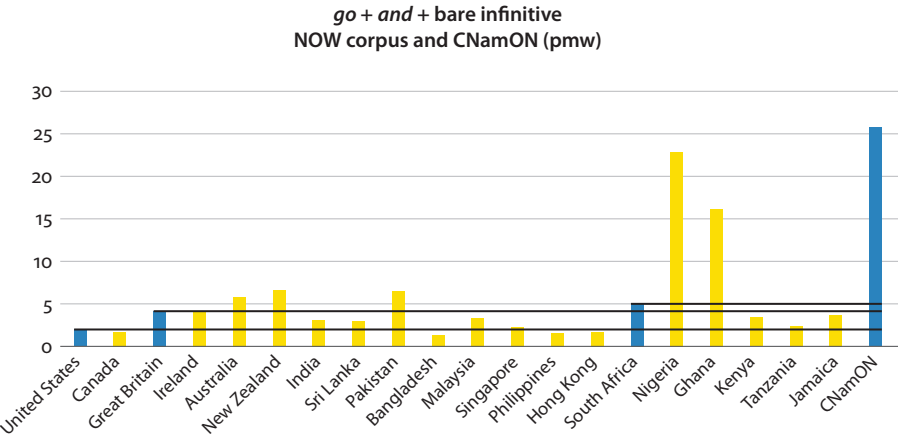


Figure 2. *go* + *and* + bare infinitive in NOW and CNameON

Adding an *and* between *go* and the bare infinitive, however, results in an interestingly different picture (cf. Figure 2). This pattern is comparatively infrequent in Englishes outside West Africa and Namibia, ranging from one (Bangladesh) to seven (Pakistan, New Zealand) occurrences per million words. Here, Namibia (26 pmw) does obviously not align with South Africa (5 pmw), but rather with Nigeria (23 pmw) and Ghana (16 pmw). A tentative explanation for this might be cross linguistic influence from Oshiwambo, a language not spoken in South Africa, for which an itive marker *ka* is documented, which explicitly marks the verb phrase for a relocation from the deictic center (Nurse 2008: 242, 244). Whether itives of this sort are also found in the languages spoken in Nigeria and Ghana, which would license the parallels, needs to be investigated elsewhere.

4.4.3 Pain as a verb

My final quantitative observation entails the frequency of the verb *pain* in CNameON as compared to the varieties represented in NOW. As mentioned above, CNameON delivered examples for *pain* as a verb that, in terms of syntactic patterns, go beyond the usages found in standard English (cf. 4.1.4, Examples (37) and (38)) and seem to translate into an overall higher frequency in NameE.

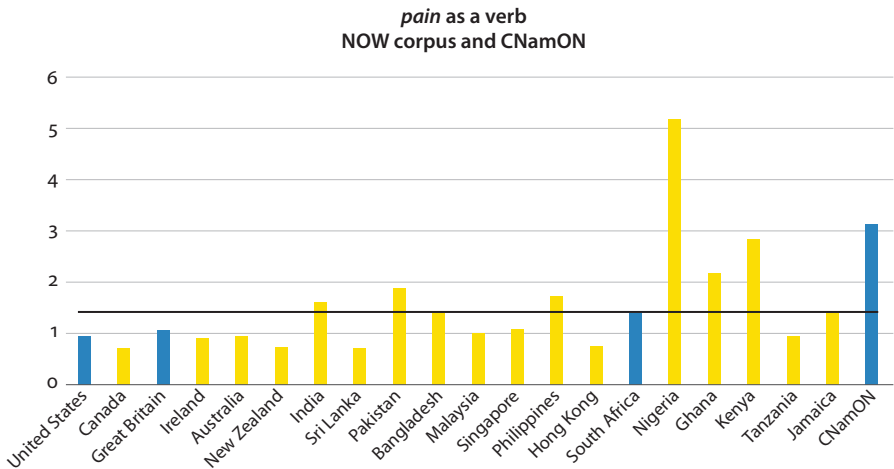


Figure 3. *Pain* as a verb in NOW and CNameON

This hypothesis is supported by Figure 3, which shows that *pain* occurs as a verb in all varieties covered by NOW, but more frequently so in Namibia. Interestingly enough, Namibia is again different from South Africa in this respect, once more aligning rather with Nigeria and Ghana (and here also Kenya).

5. Summary and discussion

Pulling together both the qualitative and quantitative findings based on CNamON as presented in Section 4, this corpus of Namibian news online has proved to be an excellent starting point for an exploration of structural peculiarities of English in Namibia, a so far neglected emerging variety of English outside a British or an American colonial context.

Evidence for unique Namibian usages were provided on the levels of lexis and (morpho-)syntax. As regards lexis, NamE contains a large number of loan words, most notably proper names from German, the language of the first colonizers, and words with reference to objects of daily life, forms of address, or terms for food and drink from Afrikaans, the language most widely used during South African occupation, and Oshiwambo, the largest indigenous language. In addition, creativity in forming new English words or expanding the meaning of existing ones was also documented. Moreover, the three quantitative accounts (4.4) provided some evidence for Namibian grammatical characteristics with respect to indefinite article usage, the pattern *go + and + bare infinitive*, and the syntactic structures involving the verb *pain*.

Like numerous other new Englishes, NamE also displays general features of L2 English, e.g. the presence of new prepositional verbs or the usage of *very much* as intensifiers. Beyond the mere identification of features, this investigation has also helped to place NamE within the framework of World Englishes research, especially in rejecting the erstwhile hypothesis that the English spoken in Namibia is only an offshoot of South African English. That this is not the case has been clearly demonstrated by the accounts of *pain* as a verb and the verbal structure *go + and + bare infinitive*, where decidedly higher frequencies emerged in Namibia than in South Africa.

6. Conclusion

The present chapter set out to document lexical and morpho-syntactic features of English in Namibia, a new variety of English in a non-postcolonial context which gained momentum after English was introduced as the only official language of the country with its independence in 1990. After more than twenty-five years during which English has made inroads in almost all domains of daily life, probably excluding the family, it was high time to initiate some stock-taking. For this purpose, the Corpus of Namibian Online News (CNamON) was compiled. It consists of 44 million words of English texts available on the websites of seventeen news services, covering both tabloid and quality content and a time span of about ten years (from

2006 to 2016). Both qualitative and quantitative evidence has been found that NamE may be in a phase of structural nativization, displaying on the levels of lexis and (morpho-)syntax utterly Namibian characteristics but also general features of L2 English. The fact that this can be documented on the basis of written data suggests that the emergence and incorporation of localized usages is fairly advanced, not lagging behind other Englishes with a colonization context. Martin Ipinge, an Oshakati resident in the far North of Namibia, is quoted in one of the news sources, *Informante* (2009-02-04), summing up the linguistic situation in Namibia from a very reasonable lay person's perspective:

As long as we have Bantu and Khoisan languages in the land of the brave, Namlish will continue to develop its own set of English words. And why shouldn't we speak Namlish? We will never be English or Americans but Africans until the day we die.
(*Informante*, 2009-02-04)

The task for us as linguists now is to keep track of these developments in Namibia, one of the most fascinating environments of emerging Englishes world-wide. CNamON, for the time being, will serve as a very fine starting point, but to grasp the full picture, other sources of data will have to be integrated, too.

References

- Anthony, Laurence. 2014. AntConc (Version 3.4.4w) [*Computer Software*]. Tokyo: Waseda University. <<http://www.laurenceanthony.net/>> (2 August 2017).
- Anthony, Laurence. 2015. TagAnt (Version 1.2.0) [*Computer Software*]. Tokyo: Waseda University. <<http://www.laurenceanthony.net/>> (2 August 2017).
- Baroni, Marco & Bernardini, Silvia. 2004. BootCaT: Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*. <http://sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf> (2 August 2017).
- Böhm, Michael Anton. 2003. Deutsch in Afrika. Die Stellung der deutschen Sprache in Afrika vor dem Hintergrund der bildungs- und sprachpolitischen Gegebenheiten sowie der deutschen Auswärtigen Kulturpolitik. *Duisburger Arbeiten zur Sprach- und Kulturwissenschaft*, Bd. 52. Frankfurt: Peter Lang.
- Buschfeld, Sarah & Kautzsch, Alexander. 2014. English in Namibia: A first approach. *English World-Wide* 33(2): 121–160. <https://doi.org/10.1075/eww.35.2.01bus>
- Buschfeld, Sarah & Kautzsch, Alexander. 2017. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36(1): 104–126. <https://doi.org/10.1111/weng.12203>
- CIA. 2017. *The World Factbook. Namibia*. <<https://www.cia.gov/library/publications/the-worldfactbook/geos/wa.html>> (2 August 2017).
- Crane, Thera, Lindgren-Streicher, Karl & Wingo, Andy. 2004. *Hai ti! A Beginner's Guide to Oshikwanyama*. <<https://wingolog.org/pub/hai-ti/hai-ti.pdf>> (2 August 2017).

- Davies, Mark. 2013a. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). <<http://corpus.byu.edu/glowbe/>> (2 August 2017).
- Davies, Mark. 2013b. Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. <<http://corpus.byu.edu/now/>> (2 August 2017).
- de Klerk, Vivian & Gough, David. 2002. Black South African English. In *Language in South Africa*, Rajend Mesthrie (ed.), 356–378. Cambridge: CUP.
<https://doi.org/10.1017/CBO9780511486692.019>
- Donaldson, Bruce C. 1993. *A Grammar of Afrikaans*. Berlin: Mouton de Gruyter.
<https://doi.org/10.1515/9783110863154>
- Edwards, Alison. 2016. *English in the Netherlands. Functions, Forms and Attitudes* [Varieties of English Around the World G56]. Amsterdam: John Benjamins.
<https://doi.org/10.1075/veaw.g56>
- Friedrich-Ebert-Stiftung. 2015. *African Media Barometer. The First Home Grown Analysis of the Media Landscape in Africa. Namibia 2015*. Windhoek: Fesmedia Africa.
- Frydman, Jenna. 2011. A critical analysis of Namibia's English-only language policy. In *Selected Proceedings of the 40th Annual Conference on African Linguistics: African Languages and Linguistics Today*, Eyamba G. Bokamba, Ryan K. Shosted & Bezza Tesfaw Ayalew (eds), 178–189. Somerville MA: Cascadia Proceedings Project. <<http://www.lingref.com/cpp/acial/40/paper2574.pdf>> (2 August 2017).
- Gries, Stefan T. & Bernaisch, Tobias. 2016. Exploring epicentres empirically: Focus on South Asian Englishes. *English World-Wide* 37(1): 1–25. <https://doi.org/10.1075/eww.37.1.01gri>
- Groves, Julie. 2010. Error or feature? The issue of interlanguage and deviations in non-native varieties of English. *HKBU Papers in Applied Language Studies* 14: 108–129.
- Harlech-Jones, Brian. 1995. Language policy and language planning in Namibia. In *Discrimination Through Language in Africa? Perspectives on the Namibian Experience*, Martin Pütz (ed.), 181–206. Berlin: Mouton de Gruyter.
- Harlech-Jones, Brian. 2006. Why the future is illiterate: Where did education reforms go wrong? *Insight Namibia* July, 30–31.
- Kamwangamalu, Nkonko M. 2013. Multilingualism in Southern Africa. In *The Handbook of Bilingualism and Multilingualism*, Tej K. Bhatia & William C. Ritchie (eds), 791–812. Malden, MA: Wiley-Blackwell.
- Kautzsch, Alexander. Forthcoming. English in Germany and Namibia: The effects of extra- and intra-territorial forces.
- Kautzsch, Alexander & Schröder, Anne. 2016. English in multilingual and multiethnic Namibia: Some evidence on language attitudes and on the pronunciation of vowels. In *Anglistentag 2015 Paderborn, Proceedings*, Christoph Ehland, Ilka Mindt & Merle Tönnies (eds), 277–288. Trier: Wiss. Verlag Trier.
- Kautzsch, Alexander, Schröder, Anne & Zähres, Frederic. In preparation. Ethnic variation in Namibian English: A first approach to Baster English.
- Klein, Norbert. 1984. *Deutsche Sprache im Kontakt in Südwestafrika. Der heutige Gebrauch der Sprachen Deutsch, Afrikaans und Englisch in Namibia*. Wiesbaden: Steiner.
- Larsen, Martin Buch. 2007. *The Media Environment in Namibia, 1990–2007*. Windhoek: Media Institute of Southern Africa Namibia (MISA).
- Maho, Jouni Filip. 1998. *Few People, Many Tongues. The Languages of Namibia*. Windhoek: Gamsberg Macmillan.

- McCormick, Kay. 2002. Code-switching, mixing and convergence in Cape Town. In *Language in South Africa*, Rajend Mesthrie (ed.), 216–234. Cambridge: CUP.
<https://doi.org/10.1017/CBO9780511486692.012>
- Mesthrie, Rajend (ed.). 2002. *Language in South Africa*. Cambridge: CUP.
<https://doi.org/10.1017/CBO9780511486692>
- Mukherjee, Joybrato. 2007. Structural nativisation in Indian English: Exploring the lexis-grammar interface. In *Rainbow of Linguistics*, Niladn Sekhar Dash, Probal Dasgupta & Pabitra Sarkar (eds), 98–116. Calcutta: T. Media Publication.
- Mungungu, Saara Sirkka. 2010. Error Analysis: Investigating the Writing of ESL Namibian Learners. MA thesis, University of South Africa. <<http://hdl.handle.net/10500/4893>> (2 August 2017).
- Namibia Statistics Agency. 2011. *Namibia 2011 Population and Housing Census Main Report*. <<http://cms.my.na/assets/documents/p19dmn58guram30ttun89rdrp1.pdf>> (2 August 2017).
- Nesselhauf, Nadja. 2011. Exploring the phraseology of ESL and EFL varieties. In *The Phraseological View of Language: A Tribute to John Sinclair*, Thomas Herbst, Susen Faulhaber & Peter Uhrig (eds), 159–177. Berlin: Mouton de Gruyter.
- NIDS-Namibia Inter-censal Demographic Survey 2006. <<http://cms.my.na/assets/documents/p19dmqga41rhl5rk6m1qi41cfa1.pdf>> (2 August 2017).
- Nghikembua, Anneli. 2015. Error Analysis in a Learner Corpus: A Study of Errors amongst Grade 12 Oshiwambo Speaking Learners of English in Northern Namibia. MA thesis, Rhodes University, South Africa. <<http://hdl.handle.net/10962/d1018911>> (2 August 2017).
- Nurse, Derek. 2008. *Tense and Aspect in Bantu*. Oxford: OUP.
- Pütz, Martin (ed). 1995. *Discrimination Through Language in Africa? Perspectives on the Namibian Experience*. Berlin: Mouton De Gruyter. <https://doi.org/10.1515/9783110906677>
- Riddle, Charles. 1989. A profile of Namibian media: The censored debate. *International Communication Gazette* 44(1): 45–55. <https://doi.org/10.1177/001654928904400103>
- Roche, Xavier & other contributors. 2017. HTTrack. Website Copier. (Software). <<https://www.httrack.com/>> (2 August 2017).
- Rothe, Andreas. 2010. *Media System and News Selection in Namibia* [Lingener Lektüren zum Kommunikations-Management 1]. Münster: LIT Verlag.
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: CUP.
<https://doi.org/10.1017/CBO9780511618901>
- Schneider, Klaus P. & Schröder, Anne. In preparation. Variational pragmatics and the specificity of English in Namibia.
- Simons, Gary F. & Charles D. Fennig (eds). 2017. *Ethnologue: Languages of the World*, 20th edn. Dallas, Texas: SIL International. <<http://www.ethnologue.com>> (20 October 2017).
- Stell, Gerald. 2014a. Social identities in post-Apartheid intergroup communication patterns: Linguistic evidence of an emergent non-white pan-ethnicity in Namibia? *International Journal of the Sociology of Language* 230: 91–114.
- Stell, Gerald. 2014b. Uses and functions of English in Namibia's multiethnic settings. *World Englishes* 33 (2): 223–241. <https://doi.org/10.1111/weng.12082>
- TexTexin. 2011. UTF-8 Encoding Debugging Chart. <<http://www.i18nqa.com/debug/utf8-debug.html>> (2 August 2017).
- Tonchi, Victor L., Lindeke, William A. & Grotpeter, John J. 2012. *Historical Dictionary of Namibia*, 2nd edn. Lanham MD: The Scarecrow Press.

- Tötemeyer, Andree-Jeanne. 2010. Multilingualism and the language policy for Namibian schools. *PRAESA Occasional Papers* No. 37, <<http://www.praesa.org.za/files/2012/07/Paper37.pdf>> (2 August 2017).
- Trudgill, Peter & Hannah, Jean. 2002. *International English: A Guide to the Varieties of Standard English*. London: Arnold.
- United Nations Department of Economic and Social Affairs. 2011. *Demographic Yearbook 2009–2010*. <<http://unstats.un.org/unsd/demographic/products/dyb/dybsets/2009-2010.pdf>> (2 August 2017).
- Wallace, Marion & Kinahan, John. 2011. *A History of Namibia*. London: C. Hurst & Co.
- Zähres, Frederic. 2016. Vowels in Namibian English: Investigating the NURSE-WORK Vowel Split. MA thesis, Bielefeld University.

Lexical expansion in Ghanaian English from a diachronic perspective

A structural and semantic analysis

Thorsten Brato

University of Regensburg

This chapter provides a real-time structural and semantic analysis of lexical expansion in the Nativization phase of Ghanaian English based on the Historical Corpus of English in Ghana (HiCE Ghana) and the written-printed sections of the Ghanaian component of ICE. Taking a comprehensive list of previously attested ‘Ghanaianisms’ – innovative lexical items of English and local origin – as a starting point, the paper shows that traditional word-formation processes like derivation or compounding play only a subordinate role and that semantic shift is the most important process used in both periods. While the corpora are comparatively small for lexical research, the results still provide a useful starting point to better our understanding of how Ghanaian English has evolved over the past 40 years.

1. Introduction

Besides accent, lexical variation is one of the most conspicuous features of Englishes around the world, often leaving those not accustomed to the culture and customs of the country or region in which the variety is spoken largely clueless. A good example is provided by the following excerpt from a news report from Ghana:

- (1) NANA Kweku Kwaning III, Benkumhene of Akim Bosome at the week-end swore the oath of allegiance to Nana Oware Agyekum III, Omanhene of Akim Bosome Traditional Area at a ceremony held at the traditional council conference hall Akim Swedru.

Nana Kwaming, who is also Akim Afoase-Kesehene, succeeded his younger brother Nana Kofi Karikari, who was customarily destooled by his king-makers nine months ago. (HiCE-PrReg039)

This short piece shows quite a remarkable array of innovative lexical items and processes, including borrowings from a local language, such as *Omanheme* – the paramount of the Akan, derivations, for example *destool* – ‘dethrone’ and compounds, as in *king-makers*, a group responsible for choosing a new chief.

Lexical expansion is also at the heart of Schneider’s (2007) Dynamic Model of the evolution of Postcolonial Englishes. He proposes that as a variety evolves, lexical innovations are drawn from different semantic domains at different points in time and that different word-formation processes are at work. While there is a good body of research describing lexical innovations from a synchronic perspective (e.g. Bamiro 1997; Baumgardner 1998; Alabi 2000; Tent 2001; Biermeier 2008; van Rooy & Terblanche 2010), only few studies have so far taken a diachronic approach (Lambert 2014; Evans 2015) and there is as yet no empirical corpus-linguistic study on the diachronic development of lexis in a New English. This paper attempts to fill this gap by showing how Ghanaian English lexis has changed over a period of about 40 years. Comparing data from the Historical Corpus of English in Ghana (HiCE Ghana, see Brato this volume for a more detailed discussion), a 600,000-word corpus of written Ghanaian English from the years 1966–1975 and the written-printed sections of the Ghanaian component of the International Corpus of English (ICE Ghana; Huber & Dako 2013), the aim of the paper is to provide a structural and semantic analysis of old and new “Ghanaianisms”, that is innovative lexical items of English and local origin. The study is based on a comprehensive list of previously attested Ghanaianisms discussed by various authors (Sey 1973; Kirby 1998; Dako 2001, 2003; Blench 2006), items noted down as potential new lexemes or collocational constructions during the compilation of HiCE Ghana and a systematic study of a range of typical derivational affixes.

The following section will provide some theoretical background to the study and briefly summarise the key points on lexis in the Dynamic Model. Section 3 introduces the previous comprehensive compilations and analyses of Ghanaianisms. In Section 4, I will outline the methodology employed before turning to the findings section, which will provide a structural and semantic classification of Ghanaian English lexis and show how deeply entrenched these items are. The paper will be rounded off with a conclusion addressing some of the concerns raised.

2. Background: Lexis in the dynamic model

Schneider’s (2007) Dynamic Model is one of the most advanced models accounting for the evolution of English in colonial and postcolonial contexts. One of its major qualities is that it can account for developments both in native and non-native contexts at the same time. Schneider posits that there are two strands, the settlers

(STL), who speak English as their mother tongue, and the indigenous (IDG) people for whom English in the initial stages is a foreign language. The main hypothesis in Schneider's model is that as these two groups get into contact more closely over the course of time, the English language will undergo predictable changes, the result of which is a new variety of English. There are five phases: foundation, exonormative stabilisation, nativization, endonormative stabilisation, and differentiation.

The historical developments of English in Ghana and some of the challenges in applying Schneider's model, most notably the absence of a sizable STL strand, are discussed in greater detail in Brato (this volume). At this point I will therefore just provide the key dates and focus on the lexical aspects of the Dynamic Model. The foundation phase began in 1632 with the first permanent British settlement on the coast close to today's city of Cape Coast. 1844 marks the onset of the exonormative stabilization phase. In this year, some of the local chiefs yielded some legal power to the British. Colonisation proper began some 30 years later when the coastal strip was declared the Colony of the Gold Coast. In 1901, Ashantiland and the Northern Territories were integrated in the colony. Ghana became independent in 1957, which is also the onset of the nativization phase. Currently, Ghana falls between phases 3 and 4 (Huber 2014: 87–90).

Each of the phases is characterised by certain aspects of variation and innovation on all linguistic levels (Schneider 2007: 21–55). As this paper is concerned mainly with lexis I will point out some of the key issues regarding this feature. In the foundation phase, contact between the STL and IDG strand is very limited. There is the need to communicate mainly for trade and therefore is only very little influence of the local languages on English with one major exception, the borrowing of place names and other toponymic terms. In the present study, place names of local or colonial origin such as names of cities or their boroughs (e.g. Kumasi, Cape Coast, Osu, Labone) were excluded from the analysis. Place names that resulted from a word-formation process (such as the semantic shift in *Airport*, describing a neighbourhood of Accra) were included.

During phase 2, lexical borrowing continues and strengthens as it no longer only words for toponyms which are integrated into the English language but also meaningful words, such as names for the local flora and fauna. This is followed by items describing cultural practices, clothing or customs typical in the IDG community.¹

In phase 3, the increased contact of the two strands leads to heavy lexical borrowing from the local languages. Additional items from the cultural domain are borrowed, but also from other areas. In this context, it will be interesting to see how

1. Schneider (2007: 39, fn. 10) claims that in exploitation colonies, such as Ghana, cultural items like food, clothing or customs are more readily borrowed.

the near-absence of the STL strand has affected borrowings into Ghanaian English. As there is little need to pay attention to the native English speakers, as would be the case in a settlement colony, it is not unlikely that even heavier lexical borrowing occurs. A second set of innovations is found at the lexis-grammar interface. Individual words may receive particular complementation patterns or begin to co-occur in larger units or chunks which results in local idioms or phraseological units. These include fixed phrases such as *Please allow me a space*, frequently found as an opener in letters to the editor, new word-formation products resulting from usually just a small number of different processes, namely derivation and compounding (either within English or by combining an English and local item), whereas shortenings are much less common. Located rather on the fringes of lexis are innovations relating to prepositional usage and verb complementation. Both are included in this study. A final set of processes concerns semantic shift, whereby an already existing word takes on a different meaning.

As Ghanaian English falls between the nativization and endonormative stabilization phases it is important to point out that the major difference between the two phases is not so much further innovation, but rather codification and acceptance of the local linguistic norms. One important step is the creation of a dictionary so that local usages are documented and may therefore become more accepted. As of today, there is no 'official' dictionary of Ghanaian English but scholars (Kirby 1998; Dako 2003; Blench 2006) have published a guide, glossary and dictionary respectively.

3. Previous studies on the lexis of Ghanaian English

Typical lexical elements of Ghanaian English are nowadays, commonly referred to as *Ghanaianisms*, defined by Dako (2001: 23) as

lexical items of both English and local origin that are peculiar to the English spoken and written in Ghana. [...] Ghanaianisms have their origin in Ghana, formerly the Gold Coast and the Northern Territories, and are peculiar to Ghanaian usage.

Several scholars have addressed this issue, ranging from smaller (sections of) articles and collections of lexical and lexicogrammatical innovations to dictionaries and book-length treatments including examples from a range of sources. Four works are of particular importance to the current study, as they form the basis against which the Ghanaianisms found in the two corpora were compared. They are outlined in the following.

The first major work is Sey (1973). He describes differences regarding idioms and idiomatic expressions, including prepositional and phrasal verbs, borrowings from the first languages and loan translations. He also provides a list of about

300 items of Ghanaianisms of English origin, which are “still in use today, but with changed frequency” (Dako 2001: 27). In this paper, Dako provides a semantic and structural analysis of Ghanaian Englishes and suggests that about 60% of Ghanaianisms are of English origin, 30% are of local origin and the remaining 10% are made up of hybrids, i.e. they combine two or more languages, e.g. in compounds like *kente cloth*, multicoloured locally woven strips of cloth sewn together. Dako (2003) is a glossary of Ghanaianisms and lists about 2,500 items from all semantic domains and from a range of word-formation processes. There are two additional fairly comprehensive dictionaries by Kirby (1998) and Blench (2006). Kirby’s book is aimed at visitors to Ghana and provides not only items considered Ghanaianisms in the strict sense, but also lists pidgin words and gives hints on various non-linguistic matters. The latter were not included in the analysis. Blench’s dictionary is fairly comprehensive and lists the Ghanaianism, a part-of-speech tag, a gloss and for some of the items also provides examples and the etymology. Between the four major works (Sey 1973; Kirby 1998; Dako 2003; Blench 2006) approximately 3970 different Ghanaianisms have been attested.

In addition to that, several other scholars have commented on various elements of typical Ghanaian vocabulary. Very prescriptive takes on typical features of Ghanaian (Gold Coast) English, including both grammatical and lexical items come from Brown & Scragg (1950), Gyasi (1991), Ahulu (1994) and Nimako (2004). Denteh (1967: 51–56) provides some loan translations, ‘common errors’ and pidgin items, Mensah (1971) discusses various aspects of Ghanaian English newspaper writing and complains about the “linguistic showmanship” (p. 9), a tendency to build overly complex sentences and make extensive use of idiomatic constructions.² Tingley (1981) mainly focuses on grammatical deviances from the British standard and provides a few examples of Ghanaianisms using typical word-formation processes. The focus of Bamiro (1997) is on lexical innovations in fiction. He provides a small sample of items, most of which are also covered in the major works.

As the works mentioned in the previous section did not provide additional items or included pidgin vocabulary, they were not included in this part of the larger study.

2. He is not the first to point this out. There are numerous comments in the education reports from the colonial government, such as this: “An improvement is noted, but teachers and pupils need to realise that long words, flowery diction and hackneyed quotation do not constitute good writing.” (Gold Coast Colony 1930: 32)

4. Methodology

Two corpora of Ghanaian English from the onset and end of the nativization phase were analysed in this study: The Historical Corpus of English in Ghana (HiCE Ghana) and the Ghanaian component of ICE. HiCE Ghana covers the period 1966–1975 and consists of about 600,000 words of written (printed) Ghanaian English. Its basic structure follows that of the relevant sections of ICE corpora (W2A-W2F). Modifications were made regarding the sizes of each of the genres; some categories have a more fine-grained division than in the original ICE. Also, the corpus contains letters to the editors, a category not found in ICE. To ensure some comparability, ICE Ghana was supplemented with about 10,000 words of this genre (see Brato this volume for a more detailed discussion about the theoretical and methodological considerations that have guided the creation of HiCE Ghana).

The two corpora were part-of-speech (POS) tagged in Treetagger (Schmid 1994) and regular expressions (see the Appendix for some examples) were used to find all inflected forms of the previously attested Ghanaianisms. Potentially new lexical items and other linguistic features such as uncommon idiomatic usages or fixed phrases were noted down in the compilation of HiCE Ghana. In addition, a set of regular expressions was used to uncover innovative derivations in the two corpora.

As a quantitative measure of overall productivity differences between the corpora, I used the measure of ‘potential productivity’ – $P = n_1/N'$ – as suggested by Baayen and Lieber (1991) regarding the productivity of morphological affixes in word-formation. In the context of this study it was implemented as the quotient of the number of *hapax legomena* (i.e. Ghanaianisms occurring only once – n_1) of Ghanaianisms by the overall number (i.e. tokens) of Ghanaianisms in the corpus (N'). It describes the likelihood of the rate at which new types are expected to show up in the corpus when N tokens have been sampled. As this value is dependent on corpus size it cannot be used meaningfully to compare data from (sub-)corpora of different sizes. To overcome these shortcomings, Baayen (2001) proposes so-called Large Number of Rare Events (LNRE) models, which allow to extrapolate expected values for hapax legomena for arbitrary values of token numbers. They are implemented in the *ZipfR* package (Evert & Baroni 2007) for R and were used to compare the overall difference in productivity between the two corpora. In the context of the of the current study I used the Zipf-Mandelbrot implementation of LNRE to estimate the expected values of types (EV), i.e. distinct Ghanaianisms and expected values of hapax legomena (EV1) for $N = 10,000$, a number set arbitrarily to represent equal token numbers (the number of all Ghanaianisms) for the two corpora. The formula used to determine the potential productivity therefore is $P = EV1/10,000$.

The focus of the paper, however, is on a descriptive qualitative and quantitative analysis of Ghanaianisms on four levels: word class, origin, word-formation process and semantic domain. In a first step, the items were grouped according to their

origin: English, Ghanaian (with subdivisions into the local languages where this information was available), West African (mainly Nigerian), foreign and hybrids, a combination of items of two or more languages.

For English items I used the taxonomy proposed by Quirk et al. (1985: 1515–85) as a starting point. It includes affixation (prefixation and suffixation), compounding, conversion and shortenings (back-formation, abbreviation, clipping, acronym, blend). In addition, I coded for semantic shift, understood here as an already existing word taking on a new (locally confined) meaning.³ Three main types can be distinguished: an already existing word or compound takes on an additional connotation (semantic widening), the meaning becomes more specific than in general usage (semantic narrowing) or that there is no (or only very little) relationship between the new and the original meanings. This is referred to in this paper as general semantic shift. Reiterations, i.e. the repetition of a word (as in *big big* = very big) were far and few but it was felt that they should be included. A category located at the lexis-grammar interface and not so much a word-formation process as such are multiword expressions. These include idiomatic formations and items with a collocational preference and innovative phrasal and prepositional verbs (see Figure 1 for an overview of categories with subcategories).

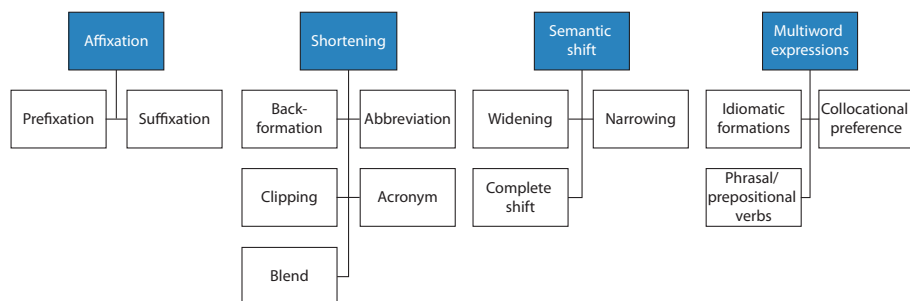


Figure 1. Word-formation processes of English items

Items with their origin in one or more Ghanaian languages were initially classified using the following criteria: affixation, borrowing, compounding, place names, proper names. These were later combined under the heading borrowings. Items of foreign origin were all borrowings. West African items are almost exclusively made up of borrowings from Hausa. A few items are also place and proper names. The hybrids were classified according to their source languages and word-formation process (affixation and compounding).

3. Strictly speaking, semantic shift is not a word-formation process as there is neither a formal change, nor a change in word class. However, as it adds to the word stock specific to Ghanaian English it was included here.

In a final step, I coded the semantic domain of each Ghanaianism. As mentioned above, the Dynamic Model assumes that in the initial stages local innovations are mainly drawn from a small range of semantic domains, such as places names, flora and fauna and cultural items and processes. In later stages, Ghanaianisms should therefore be found in a much wider range of domains. All Ghanaianisms (items of non-English origin were translated or glossed) were subjected to the UCREL Semantic Analysis System (USAS; Archer, Wilson, and Rayson 2002) for semantic tagging. USAS is a system for the automatic semantic annotation of texts. It provides 21 major semantic fields and also allows for more fine-grained analyses. For example, category *B – The body and the individual* is subdivided into five sub-categories *anatomy and physiology, health and disease, medicines and medical treatment, cleaning and personal care and clothes and personal belongings*. The USAS output was checked and corrected manually for miscategorisations and cases in which the system suggested several tags. All items were categorised down to the second level of granularity. They were conflated to avoid too many small cells.

In their study on word-formation processes in Black South African English van Rooy and Terblanche (2010: 411) suggest that apart from codification two additional criteria should be considered in judging if a variety has progressed to the endonormative stabilisation phase: generality and acceptability. *Generality* describes the “range of semantic domains in which innovations occur”. The more different lexical fields are affected the more advanced is a variety. This can be measured relatively straightforwardly using the semantic classification described in the previous paragraph. The data from ICE Ghana should potentially be more varied than that of HiCE Ghana, i.e. we should find more innovations from a wider range of lexical fields. With the current data, one major caveat is the different sizes of the corpora. Being almost twice as large as ICE Ghana makes it more likely to find innovative items in HiCE Ghana. *Acceptability* measures the usage of an innovation “beyond the innovating segment of the community.” In the South African context, this meant that the innovations that were found in a newspaper corpus directed at the IDG strand of the population were also searched for in a comparable corpus aimed at readers from the STL strand. If an innovation was found in there as well it was assumed to have gained wider acceptance in the community.

This kind of comparison is not meaningful within the Ghanaian context. As mentioned above there is virtually no STL strand and there never has been one. Therefore, I suggest that in the Ghanaian context acceptability should be measured by looking at the distribution of a Ghanaianism across the different genres, which can be grouped according to levels of formality.

Table 1 shows the formality scale of the different genres. Similarly to Fuchs and Gut's (2015: 378) classification in the context of ICE Nigeria, the scale is based mainly on my intuitions from the compilation of HiCE Ghana “rather than measurements of linguistic features, since it has not yet been established which linguistic

features correlate with formality in New Englishes.” As all data was published, the *informal* style was not attested. Novels and stories are classified as *rather informal* because they contain large parts representing speech in which more colloquial language can be expected. Texts from the skills and hobbies section and letters to the editor were classified as intermediate. The writing style in the skills & hobbies section is varied, ranging from short pieces written in a rather informal style aimed at a wider audience, such as fashion tips or Q and A sections of magazines and journals to those which use more formulaic language such as recipes. Most of the letters to the editor are rather short in length and despite following some basic formalities such as salutations, the writing style of the main text is often rather informal. Popular writing, news reports and press editorials were classified as rather formal. They all use a range of formulaic sequences and – at least impressionistically – the syntax is more complex than in the categories labelled intermediate. Academic and administrative writing shows the most formal writing style.

Table 1. Degree of formality of text categories in HiCE Ghana and ICE Ghana

1	Informal	n. a.
2	Rather informal	Novels, Stories
3	Intermediate	Skills & Hobbies, Letters to the editor
4	Rather formal	Popular writing; News reports, Press editorials
5	Formal	Academic writing, Administrative Writing

5. Findings

This section first provides a general summary of the main findings of the distribution of Ghanaianisms across the two corpora. I will then provide a structural analysis, showing which word-formation processes are most commonly used in the expansion of the vocabulary. In the third part I will discuss the distribution of Ghanaianisms across semantic domains and highlight how this varies and what the underlying reasons. The final part will address the question of how entrenched lexical innovations are, by looking at some higher-frequency items.

5.1 General findings

As mentioned in the previous section, different methods were employed to analyse the data quantitatively. I will first briefly outline the findings using Baayen & Lieber’s (1991) and Baayen’s (2001) productivity measurements before turning to a more fine-grained discussion of based on relative frequencies in a second step. As can be seen from Table 2, the application of the Baayen & Lieber formula shows no

difference in the relation of hapax legomina to types, which stands at about 41% in both corpora. This is in stark contrast to the relation between hapaxes and tokens, which underlies this form of the productivity formula. Here, the data suggests that the productivity is considerably higher in ICE (0.1) than HiCE (0.06).

Table 2. Uncorrected (raw) potential productivity rate by corpus

Corpus	Types	Tokens (N')	Hapax legomena (n_1)	<i>P</i>
HiCE Ghana	923	5,777	374	0.06
ICE Ghana	504	2,075	208	0.10

However, once the corrected version of *P* is calculated, this picture is – at first rather counterintuitively – reversed, i.e. we now find that the value for HiCE remains at approximately 0.06, while the value for ICE has dropped to only 0.03 (see Table 3). This would suggest the early nativization phase was characterised by large-scale lexical expansion that has slowed down considerably over time. This could potentially be explained by a greater need or will to ‘indigenise’ English after the colonial ties were finally capped. This would also go hand-in-hand with Schneider’s (2007: 46; emphasis added) claim that

[in] descriptive terms, it is interesting to observe that in its *early stages* this indigenization of language structure mostly occurs at the interface between grammar and lexis, affecting the syntactic behaviour of certain lexical elements. Individual words, typically high-frequency items, adopt characteristic but marked usage and complementation patterns.

Table 3. Corrected potential productivity rate based on a potential sample of 10,000 Ghanaianisms

Corpus	Expected types	Expected hapax legomena (n_1)	<i>P</i> (corrected)
HiCE Ghana	1,219	614	0.06
ICE Ghana	936	343	0.03

Another potential factor is the education expansion in the 1950s following the 1951 Accelerated Development Plan for Education, which saw numbers of primary school pupils increase significantly so that there was a severe shortage of qualified teachers. For this reason primary school leavers were employed to teach, which led to complaints about rapidly falling standards of proficiency in English (cf. Huber 2004: 844). Together with the absence of a norm-providing (and -controlling) STL strand this could explain why writers represented in HiCE might show a higher rate of idiosyncrasies, which have since then developed into Ghanaianisms. However,

there is another – more likely explanation – that lies in the structure and contents of the two corpora, which will be discussed in the following paragraphs before briefly returning the issue of productivity in Baayen's terms.

Table 4 shows that 1,078 distinct Ghanaianisms were attested. This accounts for about 11.9 types per ten thousand words (pttw). The token frequency is considerably higher – frequency at 7,852 (or 86.3 pttw). More than 85% of all types were attested in HiCE; the figure for ICE is much lower at about 47%. Of course, we must keep in mind that the corpora are of different sizes and therefore it is more likely to find a Ghanaianism in the HiCE data. In fact, there are many items which are frequently used today, which just did not happen to occur in ICE such as *enstool* 'enthroned a chief' or *gate fee* 'entrance fee'. A quick search in GloWbE shows 78 hits for the former and 19 hits for the latter.⁴

Table 4. Number of types and tokens of Ghanaianisms (pttw) in the two corpora

Corpus	All Ghanaianisms		Not previously attested	
	Types	Tokens	Types	Tokens
HiCE	923 (15.4)	5,777 (96.3)	234 (3.9)	499 (8.3)
ICE	504 (16.3)	2,075 (66.9)	21 (0.7)	73 (2.4)
Total	1078 (11.9)	7852 (86.3)	239 (2.6)	572 (6.3)

Of the 1,078 types, 239 (about 22.2%) had not been documented in previous collections of Ghanaianisms. It is quite striking that 234 of these were attested in HiCE and only 21 in ICE. This is partly due to the topics covered in the two corpora. Many of the texts in the humanities and social sciences sections of HiCE discuss items of Ghanaian culture or social practice, such as religious festivals, dances, or family structure. It is here that we commonly find a reference in a Ghanaian language to signify the item as in (2).

- (2) These include such instruments as the Akan "*prempensua*"⁵ which consist of five wooden tongues mounted on a flat wooden board, the tongues, graduated in tone, being plucked by the player. (HiCE-PopHum010)

4. The same pattern is, of course also found in reverse. For example, *thanksgiving service* 'a church service to end the funeral celebrations' is already mentioned by (Sey 1973) but is only found in the ICE data.

5. At this stage, no distinction was made between items that were mentioned only, i.e. where the Ghanaian (or foreign) name was only used in the text as a label. Of course, this increases the overall likelihood of these items and of Ghanaianisms as a whole.

Sometimes, there even is an English gloss, which may also be a Ghanaianism. The following example, drawn from a story, illustrates this point nicely. The Ewe term for *chief* (Togbi) is translated as *stool leader*:

(3) She went to see *Togbi* Dogbo, the *stool leader*. (HiCE-Story026)

Table 5. Language of origin

Language	Corpus	All Ghanaianisms		Not previously attested	
		Types	Tokens	Types	Tokens
English	HiCE	549 (9.2)	3,844 (64.1)	42 (0.7)	185 (3.1)
	ICE	381 (6.4)	1,378 (23)	15 (0.3)	62 (1)
Ghanaian	HiCE	321 (5.4)	1,671 (27.9)	186 (3.1)	307 (501)
	ICE	96 (1.6)	578 (9.6)	6 (0.1)	11 (0.2)
Hybrid	HiCE	24 (0.4)	67 (1.1)	4 (0.1)	5 (0.1)
	ICE	13 (0)	71 (0)		
other	HiCE	29 (0.5)	195 (3.3)	2 (0)	2 (0)
	ICE	14 (0.2)	48 (0.8)		

Table 5 shows the distribution of Ghanaianisms by their source language. English items (including calques) make up about 60% of types and two thirds of all tokens in HiCE. In ICE, English types are much more common (about 76%) but are found similarly frequent regarding tokens. This in turn bears directly on the number of borrowings from a Ghanaian language, which are about twice as likely in the older data (34.8%) than in ICE (19%). The other two sources, hybrids and borrowings from other language only play marginal roles and account for 2.6 to 3.1% in both corpora. By far the largest number of new Ghanaianisms are borrowings from a Ghanaian language, thus suggesting that “heavy lexical borrowing” (Schneider 2007: 44) is indeed a very widespread and common process in the early nativization phase that gradually decreases over time as English takes over. They account for about 80% of all types and about 54% of all tokens. However, of the 189 newly-attested borrowings 123 are hapax legomena and a mere 13 are found in more than one text. These include:

awoamefia (noun) *Ewe chief*: Mr Francis Nyonyo Agboada, alias Togbui Sri III, whose purported installation as the *awoamefia* of Anlo stirred a bloody civil strife in the area, has finally renounced his claim to the Anlo Stool. (ICE-W2C-009)
Kple (noun) *fetish dance*: These are a combination of secular dance movements and the gestures are from the religious dances like *Kple* and *La kpa* linked with lively clowning for the fun of it. (HiCE-AcHum013)

Tarana (noun) *a state official assisting the nayiri [supreme ruler] of the Mamprusi:*
 He was assisted by a council of state elders in which sat officials like the *Tarana*,
 the *Wudana*, the *Kpanara* and the *Nazo*. (HiCE-PopSocSci012)

English items make up about 19% of the types, but have a remarkably high token frequency (about 44%). Hybrids contribute only insignificantly. There are four types and five tokens. Examples for both are discussed in the following section.

In the previous paragraphs I have shown that the earlier data is characterised very strongly by a large number of hapax legomina of Ghanaian origin, as authors describe and discuss Ghanaian musical instruments, dances or religious and social concepts. This, of course, bears directly on the potential productivity rate, which therefore is so much larger in HiCE. Taking into account only the English language items (see Table 7) shows that now the productivity for a potential sample of 10,000 Ghanaianisms is virtually the same.

Table 6. Corrected potential productivity rate based on a potential sample of 10,000 Ghanaianisms (English items only)

Corpus	Expected types	Expected hapax legomena (n_1)	P (corrected)
HiCE Ghana	813	326	0.03
ICE Ghana	872	338	0.03

5.2 Word-formation analysis

As outlined above, the English items were categorised formally on the basis of Quirk et al.'s (1985) classification of word-formation processes supplemented mainly by semantic shift and multiword expressions. The traditional word-formation processes only play a very minor role in Ghanaian English both regarding types and tokens. An overview is shown in Table 7. As can be seen, as for the newly-attested items, affixation plays only a very subordinate role and accounts for only 6% of the types in HiCE and even less in ICE. The token frequencies are similarly insignificantly small. Compounds are more common. About a quarter of all types in both corpora were created in this way, making up about 19% of tokens in HiCE and 17% in ICE. Multiword expressions come in at about 14% of the type frequencies and 8% and 11% of the tokens respectively. By far the most productive means of expanding the vocabulary is not a word-formation process as such, but semantic shift. About 46% of the types and 66% of the tokens in the historical data and 50% of the types and 61% of the tokens in the current data have been created in this way. Other processes only play a minor role.

Table 7. Word-formation processes (English items)

Word-formation process	Corpus	All Ghanaianisms		Not previously attested	
		Types	Tokens	Types	Tokens
Affixation	HiCE	33 (0.6)	126 (2.1)	1 (0)	1 (0)
	ICE	21 (0.7)	52 (1.7)	1 (0)	1 (0)
Compounding	HiCE	153 (2.6)	722 (12)	14 (0.2)	79 (1.3)
	ICE	90 (2.9)	237 (7.6)	4 (0.1)	11 (0.4)
Semantic shift	HiCE	254 (4.2)	2,526 (42.1)	3 (0.1)	7 (0.1)
	ICE	190 (6.1)	846 (27.3)	1 (0)	1 (0)
Multiword expression	HiCE	29 (0.5)	151 (2.5)	19 (0.3)	91 (1.5)
	ICE	25 (0.8)	88 (2.8)	7 (0.2)	37 (1.2)
Other	HiCE	254 (4.2)	2,526 (42.1)	4 (0.1)	5 (0.1)
	ICE	190 (6.1)	846 (27.3)	2 (0.1)	12 (0.4)

Only three new derivatives were found, all used only by a single writer. In addition, the first item only occurs in the compound *village ariser*.

ariser (noun) *Someone, whose task it is to wake up others*: He was brought back to reality when he heard the village *ariser* summoning all members of the Ogbia group and the fetish priests to meet in the chief's house. (HiCE-Story012)

commandist (adjective) *authoritative*: This is not the way to build a party of opportunity for others with ambition; the approach is *commandist*, it breeds tension and possible fracture and, it offends the dignity and honour of others. (ICE-LetED036)

non-stool (adjective) *not belonging to the stool*: A *non-stool* subject however has the necessary customary rites to follow in order to obtain land. (HiCE-SkiHob021)

Compounding was much more commonly found. 15 new compounds of English origin were attested in the data. They can be sub-classified into three groups: loan translations, a new concept, and a combining form, the latter of which is very productive in Ghanaian English. An example for each is given below.

advance guard (noun) *a divisional chief in Akan warfare; Twafoohene*.⁶ Obiri Yeboa had the scouts, the *advance guard* and the main fighting body (ICE-W2A-006)

6. *-hene* is the Akan affix denoting a chief, e.g. *Omanhene* (paramount chief)

boo boy (noun) *young man engaged to boo at political opponent*: The Campaign was marked by the resurgence of *Boo Boys*, organised by PP activists, to hoot and boo at leaders of the National Alliance of Liberals and to insult them in the most indignified language possible. (HiCE-PressEd025)

ill- (combining form) *wrong/bad*: But as *ill-luck* would have it, my mother took the cue and jumped on the subject. (HiCE-Story004)

Conversion is not very productive in Ghanaian English. Only two innovations were found.

booze s.o. (conversion) *provide drinks for s.o.*: And thank you for *boozing us*. (HiCE-Novel020)

jungle (adjective) *uncivilised*: THE University of Science and Technology branch of the Progress Club, on Thursday called on the National Union of Ghana Students (NUGS) to concentrate on matters affecting students' welfare and not to drag the NUGS into "*jungle*" national politics. (HiCE-PrPol112)

One important element in the nativization of English is the preference for certain words to collocate. These are referred to as *idiomatic formations* by Dako (2001, 2003). In this paper, I refer to these constructions as *multiword expressions* and subdivide them into idiomatic formations and phrasal and prepositional verbs. For the former it is not so much that these constructions are completely new; it is rather the fact that they are used more frequently in the Ghanaian context. Nine new constructions are documented in the data.

One such example is the construction *condone and connive* (or vice versa) and its various forms, which seems to be a feature almost exclusive to Ghanaian news report or editorial writing. It occurs only four times in HiCE Ghana, where it is found in three items from different newspapers and magazines. There is only a single token in ICE Ghana (part of a quotation). However, the construction is found several times in the *Newspapers on the Web* (NOW, Davies 2016) corpus. Of the 28 hits in NOW, 27 are from Ghana, one from Sri Lanka. It is not found at all in the two major current reference corpora (BNC and COCA) and the TIME corpus. There is one hit in COHA.

condone and connive (idiomatic formation): They became the first club in Ghana to play a match with 13 players and the man who *condoned and connived* at this was recently promoted F.I.F.A. referee, A. Q. Hulede. (HiCE-PrSpo061)

There are several instances of innovative verb complementations. Some of these have not been mentioned by the four works of reference, such as *consider as*, but are commonly found in World Englishes. Several other items rather seem idiosyncrasies of the individual writers than new verb complementations as in this example:

stay as (prepositional verb): *stay, remain*: Do you wish to *stay as* a virgin all your life? (HiCE-Novel006)

Semantic shift is a productive process in Ghanaian English. In that respect, it is quite remarkable that only three items not previously attested were found in the corpora.⁷ These are:

inmate (noun) *any other person in the house; a tenant or neighbour living in the same house*: There are other reasons why the Ghanaian host leads his guest in: he, by word or action, can prepare the other *inmates*, if any, for a welcoming honour due to the guest; for example, the announcement: “Nana Asamasi is here!” (HiCE-SkiHob001)

man (noun) *husband*:⁸ In situations like this, both man and wife would be like two strangers under the same roof. (ICE-W2B-020)

much (+ adjective) (adverb) *very*: I am not *much happy* that it is Eleven Wise who are to welcome Kotoko back into the league. (HiCE-PrSpo038)

5.3 Semantic classification

All Ghanaianisms were subjected to the UCREL Semantic Tagger, which groups items into 21 major semantic domains. The results are summarised in Figure 2,⁹ which shows that Ghanaianisms show up in 20 of these. The only domain in which no Ghanaianism was attested is science & technology. By far the largest number of Ghanaianisms comes from the social domain. This is partly due to its rather fine-grained division, but of course it also covers various important aspects of life. Food & farming items are also very frequent. Again, this can be expected as these items also play a vital role in everyday life. Names & grammatical words is the third-most frequent domain. This includes common appellations such as *nana* or the affix *hene*, as shown in (1) both used in the context of the Akan chieftaincy system. Of the other categories, arts & crafts and entertainment, sports & games

7. On the other hand, these are of course notoriously difficult to uncover as they cannot be found using any systematic or quantitative approach and usually require a lot of insider’s knowledge of the variety so that some may have gone unnoticed in my analysis.

8. As was pointed out quite rightly by one of the reviewers, a likely source for this Ghanaianism is the fixed expression: ‘I now pronounce you man and wife’.

9. To get the full picture including raw frequencies and values for items not previously attested, the reader is referred to Table 10 in the Appendix.

stand out because of their large number of newly attested neologisms, the reasons for which I will address below.

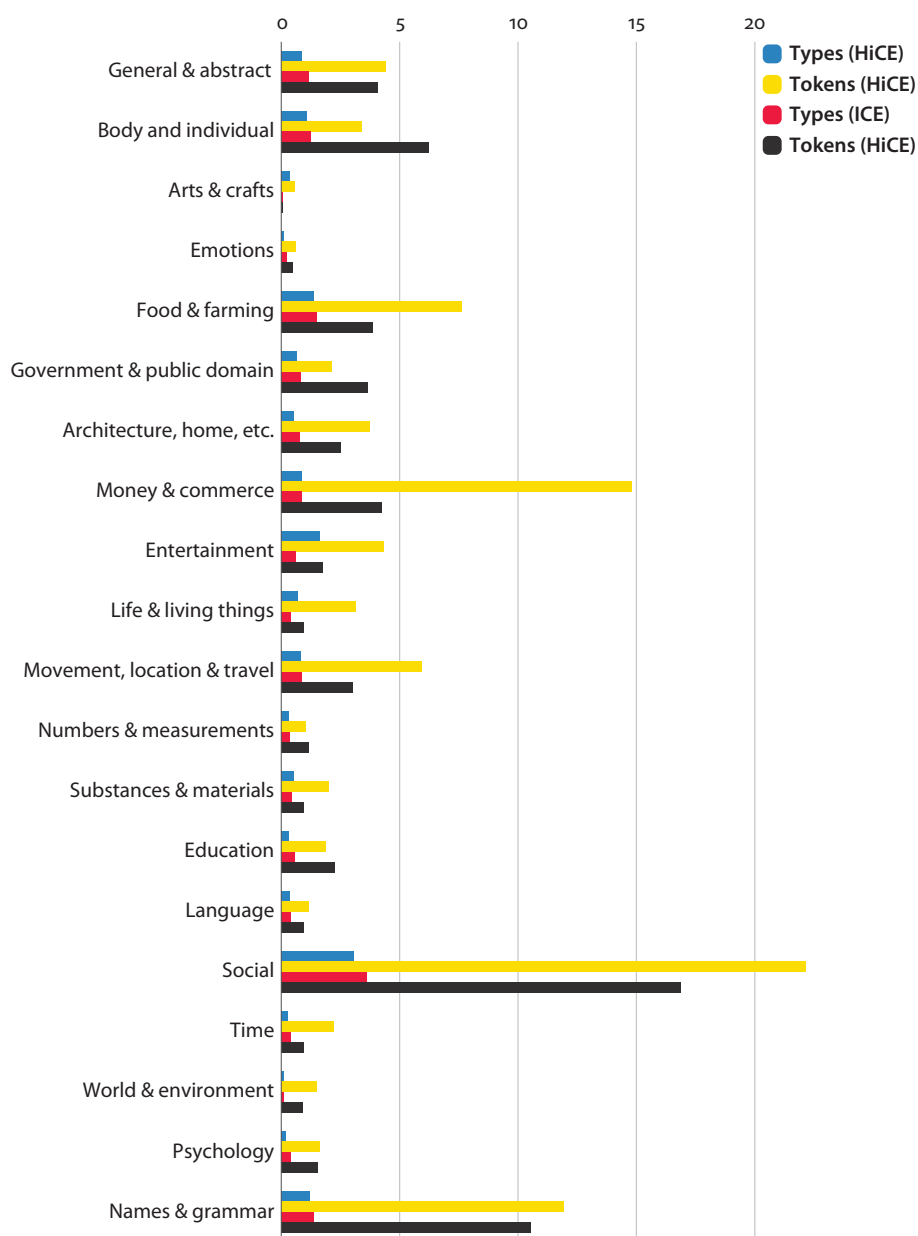


Figure 2. Semantic classification based on the USAS tagset (relative frequencies only)

I will now illustrate a range of newly attested Ghanaianisms from the different domains. Three new types were found in the general category, two (ill- and much + adjective) of which have already been mentioned above. The other construction is a multiword expression, which however is unique but shows another typical feature of Ghanaian English, the absence of the indefinite article:

shake oneself from (idiomatic formation) *get rid of*: ABUAKWA Susu Biribi, who for the past four weeks have been in a fix of uncertainty yesterday *shook themselves* from their slumber to secure 2–1 victory over Kumasi Cornerstone. (HiCE-PrSpo053).

There are eight new items, seven of Ghanaian origin and all only found in HiCE, in section B. They, for example, include the following:

akɔm (noun) *possession (traditional medicine)*: Akesson [...] points out that throughout his training the would-be medicine man is instructed in the art of possession or *akɔm*. (HiCE-PopNatSci006)

doso (noun) *robe of a medicine man*: On that day, he will be asked to dress in his full robe (*doso*), with all charms upon him (HiCE-PopNatSci006)

oboroniwawu/oboroni wawu (noun) *second hand clothes; literally: the white man has died*: SOME police officers of the Teshie Police Station, in Accra, were alleged to have taken a total sum of NC1,000 as bribe from an alleged smuggler, to “kill” a case in which an army officer, Lieutenant E. M. Takyi of the Field Engineers, Teshie, intercepted a Bedford truck loaded with 40 bales of “*Oboroniwawu*”, second hand clothings two years ago. (HiCE-PrReg048)

heavy backs (idiomatic formation) *sizeable lower backside*: The lady with *heavy backs* will also do well to avoid the mini (HiCE-SkiHob032)

20 of the 21 types attested for HiCE in the Arts & crafts section are new. However, they are all borrowings from Ewe, a Ghanaian language spoken in the eastern part of the country, stem from a single text and are all types of beads that the author describes, as in:

hotsui (noun) *type of bead*: Strings of *Hotsui* (cowries) are used on the ankles and arms. (HiCE-AcPopSci006)

Food and farming is a domain in which we expect a large number of borrowings from Ghanaian languages to describe typical food items endemic to the country. 17 new items were attested, all occur in HiCE, only one type is also found in ICE. Most of these come from only two texts, a recipe book and a page labelled “Have a Ghanaian dish” from a magazine. They have in common that usually the Ghanaian name is only used in the heading and then glossed in English as in the first example below:

oblayo (noun) *boiled crushed corn, porridge*: 1 cup leftover *oblayo* (boiled crushed corn). (HiCE-SkiHob009)

sokode (noun) *some type of food*:¹⁰ As one would who had met in the kitchen of an ultra modern hotel a fat woman famous for preparing *sokode* at the gutter's edge at Kokompe (HiCE-Story010)

apiti (noun) *meal of cooked plantains*: Same recipe as *apiti*, but Kakro [plantain cakes] is deep-fried in oil (HiCE-SkiHob012)

Section G: Government and the public domain includes ten new types, only one (advance guard, described above) of which is English. They come from a small range of sources. Examples include

dwanetoea (noun) *a customary law by which an offender may ask someone to see the offended with an apology on the offender's behalf*: Strangers are bewildered when, instead of a man approaching them directly, he asks someone else to do so on his behalf, it is the "*dwanetoea*" method of approach that has been employed. (HiCE-SkiHob001)

nhenkwaa (noun) *a civil servant in the Akan administrative system*: The *Nhenkwaa* had the right to go into any one's house and take the sheep, fowls and eggs or to take them off the street, no matter whose property they were. (HiCE-AcHum-012)

ayefare (noun) *a claim of damage made by the husband whose wife has committed adultery*: Let him be assured that I have not the slightest difficulty with understanding the meaning of things like '*Ayefare*'. (HiCE-LetEd055)

The section on architecture, buildings, houses & the home only contains four new Ghanaianisms from two texts. Three refer to the Ewe names of building materials as in:

heleti (noun) *spine of coconut palm branch*: HELETI – spine of coconut palm branch which is used for purlins and horizontal wall members (HiCE-AcTech001)

Two innovative Ghanaian types were documented in the money & commerce category, each occurring several times but in the same texts. The first item is a hybrid compound. They are:

abusa tenancy (noun) *a farm renting system*: *Abusa tenancy* is for the growing of tree crops such as cocoa, coffee, oil palm, coconut, citrus and so forth. (HiCE-SkiHob021)

10. No suitable translation or description was found.

tindana (noun) *a custodian who administered land*: The *Tindana's* office was not abolished; he was consulted on matters of importance and he played an important role in the installation of a paramount chief. (HiCE-PopSocSci012)

In addition, there were four items of English origin, the idiomatic formation *meet the pockets*, a new phrasal and prepositional verb each and a compound incorporating another Ghanaianism. A *lorry park* is a bus station, at which the traveller buys their ticket from the *bookman*, someone who makes travel arrangements. A *lorry park overseer* is synonymous of *bookman*:

meet the pockets (idiomatic formation) *to be financially affordable*: Depending on the availability of adequate financial assistance, the factory hopes to bottle its products in various sizes to *meet the pockets* of all classes of workers.

(HiCE-PopTech002)

pay in (phrasal verb) *pay*: A form of entertainment in which once you *pay in* at the gate you are sure of getting at least three things: a long, good night's laugh; loads of educative sayings and moralisations and the inevitable music session that follows, or often precedes the real concert show. (HiCE-PopHum018)

sell out (prepositional verb) *sell*: About 6,500 seeds of the plant *were sold out* or distributed from Aburi in 1899. (HiCE-AcTech010)

lorry park overseer (noun) *bookman*: THE Accra-Tema City Council is to re-engage all *lorry park overseers* (bookmen)

Entertainment, sports & games is the semantic domain in which by far the most new Ghanaianisms are documented, 63 items in total. 58 are of Ghanaian origin and denote cultural phenomena like dances or musical instruments. Of these, 38 types and 57 tokens stem from only two sources in HiCE, describing typical Ghanaian customs. It is therefore questionable how widespread knowledge and use of these items really is. Two examples are shown here:

siky (noun) *an Akan dance*: *Siyki*, another Akan dance is popular in the thick evergreen forest belt of Ashanti. (HiCE-PopHum007)

husago (noun) *a funeral dance*: But the thought of not seeing him physically, in facing life together, brings sorrow and the *Husago* returns; and the priests and priestess or heroes and heroines in this case, lead the procession to prepare for the burial. (HiCE-PopHum013)

One of the few English examples is an idiomatic formation found in two sports reports:

connect home (idiomatic formation) *score a goal*: They were rewarded in the 76th minute when Barnieh (senior) *connected home* Kwakye-Aliman combination. (HiCE-PrSpo005)

The category of life & living things has remarkably few types; particularly in ICE it is rather underrepresented. Given that local names for flora and fauna are among the most commonly borrowed we should expect to find many more. There are 13 newly attested types for plants and another 3 for animals, as in:

akakapempe (noun) *a tree*: A medicine man cuts a piece of the root of the tree called *akakapempe*, washes it, scrapes off the thin layers, puts it in a piece of linen cloth or towel, soaks it with water, and then he squeezes it to allow the juice to drop in his eyes. (HiCE-PopNatSci006)

konte (noun) *elephant*: The trophies taken from the animals are as follows:
Konte: lower jaw, tusks, tail and trunk (HiCE-AcHum009)

In addition, there is an English item which may be a loan translation from one of the Ghanaian languages, the context unfortunately does not provide any helpful details:

summary burial (idiomatic formation) ?: Ayele was given a *summary burial*. (HiCE-Story003)

There are only two neologisms in the domain Movement, location, travel & transport not documented before. Both come from the same text and describe different quarters of a town in Northern Ghana.

yagra-forng (noun) *the trader's section in the town of Kumbungu in Northern Ghana*

Nayilli-forng (noun) *the chief's section in the town of Kumbungu in Northern Ghana*

Both occur in one sentence: To the east is the *Yagra-forng* market section in which most of the traders live, and to the west is *Nayilli-forng*, or Chief's section where the Kumbun-Na resides. (HiCE-AcTech013)

Two types of Ewe measurements are newly attested for semantic domain N:

abo (noun) *the distance between the two middle fingers of a man standing with his arms outstretched*

afo (noun) *the length of a man's foot*: All measurements are based on the human body, the distance between the two middle fingers of a man standing with his arms outstretched (*abo*) is used for long measurements, while a man's foot (*afo*) or fingers are used for shorter distances. (HiCE-AcTech001)

In addition, there are two prepositional verbs of English origin:

speeden up (prepositional verb) *speed up*: It therefore calls on all African leaders to join arms against Racist South Africa to *speeden up* the process of destroying partheid [sic!] and all the enemies of mother Africa (HiCE-PressEd068)

cake-up (prepositional verb) *pile up*: Knowledgeable sources close to the Mill disclosed that the Mill would suspend the production of flour since its supply circulating in the country was far in excess of demand, and since some of the flour had *caked-up* because it had been purchased from the Mill about four months earlier. (HiCE-PrReg071)

There are five Ghanaian neologisms in domain O, all of which occur in only one text each. They include:

abaxe (noun) *a special mat for women of high social standing*: For such a person, a stool (for men) or a special mat (*abaxe*) (for women), is sanctified in his/her memory. (HiCE-AcSocSci-006)

ntsuma (noun) *a type of red clay*: Earlier in the afternoon of this Thursday, the womenfolk in town collect a special type of red clay (*ntsuma*), which they take to the gathering point (Mukpono); there they join the Soobii procession. (HiCE-PopHum004).

There are only very few types in the education domain and only one neologism:

akomyi (noun) *a graduation ceremony*: The days [sic!] activities divide themselves into five clear stages, and again, my field notes taken at this year's "akomyi", may serve as an illustration. (HiCE-AcHum015)

Category Q, denoting linguistic actions, states and processes, contains very few Ghanaianisms overall. Three new types were found, all are hapaxes:

din-wee (noun) *bare name*: The Ghanaian code of etiquette demands that you do not call your elders by their "bare" name (*din-wee*). (HiCE-SkiHob001)

discuss about (prepositional verb) *discuss*: Mansa came from the kitchen to see her two sisters holding their new cloths and *discussing about* the forthcoming festival. (HiCE-AcHum003)

nsuae (noun) *pledge to the village chief*: The heads of families make their pledges (*nsuae*) to village chiefs. (HiCE-AcHum012)

Social actions, states and processes are by far the most productive group. This is partly because in the USAS tagset this category combines many different aspects and partly because of its importance in everyday life. It is therefore not surprising that the number of newly attested neologisms is the among the highest. 42 new types were found, covering seven of the nine subdivisions. Examples are drawn from each of those.

A neologism found in the most general category is this one:

jungle (adjective) *uncivilised*: THE University of Science and Technology branch of the Progress Club, on Thursday called on the National Union of Ghana Students (NUGS) to concentrate on matters affecting students' welfare and not to drag the NUGS into "*jungle*" national politics. (HiCE-PrPol112)

S2 denotes people. There are several items in this group, such as the *ariser*, *boo boy* and *inmate* already discussed above. Commonly found in the older data is the compound *youngman*, written as one word. What first seemed like an error, is more likely a neologism just because of how frequently it occurs. There are 14 tokens in HiCE, two in ICE and about 30 in GloWbE. It often occurs in a negative context, so that it could be interpreted in the sense of rowdy, but then other examples suggest a more general preference for this word which is also found in one example of *youngwomen*:

youngman (noun) *man of young age (may carry negative connotations)*: As soon as one bends down to cut a piece of the meat the other *youngmen* rain blows and strokes upon him to deter him. (HiCE-PopHum-004)

youngwoman (noun) *woman of young age*: Again at Cape Coast, a few educated *youngwomen* – teachers and housewives – had in 1897 formed their own Young Ladies Christian Association. (HiCE-PopHum016)

An example from category S4 (kin) is the following:

ayete (noun) *arranged engagement*: Marriage is preceded either by courtship or parental agreement (*ayete*) (HiCE-AcSocSci007)

S5 describes groups and affiliations. There are four new types, two English and two Ghanaian. The adjective *non-stool* has been described above as an example from this category.

An example from category S7: Power relations has also already been mentioned, *stool leader*. There are five further neologisms.

Religion and the supernatural (S9) is the most productive subcategory. 21 neologisms were found. Most of these, however, were restricted to HiCE Ghana and occurred almost exclusively in only one text. An exception is the following:

ntoro (noun) *the father's soul/spirit which is passed down the generations* During my field-work, I asked some of my informants about the concept of *ntoro*, and I was told that the concept was foreign to the Akwapems, and that they did not believe that blood and spirit, which are the main constituents of the individual, could be designated in that way.

There are only four new items in the Time domain, all of which are hapaxes. An example can be provided by this new prepositional verb:

begin from ... to (prepositional verb) *last from ... to*: FIVE teams will compete in the four-day international polo contest which *begins in Accra from April 9 to April 12*. (HiCE-PrSpo022)

There is only one item in the weather domain, which may just be a fossilised spelling error, as nine of the 11 tokens stem from one text:

lightening (noun) *lightning*: However, he posited that not all rains produce lightening but some types of clouds facilitate the production of *lightening* when it rains and, as a consequence, it will be prudent on the part of people to put away cell phones whenever there is a downpour. (ICE-W2C008)

Psychological actions, states & processes is a large group in USAS but plays almost no role in the current data. The example *consider as* has already been mentioned above. There are two more types, which may rather be preferences than real neologisms. They are:

aim at + progressive (prepositional verb) *intend to*: In a Radio and Television broadcast, last Monday, Mr Ampaw said the campaign *aims at awakening* national consciousness to the need for good sanitation. (HiCE-PrPol015)

clear the air (idiomatic formation) *shed light, illustrate*: I COMMEND Mr Emmanuel K. Kwamong for *clearing the air* on the usefulness of honey to the body. (ICE-LetEd024)

The final category, Names and grammatical words, contains several examples of interjections of Ghanaian origin, but more interesting is the construction *a certain*, which may have its origin in Akan and Ewe. Both languages have an indefiniteness marker (*bi* in Akan, *ade* in Ewe) and there is a corresponding construction in English:

a certain (idiomatic formation) *indefiniteness marker*: In the old, old days, long before those who gave birth to our great grandfathers were born, there existed *a certain* village in the land of Asante. (HiCE-Novel010); In the course of her wandering she met and married *a certain* happy-go-lucky bachelor. (HiCE-AcHum013)

5.4 Entrenchment

If we want to find out how deeply a Ghanaianism is rooted in the variety we should look at how often it occurs across time, semantic domains and formality levels. I already briefly mentioned above that many more Ghanaianisms are attested in HiCE than in ICE and that this is partly due to the number of words of the two corpora. 573 types, accounting for 1,368 tokens were only attested in HiCE; 154 types (327 tokens) only occur in ICE. But we also find that Ghanaianisms may fall out of use because they refer to a concept that is no longer relevant. There are several such examples in HiCE, such as *Operation Feed Yourself* noted as obsolete by (Dako 2003: 159) and referring to a “campaign during the Acheampong regime [between 1972 and 1978] to become self-sufficient in food. Another is *letter writer*, ‘someone whose profession it is to write letters for other people’. This job has basically died out with the expansion of the education system and technological advances. On the other hand, there are items in ICE, which just cannot show up in HiCE, because they were founded much later, as the *Bureau of National Investigation (BNI)*, the internal intelligence agency in 1996 and the notion of HIPC – Heavily indebted poor countries – from the same year.

Overall, however, there is a remarkable stability in the usage of Ghanaianisms over time indicating that many of these mentioned by the various sources from the 1990s and 2000s were in fact well-established much earlier. In the following, I will only consider Ghanaianisms which were found at least five times and in both corpora. Applying this criterion, there are 259 types, which overall account for 5,892 tokens, i.e. 24% of the types make up about 75% of all tokens.

It is interesting to note the distribution across source language in this respect. More than 75% of these well-established Ghanaianisms are of English origin, about 19% are local items, 2.3% are hybrids; foreign words, place and proper names and items where no source language could be determined account for 3.1%. If we compare this to the overall distribution by source language, we find that interestingly it matches the general ICE patterns (76%, English, 19% Ghanaian, 2.6% hybrids, 2.8% other), but not those of HiCE, where about 60% of all types are of English origin, compared to 35% of local source and 2% and 3% respectively for hybrids and other.¹¹ This suggests that unlike what was hypothesised before, the absence of an STL strand does not lead to a stronger adoption of local words or constructions in writing. It is much more so that the larger number of local items in the early data

11. There is also a big discrepancy between the values from the corpora and those suggested by Dako (2001: 27) cited above. One explanation lies in the methodology. The corpora are very limited in size and structure and contain only printed data. Dako, on the other hand collected data from a much larger range of sources (including student writing) and over a longer period of time.

is caused by the choice of topics covered. As was mentioned above, most of the texts from the humanities and social sciences sections of HiCE describe in great detail traditional customs and social practices from the different tribes and regions of the country, therefore referring much frequently to words of local origin.

Table 8. Semantic classification according to USAS tagset of types attested five or times

	Types (pttw)	Tokens (pttw)	
		HiCE	ICE
S: Social actions, states & processes	54 (0.6)	1,038 (17.3)	415 (13.4)
Z: Names & grammatical words	29 (0.3)	589 (9.8)	299 (9.6)
M: Movement, location, travel & transport	23 (0.3)	281 (4.7)	88 (2.8)
A: General & Abstract terms	22 (0.2)	221 (3.7)	110 (3.5)
F: Food & farming	21 (0.2)	322 (5.4)	82 (2.6)
H: Architecture, buildings, houses & the home	16 (0.2)	200 (3.3)	67 (2.2)
B: The body and the individual	14 (0.2)	111 (1.9)	151 (4.9)
I: Money & commerce	13 (0.1)	769 (12.8)	99 (3.2)
K: Entertainment, sports & games	12 (0.1)	84 (1.4)	39 (1.3)
G: Government & the public domain	10 (0.1)	51 (0.9)	52 (1.7)
P: Education	9 (0.1)	85 (1.4)	55 (1.8)
L: Life & Living things	7 (0.1)	111 (1.9)	24 (0.8)
O: Substances, materials, objects & equipment	6 (0.1)	75 (1.3)	17 (0.5)
X: Psychological actions, states & processes	5 (0.1)	80 (1.3)	37 (1.2)
T: Time	4 (0)	67 (1.1)	17 (0.5)
W: The world & our environment	4 (0)	78 (1.3)	28 (0.9)
Q: Linguistic actions, states & processes	4 (0)	42 (0.7)	20 (0.6)
N: Numbers & measurements	4 (0)	38 (0.6)	22 (0.7)
E: Emotional actions, states & processes	2 (0)	22 (0.4)	6 (0.2)
Total	259 (2.8)	4,264 (71.1)	1,628 (52.5)

The range of semantic domains of these common items is shown in Table 8, sorted by type frequency. Compared to the newly-attested items, we find that the Entertainment, sports and games category, which was the most productive, is much less important overall. Similarly, life and living things and arts & crafts, which does not even show up in the second list, are much less important. This once more supports the claim that the earlier data was strongly characterised by borrowings typical referring to flora and fauna as well as cultural aspects. These were described in detail by scholars, but only very few made into the common vocabulary. This is different for the social domain, which not only showed the highest productive regarding new items, but also is the most important overall. 54 types account for over 1,000 tokens in HiCE and 415 in ICE. The term *Asante/Ashanti*, denoting both the most populous ethnic tribe and a region, is by far the most frequent,

found 281 times in HiCE and 89 times in ICE. *Stool* 'seat & authority of the chief' comes in second (62/21 tokens) and *gods* is the third most frequently found item (55/21 tokens). The second-most frequent category is Names & grammatical words. Here, the name of a well-known Ghanaian football club – *Kotoko* – which literally translates as porcupine is found most frequently, particularly in the HiCE data (118/13 tokens).¹² *Nana*, the appellation for an Akan chief is the second-most frequent (74/42 tokens) and the suffix *-hene* 'chief' in the Akan system (excluding *omanhene* 'paramount chief' and *asantehene* 'King of the Ashanti') comes in third (70/14 tokens). The three most frequent items in the Movement, location, travel & transport domain are *reach* 'arrive (at)' (53/14 tokens), *tro-tro* 'converted lorry/minibus, used for passenger transport' (58/2¹³ tokens) and *overseas* 'abroad' (26/3 tokens). A salient feature of Ghanaian English, the usage of *would* for *will* comes in first in the general & abstract domain. At present, 30 tokens are documented in each corpus.¹⁴ The prefix *ill-*, described above was found 26 times in HiCE and five times in ICE. *Going on*, in the sense of *taking place* was the third most frequent type (17/9 tokens). Food and farming items, unsurprisingly, mainly contain common groceries and Ghanaian dishes. The three most frequent are *yam*, a starchy tuber common across West Africa (52/21 tokens), *groundnut* (49/6 tokens), which is preferred over *peanut* and *cassava* 'manioc' (37/4 tokens).

There is little noteworthy in the other categories with two exceptions. In the Body and the individual domain, there are more tokens in ICE than HiCE. This is due to the terms *kente* and *kente cloth*, which show up 61 times between two texts in ICE. The other is the high token frequency in the Money & Commerce domain. This is due to two items: the *workers* mentioned above and the word *cedi*, denoting the local currency (158/26 tokens).

A final measurement of the entrenchment of Ghanaianisms briefly to be discussed here is the occurrence across different levels of formality, summarised in Table 9. The least formal level covers creative writing. As mentioned above, this includes the representation of speech and is usually deeply rooted in a society. Therefore, it is not surprising that in both corpora the type frequency is among the highest. The intermediate category covers the skills & hobbies section and the

12. This is likely so because of the division of the news category into four sub-sections in HiCE: Political, Cultural, Regional, Sports (also see Brato this volume)

13. This is clearly due to corpus size. *Tro-tro* is a very common concept as it forms the backbone of public transport in the country.

14. The analysis of modal verbs is currently still ongoing, so this figure is subject to change. After having coded about a third of the approximately 13,000 modals in the corpora, the picture that emerges is that using the past tense forms in contexts in which the present tense would be expected (or the other way around) is a highly salient (Huber & Dako 2004), but rather infrequent feature.

letters to the editor. In HiCE, the normalised frequencies are only marginally different to the former. Breaking this down further, we find that this is mainly due to the skills & hobbies section and herein the dominance of items from the food domain. The differences between the corpora regarding the rather formal style (covering popular writing, news, and editorials are only small. Rather more interesting is what we find in the formal domain, containing the academic and administrative writing sections. The type frequency is about 50% larger in HiCE, but more striking is the massive difference in the tokens. The reason for this was already alluded to above. Many of the early texts from the humanities and social sciences describe various cultural practices and therefore individual types, e.g. denoting dances, types of drums or kinship systems are extraordinarily frequent.

Table 9. Distribution across formality levels

Formality level	Corpus	Types	Tokens
Rather informal	HiCE	323 (40.4)	1,088 (136)
	ICE	168 (42.0)	321 (80.3)
Intermediate	HiCE	282 (40.3)	868 (124)
	ICE	106 (35.3)	181 (60.3)
Rather formal	HiCE	527 (19.5)	2,384 (88.3)
	ICE	315 (22.5)	1,039 (74.2)
Formal	HiCE	349 (24.9)	1,437 (102.6)
	ICE	161 (16.1)	534 (53.4)

6. Conclusion

This paper set out to provide a real-time structural and semantic analysis of lexical expansion in the Nativization phase of Ghanaian English. Taking a comprehensive list of previously attested Ghanaianisms as a starting point it showed that only a very small number of word-formation processes are employed in vocabulary growth in Ghanaian English and that these are relatively equally distributed across time.

What was clearly shown here is that it is important to not only consider typical word-formation processes like derivation or compounding if one wants to fully capture how the vocabulary is expanded in New Englishes. It is vital that we not only take a quantitative-structural approach but that we rely on qualitative assessment of the data to discover semantic shifts. Therefore, it may even be more helpful to work with smaller corpora (or excerpts from larger ones) so that the researcher has a chance to actually read through the data and get a better picture.

It has also been shown that although Ghanaianisms come from a wide range of local and foreign languages, it is the English items which make up the core, both

regarding overall frequency but also for items more deeply entrenched. Heavy lexical borrowing was clearly a feature of the early phase. However, as was shown, the dominance of Ghanaian items in the early corpus is very much due to the different contents addressed in some of the categories. These are much broader in the ICE data and do not so much address Ghanaian customs, a major source for borrowings in the older data. At the same time, the comparatively small number of neologisms of Ghanaian origin was not expected. It was hypothesized that the absence of an STL strand would rather favour the use of Ghanaian words and phrases, particularly of Akan origin as this group is the most dominant, in terms of population, exposure to English and power.

Regarding the English items it is noteworthy that so many Ghanaianisms attested by scholars over the last twenty years were already present in the early data and furthermore that they were found across all semantic domains. This early entrenchment can be seen as an indicator that a core stock of Ghanaianisms was already well-established in the early stages of Nativization. Whether this means that political independence was the trigger to short-term and large-scale lexical expansion or that many of these items have been around even longer, cannot be addressed here in the absence of any empirical data from the colonial period. Hopefully, colleagues working on historical and diachronic corpora going back to the 1930s (Hoffmann, Sand, and Tan 2012; Biewer et al. 2014) can shed more light on this – at least on a qualitative level.

Of course, corpora like HiCE Ghana are comparatively small for lexical research, but the results still provide a useful starting point for our understanding of how Ghanaian English has evolved over 40 years. The overall frequency has increased slightly, yet there is a remarkable homogeneity regarding the processes employed. English-based forms are clearly dominant; affixation only plays a subordinate role in the creation of new lexemes and while Ghanaianisms are found across the range of semantic domains, there are some which clearly dominate.

References

- Ahulu, Samuel. 1994. How Ghanaian Is Ghanaian English? *English Today* 10(2): 25–29. <https://doi.org/10.1017/S0266078400007471>
- Alabi, Victoria A. 2000. Semantics of occupational lexis in Nigerian English. *World Englishes* 19(1): 107–12. <https://doi.org/10.1111/1467-971X.00158>
- Archer, Dawn, Andrew Wilson, & Rayson, Paul. 2002. *Introduction to the USAS category system*. Ms, last modified June 12, 2016. <<http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf>>
- Baayen, Rolf Harald. 2001. Word frequency distributions. *Text, Speech and Language Technology* 18. Dordrecht: Kluwer.
- Baayen, Rolf Harald & Lieber, Rochelle. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29(5): 801–44. <https://doi.org/10.1515/ling.1991.29.5.801>

- Bamiro, Edmund O. 1997. Lexical innovation in Ghanaian English: Some examples from recent fiction. *American Speech* 72(1): 105–112. <https://doi.org/10.2307/455612>
- Baumgardner, Robert J. 1998. Word-formation In Pakistani English. *English World-Wide* 19(2): 205–46. <https://doi.org/10.1075/eww.19.2.04bau>
- Biermeier, Thomas. 2008. *Word-formation in new Englishes: A corpus-based analysis* [Anglistik, Amerikanistik 29]. Münster: LIT.
- Biewer, Carolin, Bernaisch, Tobias, Berger, Mike & Heller, Benedikt. 2014. Compiling The Diachronic Corpus of Hong Kong English (DC-HKE): Motivation, progress and challenges. *ICAME* 35. Nottingham, April 30.
- Blench, Roger. 2006. *A Dictionary of Ghanaian English*. <<http://www.rogerblench.info/Language/English/Ghana%20English%20dictionary.pdf>> (13 July 2016).
- Brown, Philip Penton & Scragg, John. 1950. *Common Errors in Gold Coast English: Their Cause and Correction*. London: Macmillan.
- Dako, Kari. 2001. Ghanaianisms: Towards a semantic and formal classification. *English World-Wide* 22(1): 23–53. <https://doi.org/10.1075/eww.22.1.03dak>
- Dako, Kari. 2003. *Ghanaianisms: A Glossary*. Accra: Ghana University Press.
- Davies, Mark. 2016. NOW Corpus (News On the Web). <<http://corpus.byu.edu/now/>> (22 June 2016).
- Denteh, Andrew C. 1967. *A Visitor's Guide to Ghanaian Customs and Other Matters*. Accra: Self-published.
- Evans, Stephen. 2015. Word-formation in Hong Kong English: Diachronic and synchronic perspectives. *Asian Englishes* 17(2): 116–31. <https://doi.org/10.1080/13488678.2015.1036510>
- Evert, Stefan & Baroni, Marco. 2007. zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29–32. Prague.
- Fuchs, Robert & Gut, Ulrike. 2015. An apparent time study of the progressive in Nigerian English. In *Grammatical Change in English World-wide* [Studies in Corpus Linguistics 67] Peter C. Collins (ed.), 373–87. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.67.15fuc>
- Gold Coast Colony. 1930. Report on the Education Department for the Year 1929–1930.
- Gyasi, Ibrahim Kwaku. 1991. Aspects of English in Ghana. *English Today* 26: 26–31. <https://doi.org/10.1017/S0266078400005502>
- Hoffmann, Sebastian, Sand, Andrea & Tan, Peter K. W. 2012. *The Corpus of Historical Singapore English: A first pilot study on data from the 1950s and 1960s*. Leuven, 30 May.
- Huber, Magnus. 2004. Ghanaian English: Phonology. In *A Handbook of Varieties of English: A Multimedia Reference Tool*, 3 Vols, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 842–865. Berlin: De Gruyter.
- Huber, Magnus. 2014. Stylistic and sociolinguistic variation in Schneider's nativization phase: The case of Ghanaian English. In *The Evolution of Englishes: The Dynamic Model and Beyond* [Varieties of English around the World G49], Sarah Buschfeld, Magnus Huber, Thomas Hoffmann & Alexander Kautzsch (eds), 86–106. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g49.06hub>
- Huber, Magnus & Dako, Kari. 2004. Ghanaian English – Morphology and syntax. In *A Handbook of Varieties of English: A Multimedia Reference Tool*, 3 Vols, Edgar Schneider, Kate Burridge, Bernd Kortmann, Rajend Mesthrie & Clive Upton (eds), 854–865. Berlin: De Gruyter.
- Huber, Magnus & Dako, Kari. 2013. *International Corpus of English, Ghana Component*. Gießen: University of Giessen.

- Kirby, Jon P. 1998. *A North American's Guide to Ghanaian English*. Tamale: Tamale Institute of Cross-Cultural Studies.
- Lambert, James. 2014. Diachronic stability in Indian English lexis. *World Englishes* 33(1): 112–27. <https://doi.org/10.1111/weng.12072>
- Mensah, A. N. 1971. The language of Ghanaian newspapers. *Legon Observer*, 22 October.
- Nimako, D. Annor. 2004. *Mind Your Language: Educated Ghanaian English*. Tema: Ronna Publishers.
- Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511618901>
- Schneider, Edgar W., Burridge, Kate, Kortmann, Bernd, Mesthrie, Rajend & Upton, Clive (eds). 2004. *A Handbook of Varieties of English: A Multimedia Reference Tool*, 3 Vols. Berlin: Mouton de Gruyter.
- Sey, Kofi Abakah. 1973. *Ghanaian English: An exploratory survey*. London: Macmillan.
- Tent, Jan. 2001. A profile of the Fiji English lexis. *English World-Wide* 22(2): 209–45. <https://doi.org/10.1075/eww.22.2.03ten>
- Tingley, Christopher. 1981. Deviance in the English of Ghanaian newspapers. *English World-Wide* 2(1): 39–62. <https://doi.org/10.1075/eww.2.1.04tin>
- van Rooy, Bertus & Terblanche, Lize. 2010. Complexity in word-formation processes in New Varieties of South African English. *Southern African Linguistics and Applied Language Studies* 28(4): 407–24. <https://doi.org/10.2989/16073614.2010.548022>

Appendix

Examples of regular expressions used

Regular expressions allow for relatively complex search queries in AntConc and other software. They were used in this study in several ways to find both previously attested Ghanaianisms and potential new items. Some examples are outlined below. For items of Ghanaian origin relatively broad regular expressions were used, i.e. unlike for most English items I included a placeholder for the POS tag, which was not always meaningful. (4) is an example of a Ghanaian item with multiple possible spellings for the item *abusuapanyin/Abusua Panin*, the head of the matrilineage.

- (4) (i)\babusua(_\w+)?(-|\s)?pany?in_\w+\b
 (i) – this means case-insensitive
 \b – word boundary
 abusua(_\w+)? – finds the sequence of letters *abusua*, followed 0 or 1 times (indicated by the final question mark) by the expression in the brackets. This expression is a placeholder for all sorts of POS tags created by Treetagger: Underscore followed by one or more (the plus sign) letters (\w shorthand for [a-zA-Z])
 (-|\s)? – an optional hyphen or space (\s)
 pany?in_\w+\b – the first element finds both the forms *panyin* and *panin* followed by an underscore, a POS tag and a word boundary

Example (5) is the regex for the compound *lorry park*, a (bus station).

- (5) (i)\blorry(_N[NP])?(-|s)?parks?_N[NP]S?\b
- As above, (i) marks case-insensitiveness; \b indicates word boundaries, (-|s)? an optional hyphen or space
- lorry(_N[NP])? – This part finds the sequence *lorry*, optionally followed by a POS tag for a noun (NN) or proper noun (NP). It was found necessary to search for both the NN and NP tag, as Treetagger used tagged ‘normal’ nouns as proper nouns when the spelling was in capitals (as in headings).
- parks?_N[NP]S? – finds the sequence *park* or *parks*, followed by a noun tag in the singular or plural (indicated by the final S?)

Example (6) shows the regex for the prepositional verb *connect home*, meaning to score a goal

- (6) (i)\bconnect\w*_V\w+\shome_NN\b
- As above, (i) marks case-insensitiveness; \b indicates word boundaries
- connect\w*_V\w+ – This part finds the sequence *connect*, optionally followed by zero or letters (\w*), an underscore and any verb tag (V\w+)
- \s – a compulsory space
- home_NN – the noun *home*

Several very basic regular expressions, e.g. as shown in (7) for noun-to-noun suffixation, were used to discover potential new words using suffixations. The results, of course, largely yielded non-innovative words, which were discarded.

- (7) (i)\bw+(ese|i?an|is[mt]|ie|y|ship)_N\w+\b
- finds any words ending in *ese*, *ian*, *an*, *ism*, *ist*, *ie*, *y* or *ship*, followed by POS tag for nouns.

Additional table

Table 10. Semantic classification based on the USAS tagset

USAS	Corpus	All items		Not previously attested	
		Types	Tokens	Types	Tokens
A: General & Abstract terms	HiCE	51 (0.9)	266 (4.4)	3 (0.1)	29 (0.5)
	ICE	36 (1.2)	126 (4.1)	1 (0)	5 (0.2)
B: The body and the individual	HiCE	65 (1.1)	205 (3.4)	8 (0.1)	10 (0.2)
	ICE	39 (1.3)	193 (6.2)		
C: Arts & crafts	HiCE	21 (0.4)	35 (0.6)	20 (0.3)	34 (0.6)
	ICE	2 (0.1)	2 (0.1)		
E: Emotional actions, states & processes	HiCE	7 (0.1)	38 (0.6)		
	ICE	7 (0.2)	15 (0.5)		
F: Food & farming	HiCE	83 (1.4)	457 (7.6)	17 (0.3)	24 (0.4)
	ICE	46 (1.5)	120 (3.9)	1 (0)	1 (0)

Table 10. (*continued*)

USAS	Corpus	All items		Not previously attested	
		Types	Tokens	Types	Tokens
G: Government & the public domain	HiCE	40 (0.7)	127 (2.1)	9 (0.2)	16 (0.3)
	ICE	26 (0.8)	113 (3.6)	3 (0.1)	4 (0.1)
H: Architecture, buildings, houses & the home	HiCE	32 (0.5)	225 (3.8)	4 (0.1)	4 (0.1)
	ICE	24 (0.8)	78 (2.5)		
I: Money & commerce	HiCE	52 (0.9)	889 (14.8)	7 (0.1)	15 (0.3)
	ICE	27 (0.9)	131 (4.2)		
K: Entertainment, sports & games	HiCE	99 (1.7)	260 (4.3)	62 (1.0)	99 (1.7)
	ICE	19 (0.6)	54 (1.7)	1 (0)	4 (0.1)
L: Life & Living things	HiCE	41 (0.7)	188 (3.1)	21 (0.4)	33 (0.6)
	ICE	12 (0.4)	30 (1.0)		
M: Movement, location, travel & transport	HiCE	49 (0.8)	355 (5.9)	4 (0.1)	4 (0.1)
	ICE	27 (0.9)	94 (3.0)		
N: Numbers & measurements	HiCE	20 (0.3)	61 (1.0)	5 (0.1)	10 (0.2)
	ICE	11 (0.4)	36 (1.2)	1 (0)	1 (0)
O: Substances, materials, objects & equipment	HiCE	31 (0.5)	121 (2.0)	5 (0.1)	9 (0.2)
	ICE	14 (0.5)	29 (0.9)		
P: Education	HiCE	20 (0.3)	114 (1.9)	1 (0)	3 (0.1)
	ICE	18 (0.6)	70 (2.3)		
Q: Linguistic actions, states & processes	HiCE	22 (0.4)	70 (1.2)	3 (0.1)	3 (0.1)
	ICE	13 (0.4)	30 (1)		
S: Social actions, states & processes	HiCE	183 (3.1)	1328 (22.1)	41 (0.7)	88 (1.5)
	ICE	112 (1.9)	523 (8.7)	7 (0.2)	10 (0.3)
T: Time	HiCE	17 (0.3)	134 (2.2)	4 (0.1)	4 (0.1)
	ICE	13 (0.2)	29 (0.5)		
W: The world & our environment	HiCE	7 (0.1)	91 (1.5)	1 (0)	1 (0)
	ICE	4 (0.1)	28 (0.5)	1 (0)	10 (0.3)
X: Psychological actions, states & processes	HiCE	12 (0.2)	97 (1.6)	2 (0)	60 (1.0)
	ICE	12 (0.2)	48 (0.8)	3 (0.1)	32 (1.0)
Z: Names & grammatical words	HiCE	71 (1.2)	716 (11.9)	17 (0.3)	53 (0.9)
	ICE	42 (0.7)	326 (5.4)	3 (0.1)	6 (0.2)

Capturing the lexicon of Ugandan English

ICE-Uganda, its limitations, and effective complements

Bebwa Isingoma and Christiane Meierkord

Gulu University / Ruhr-University of Bochum

This paper discusses corpus linguistics as one method to investigate the lexicon of Ugandan English, which is characterised by borrowing, calquing, semantic extension, narrowing, and shift. It documents how analysing a well-balanced corpus, such as the Uganda component of the *International Corpus of English*, allows for a contextualisation of observations made from current uses of English and for an assessment of the textual genres in which such innovations occur. At the same time, it critically discusses the limitations and biases associated with a comparatively small corpus and argues for a multi-method approach that involves using larger, though less well controlled, corpora as well as supplementing corpus analyses with experimental data, which tap into the spread of lexical innovations.

1. Introduction

The level of the lexicon has often been said to be one of the most obvious to display particularities of World Englishes, and Ugandan English (UgE) would seem to be no exception. However, in individual textbooks, the lexicon is simply described through a brief list of borrowings from the indigenous language, for example in Trudgill and Hannah (2017).

For some varieties, the lexicon has been captured in dictionaries (e.g. South African English in Branford & Branford 1991 or Silva et al. 1996; or Nigerian English in Igboanusi 2010). Mostly, however, there exist individual glossaries or research papers such as Dako (2003) for Ghanaian English, Kouega (2006, 2007) for Cameroonian English and Skandera (2003) for Kenyan English. As Wolf and Polzenhagen (2009: 10) point out, typically, “[i]tems have been identified that are specific or typical of a particular country, yet it is largely unexplored whether these features are also exclusive, i.e., don’t occur elsewhere in the region, or even beyond.” Similarly, Melchers and Shaw (2013: 141) caution that “[s]ometimes specific items thought to be typical of a particular variety are actually more widespread” and sometimes are “just a natural simplification” (in the case of *fill a form* instead of

fill out/ in a form). Today, the comparisons that are necessary to address these issues are easily possible through a comparison of corpora, particularly through the *Corpus of Global Web-based English* (GloWbE; see Section 4.2 for details). However, as Wolf and Polzenhagen (2009) argue, such a mere descriptive approach might not always possess the full explanatory power necessary to interpret individual phenomena.

Therefore, this paper starts from a traditional description of the characteristics of UgE's vocabulary which is based on evidence collected from newspapers, books and other writings authored by Ugandans. It then moves on to examine a controlled corpus of UgE, the Uganda component of the *International Corpus of English* (ICE-UG), which has recently been completed at the Ruhr-University of Bochum, Germany. Against the background of the findings of these analyses, the paper then argues that a multi-method approach is necessary to fully account for the realities of a variety's lexicon. On the one hand, evidence from a small, carefully designed corpus (such as the individual components of the ICE) is a good starting point, but studies of the lexicon often require a considerably larger data set. One method of collecting such large-scale corpora, using web-crawlers, is presented and discussed as regards its potential but also its limitations in comparison to the more carefully controlled corpora such as ICE. Furthermore, we argue that, as many particularities are more characteristic of spoken language use not reflected in written data, data collected through acceptability tests and production tasks provide a rich complement to written corpus data.

In fact, the results of the latter indicate that there are crucial discrepancies between what can be observed through long-term participant observation, the results yielded by corpus analyses, and tests that probe into Ugandans' familiarity with individual vocabulary items.

1.1 English in Uganda

The spread of English beyond the British Isles started in the early modern period (1500–1800) with the arrival and subsequent settlement of British colonists in Virginia in the current United States of America (Algeo 2010: 140). The British thrust into the world continued and brought them to what is now called Uganda, with the first Englishman, Sir Henry Morton Stanley, arriving in the area in 1874 (Meierkord 2016a). He was followed shortly by Anglican and Catholic missionaries respectively. The establishment of the Anglican and Catholic missions in the area evidently brought in people who spoke non-indigenous languages. Missionaries would soon start schools. The number of these people increased when the British government, having declared a protectorate over Buganda in 1894 and later over other areas, dispatched administrative personnel as well as teachers (cf. Meierkord

2016a). As is evident, the British personnel could not take charge of the entire protectorate; hence, there was a need to train local personnel for that purpose. Schools were thus established in different parts of the protectorate mainly by missionaries (Sekamwa 1997). This meant that two categories of speakers of English emerged: on the one hand, first language (L1) speakers comprising British personnel as well as a small number of settlers (Meierkord 2016a), on the other hand, second language (L2) speakers comprising an elite from the indigenous peoples. In both cases, the number of speakers of English was considerably small. The British population was of very limited size, as Uganda was not projected as a settlement colony. Ugandans mainly acquired English at school, but the education system of the time emphasised local languages (cf. Scanlon 1964), while English was only employed in the upper primary section and the secondary segment of education; moreover this level was primarily meant for the “sons of parents who could afford to pay school fees” (Byabazaire 1979: 64). Despite the increased number and spread of schools all over the country in the years to follow, with the resultant increase in schooling as well as the ubiquitous use of English in official settings and other domains today, English remains the language of a minority (cf. Nakayiza 2016).

As is the case with other British colonies and protectorates, English was brought to Uganda, where indigenous languages had already established themselves, a situation that has far-reaching effects on the variety of English spoken in the country. Since a precise assessment of the number of Ugandan languages is difficult, due to the problematic nature of the language versus dialect dichotomy, estimates range from 35 languages (Babito 2005) to 65 (Kabanankye & Kwagala 2007). Namyalo, Isingoma and Meierkord (2016) follow Lewis, Simons and Fennig (2015), who identify a total of 39 indigenous languages. What is clear, however, is the fact that there are four language groups in Uganda, namely Bantu (66.4%), Nilotic (27.2%), Central Sudanic (6.3%) and Kuliak (less than 1%).¹ The numeric supremacy of Bantu languages is said to have a bearing on how English is spoken in Uganda (Isingoma 2016: 215), although it is safe to state that other languages in Uganda also play a role in their own right (cf. Meierkord 2016b; Ssempuuma, Isingoma & Meierkord 2016). Statistics on the number of L2 speakers of English in Uganda are scanty as they do not provide a clear picture of the current number of speakers. Crystal (2003: 65) estimates that there are 2,500,000 speakers, while Namyalo, Isingoma and Meierkord (2016: 23) conjecture that the number could have quadrupled due to increased schooling at secondary level (cf. Uganda Bureau of Statistics 2010), as English is acquired mainly at school. Based on the assumption that attainment of secondary

1. The percentages were computed based on data provided by Simons and Fennig (2017), following the 2014 census data (but see Namyalo, Isingoma & Meierkord 2016: 21 and references therein for a caution regarding the use of census data).

level education translates into an intermediate level of proficiency in English, it might be right to maintain that the number of L2 speakers has indeed quadrupled, since current reports indicate that up to 15.4% of the population has attained the ordinary level of education (Uganda Bureau of Statistics 2016: 26). Furthermore, in urban areas, the completion of the primary cycle is sufficient to make one proficient in English, as English is used as the medium of instruction right from nursery schools (cf. Ssentanda 2016; Namyalo, Isingoma & Meierkord 2016). L1 speakers of English, on the other hand, are few and the 2014 census reports a total of 3,235 people from the UK, the USA, Canada and Australia, whom we assume are all L1 speakers of English. Despite the fact that the majority of Ugandans do not speak English, as it is associated with schooling, the language enjoys the highest status as an official language and as a medium of instruction, and it also has the highest socio-economic prestige associated with it (cf. Nakayiza 2016).

2. Ugandan English and its lexicon

2.1 Ugandan English

UgE is doubtless a variety of English spoken mainly in Uganda. It manifests itself at all linguistic levels, namely phonological, morphological, syntactic, lexico-semantic, and pragmatic (cf. Fisher 2000). Substrate influence from indigenous languages plays an important role in the realisations of UgE forms, but also levelling and acculturation have a role to play. UgE shares some of its properties with its East African neighbours: Kenyan and Tanzanian English, partly since some of its borrowings are from Kiswahili, the main substrate language that influences Kenyan and Tanzanian English. In addition, as an L2 variety, UgE naturally exhibits similar properties as other L2 varieties, both within East Africa and beyond, due to factors such as levelling, semantic change, archaisms, etc., which are typical of L2 Englishes (cf. Mesthrie & Bhatt 2008; Schneider 2007).

2.2 The lexicon of Ugandan English

2.2.1 Introduction

As is the case with other varieties of English, the lexicon of UgE is characteristically typified by borrowing and calquing, semantic change, morphological and phraseological innovations as well as archaisms. Whilst the above categories are attested in other L2 varieties, archaisms involving the same lexical items (e.g. *ever* 'always', *thrice* 'three times', *run mad* 'go mad') are quintessentially ubiquitous in L2 varieties (cf. Mesthrie & Bhatt 2008: 115). For instance, GloWbE shows that *run mad* 'go

mad' is present in many L2 varieties, for example, Kenyan, Tanzanian, Ghanaian, Nigerian English, etc., including L1 Englishes (presumably in non-standard varieties).² In the following sections, we present different categories of the UgE lexicon, with examples drawn from local national newspapers (*The Monitor*, *The New Vision* *The Observer* and *Red Pepper*) and other writings authored by Ugandans.³

2.2.2 Borrowing

Lexical borrowing is one of the most conspicuous features of L2 Englishes. Schneider's (2007: 35ff.) five-phase model of the development of Englishes shows that borrowing manifests itself in the very first phase, the foundation phase, where toponyms are borrowed, and borrowing continues exponentially in the subsequent phases.⁴ Instantiations of lexical borrowings in UgE are given in Tables 1 and 2, some of which are taken from Isingoma (2016) and Nassenstein (2016).

Table 1. Examples of nominal borrowings in Ugandan English

Word	Origin	Meaning
balokole (sing. muloke)	Ugandan Bantu	Pentecostalists
bayaye (sing. muyaye)	Ugandan Bantu	idler, vagabond, con
bodaboda	Ugandan Bantu (originally borrowed and adapted from English)	motorcycle/bicycle taxi/taxi man
boo	Ugandan Nilotic	a type of greens
bwana	Kiswahili	Mister
chapati	Hindi	a flat pancake-like bread
entandikwa	Ugandan Bantu	microfinance scheme
kabalagala	Ugandan Bantu	pancake
kadongo kamu	Ugandan Bantu	a type of Ugandan music
kisanja hakuna mchezo	Ugandan Bantu + Kiswahili	term of tenure with no jokes or laxity
kyeyo	Ugandan Bantu	menial jobs done by Ugandans or Africans in developed countries/ any extra work done to earn more money

(continued)

2. GloWbE: Corpus of Global Web-Based English

3. Due to space, not every instantiation of the lexical items will have an example.

4. The other phases are: exonormative stabilisation, nativisation, endonormative stabilisation and differentiation (Schneider 2007: 33–35).

Table 1. (continued)

Word	Origin	Meaning
mailo	Ugandan Bantu (originally borrowed and adapted from English)	land tenure system
malakwang	Ugandan Nilotic	a type of greens
matooke	Ugandan Bantu	mashed cooked bananas/bananas from which the meal is made
muzigo	Ugandan Bantu	rental house
posho	Ugandan Bantu (originally borrowed and adapted from English)	cornmeal
simsim	Arabic via Kiswahili	sesame
shamba boy	Kiswahili + English	field hand/ farmhand
sukuma wiki	Kiswahili	collard, kale
yaka	Ugandan Bantu	a prepaid electricity consumption billing system where a consumer buys units and loads them onto their meter

Examples of lexical borrowings extracted from Ugandan newspapers are given in (1).

- (1) a. [...] and particularly interrogate the prosperity theology that now underpins the new version of the *Balokole* movement. (*New Vision* 13/03/17)
- b. The doctor revealed that travelling on a *bodaboda* was dangerous for her life. (*Monitor* 11/07/17)
- c. What makes *boo* than just leaves is the fact that it weaves a thread of variety not only through its aroma, but [...] (*Monitor* 24/06/12)
- d. Today, the *chapati* franchise in Kampala and beyond is owned by Basoga youth. (*Monitor* 09/05/13)
- e. Today, *kabalagala* has evolved to not only have tastier ingredients, but also to attract more affluent people to be served in uptown eateries. (*Monitor* 22/07/12)
- f. It is a blend of lemons and passion fruit with steamed *matooke* and beef [...]. (*Observer* 21/07/17)
- g. He argues that *kyeyo* is beneficial for target workers. (*Monitor* 06/11/15)
- h. *Malakwang* is widely eaten and cooked in restaurants and eating joints that serve northern cuisine [...]. (*Monitor* 19/08/12)
- i. From a bungalow, Harriet is Alex's wife, living in a *muzigo* in Kireka. (*Monitor* 08/09/14)
- j. Some communities are to produce housemaids and *shamba boys* for other communities. (*New Vision* 10/03/17)

Despite the prevalence of Ugandan Bantu languages as sources of nominal borrowings in UgE, other indigenous languages, as well as endogenous languages, have also contributed to the UgE lexicon. For example, *malakwang* ‘type of greens’ (1h) and *boo* ‘type of greens’ (1c) are from Ugandan Nilotic languages, but their use is generally restricted to areas where Nilotic languages are spoken, that is to Northern Uganda, or restaurants associated with what is locally known as the “northern cuisine”. *Shamba* in the compound *shamba boy* ‘field hand/farmhand’ (1j) is of endogenous origin, as it comes from Kiswahili, the East African lingua franca. This explains the occurrence of this expression in both Kenyan and Tanzanian English (cf. GloWbE), where Kiswahili plays a major substrate role. The originally Hindi word *chapati* reflects the impact made by the Asian population in the past.

Not surprisingly, most of these borrowings are in the semantic fields of food, as Schmied (2012: 245) points out due to the fact that “in many African countries there is a marked contrast between European and African food (and eating habits) because Europeans in East Africa have tended not to adopt African food”. However, borrowing goes beyond nouns and includes adverbs, discourse markers and morphemes, as shown in Table 2.

Table 2. Examples of borrowings in other grammatical categories

Expression	Category	Origin	Meaning
ate	discourse marker	Ugandan Bantu	so
ka-	morpheme	Ugandan Bantu	pejorative/diminutive marker
ki-	morpheme	Ugandan Bantu	pejorative/augmentative marker
mbu	discourse marker	Ugandan Bantu	that/reportedly
nga	discourse marker	Ugandan Bantu	how come
nti	discourse marker	Ugandan Bantu	that
oyee	interjection	Lingala	viva/long live
pakalast	adjective/adverb	Kiswahili + English	indefinite/indefinitely/forever

Some examples from local newspapers illustrating the use of the borrowings listed in Table 2 are presented in (2).

- (2) a. *Mbu* his music was turning away people from the church.
(Monitor 27/05/17)
- b. Uganda Cranes *oyee*!
(New Vision 17/01/17)
- c. Bukenya okays M7 to rule *pakalast*.
(Red Pepper 26/07/16)

The examples in Table 2 are mainly cases of indigenous borrowing involving Ugandan Bantu languages, although *pakalast* presents a case of hybrid borrowing involving an endogenous language (Kiswahili) and English. The interjection *oyee* ‘viva/long live’ (2b), on the other hand, is a case of exogenous borrowing, as it

comes from Lingala, a lingua franca in the two Congos, also spoken in the Central African Republic, Angola and Gabon (cf. Bokamba 2012: 292). Two hypotheses suggest themselves for the introduction of this Lingala word in UgE: (i) the popularity of Congolese (DRC) music in the later part of the 20th century, sung in Lingala where words like *oyee* were prevalent; (ii) in the same period, the DRC was led by the dictator Mobutu Seseseko, who had compelled all Congolese to dance for him in cult ideological songs, known as “animation”, which praised him as the “savior” of the country (cf. Udo-Udo 2017: 39). The singing and dancing was obligatorily preceded by and interluded with cheers of hurray such as *Mobutu oyee* ‘viva/long live Mobutu!’ This political leitmotif then spread to neighbouring Uganda and has been used in both political and non-political domains.

2.2.3 Calques

It is a common characteristic of L2 varieties of English to have calqued expressions from indigenous languages (Mesthrie & Bhatt 2008). Instances of calques are given in Table 3.

Table 3. Examples of calques in Ugandan English

Expression	Meaning
below 18 (or any age)	under 18 (or any age)
eat big	get the lion’s share, enjoy something more
go to the bush	start/join an armed rebellion against a government
introduction	engagement/betrothal function
man eateth where he worketh	a person should use the resources at their workplace for personal gains

The calqued lexical items are found in local newspapers and writings authored by Ugandans, as the examples in (3) show.

- (3) a. An orphan is defined as anyone *below 18* years [...]
(Uganda Bureau of Statistics 2016: 21)
- b. At first my parents were okay with anything because they thought I would
have the *introduction* first and at a later time the wedding.
(Monitor 06/02/17)
- c. My maternal uncles, all of them, *went to the bush*. They were rebels.
(Observer 23/05/16)

Just like borrowings, calques are, in the main, supplied by indigenous languages. However, they differ from borrowings in that unlike borrowings, which keep the donor language words, calques use the recipient language words to translate donor language lexical items (cf. Haspelmath 2009: 39). Isingoma (2016: 149ff.) discusses

calques in UgE, where he makes mention of the use of the word *eat* in *eat money*, calqued from the various Ugandan indigenous languages to mean ‘embezzle/swindle money’. In fact, it can also be used to simply mean ‘use/misuse money’, as in (4a–b) below. In addition, the verb *eat* has other metaphorical uses in UgE as illustrated by (4c–d).

- (4) a. [...] unless they want us to give them money and they just *eat*.
(Observer 22/04/17)
- b. Katunguka [the Vice Chancellor of Kyambogo University, bi] said, “[...] others deliberately keep the money to pay it towards end of semester and end up *eating* it with their girlfriends.” (New Vision 02/08/17)
- c. UNRA, ministries *eat big* in shs523b⁵ additional budget.
(Monitor 06/04/17)
- d. This was “common” sense in Uganda, where it was said that “*man eateth where he worketh*”. This did not mean having a cup of tea or a *soda* or even lunch at the place of work but rather that a “wise” man should steal from his place of work and get rich there.⁶ (Suruma 2014: 84–85)

As these examples illustrate, UgE *eat* is semantically richer than L1 English *eat*, because the equivalent of *eat* in Ugandan indigenous languages has additional metaphorical uses that L1 English *eat* does not have. For example, in Runyoro and Rutooro (two western Bantu languages in Uganda), *kulya*, the equivalent of *to eat* can metaphorically collocate with expressions as shown in Table 4.

Table 4. Metaphorical uses of *kulya* ‘to eat’

Runyoro/Rutooro	Literal translation	Meaning
<i>kulya sente</i>	to eat money	to embezzle, swindle, (mis)use money
<i>kulya enguzi</i>	to eat a bribe	to take a bribe
<i>kulya ibanja</i>	to eat a debt	to incur a debt
<i>kulya obukama</i>	to eat the kingship	to become king
<i>kulya obukuru</i>	to eat leadership	to become a leader
<i>kulya amadaraka</i>	to eat enjoyment	to enjoy

It is this wide array of the semantics of the equivalent of *eat* in many Ugandan indigenous languages that feeds into UgE’s metaphorical extension of the word *eat*. Moreover, the UgE adage in (4c) above is an adapted calque from Ugandan languages (but cynically linked to the biblical verse in Ecclesiastes 3:9: *what profit hath*

5. Shs523b is short for 523 billion Ugandan shillings.

6. *Soda* in UgE means any soft drink.

he that worketh in that wherein he laboureth?), where, for example in Runyoro/Rutooro, it is realized as in (5).

- (5) *Embuze eriira nambere bagisubikire.*
Lit: 'A goat eats where it is tethered.'

As is the case in the UgE adage in (4c), the Runyoro/Rutooro adage (5) is used to exculpate someone from any wrongdoing after they have taken advantage of their position to involve themselves in corruption or any other felony.

2.2.4 Semantic change

There are scores of expressions in UgE whose semantics has been changed via semantic extension. Table 3 shows such expressions.

Table 5. Examples of semantic change in Ugandan English

Expression	Meaning/usage
double-decker*	bunk bed
evening	part of the day from 4.00 pm, hence the greeting <i>good evening</i> is used from this time
fallen	deceased (used without restriction to a soldier killed in battle)
grass	suffer, live in poverty
local	not sophisticated, primitive
stroke	take/have an additional subject, item

* See also Buregeya (2007) in relation to Kenyan English.

The lexical items in Table 5 occur frequently in UgE, as shown in the sentences in (6).

- (6) a. One small room is occupied by six *double-decker* beds, implying that it is occupied [...]
(*New Vision* 09/08/12)
- b. Senior medics from the ministry of health and Mulago hospital were at Entebbe International airport to receive the body of their *fallen* colleague Dr Margeret Mungherera.
(*New Vision* 07/02/17)
- c. Sources have revealed that well endowed Nakitto is *grassing* seriously after the *zungu* pensioner dumped her.
(*Red Pepper* 16/10/11)⁷

As can be seen, the instantiations in Table 5 are all cases of semantic extension: *double-decker* (6a) is extended from the Standard English *double-decker bus* to UgE *double-decker bed* ‘bunk bed’. Thus, in UgE *double-decker* means both a type of bus and a type of bed. The meaning of the word *evening* has been extended to

7. *Zungu* is informal for Ugandan English *muzungu*, which means ‘white person’.

include time which is regarded as part of *afternoon* in Standard English: while in UgE *evening* starts at 4.00pm, it starts at “about 6.00 o’clock” (OALD) in Standard English. Similarly, the preposed attributive adjective *fallen*, which is used in Standard English in reference to a soldier killed in a war, has seen its meaning extended to include cases outside military domains and engagements. Furthermore, the lexical item *grass* has been extended metaphorically from its Standard English meaning ‘crop the grass/graze’ to mean ‘suffer/live in poverty’, in the sense that the referent is so impoverished that they cannot afford to buy food and therefore have resorted “to eating grass”. While the Standard English meaning of *local* is attested in Uganda (e.g. in the noun phrase *ministry of local government*), it has also been given a pejorative meaning arising from the advent of western lifestyles, which saw local traditions and practices often denigrated (cf. Okazaki, David & Abelmann 2008: 96). As for *stroke*, its additional semantics in UgE stems from the meaning of the symbol (/), known, among others, as *slash*, *solidus*, *virgule* or *stroke* (cf. Collins Dictionary). The appellation *stroke* is the most common term used in Uganda. Usually used in fractions or to show alternatives, it is also used in Uganda to show a combination of items or subjects (e.g. history/geography, i.e. history *stroke* geography) in the sense that the last item is optionally added. Through the word formation process known as conversion, a verb was derived from the noun *stroke*, with the meaning ‘take/have an additional subject, item’.

2.2.5 Morphological and phraseological innovations

Furthermore, a number of expressions in UgE stem from innovations that involve combining two or more words that usually do not collocate in L1 English with the meaning ascribed to them in UgE, as shown in Table 6.

Table 6. Examples of morphological and phraseological innovations

Expression	Meaning/British or American English Form
borrow a leaf*	take a leaf
cold power	food prepared the previous day
give someone a push	walk/accompany someone
good enough	luckily
pay someone in his/her own currency	have a taste of one’s own medicine
special hire	taxi
turnboy*/turnman†	an assistant to a driver

* See also Buregeya (2007) in relation to Kenyan English.

† Also spelt *turn boy/turn-boy, turn man/turn-man*

Ugandan newspapers use the items presented in Table 6 as shown in the examples in (7).

- (7) a. Fifth, would Ugandans want to *borrow a leaf* from the conduct of business by the Chinese People's Congress [...] (New Vision 01/07/16)
- b. Livingstone Kawuulu, a *special hire* driver working at Wandegeya *stage*, says that killings usually [...] (New Vision 26/01/16)⁸
- c. Later that afternoon, his brother came into the cabin of the tanker as a *turn-boy* and we drove off towards Kenya (Monitor 01/09/13)
- d. Suspected iron bar hitmen on Sunday morning killed a 35-year-old and injured his *turnman* in Gulu Town. (Monitor 04/01/17)

The expression *turn-boy* (7c) (also spelt as *turn boy/turnboy*) 'assistant to a driver' is an East African lexical item, as GLoWbE shows. More importantly, another expression has emerged, that is *turn-man* (also spelt as *turn man/turnman*) as in (7d), to refer to a 'mature *turnboy*'. While *turnboy* is an exocentric compound, on a par with *boyfriend* (i.e. its meaning is idiomatic), *turnman* is an endocentric compound (i.e. its meaning is more or less literal). As is well known, a *head teacher* (endocentric) is a teacher who is a head, but a *sweetheart* (exocentric) is not a heart which is sweet. Thus, the literal interpretation of the compound *turnboy* has made Ugandans or East Africans feel that the expression is not congruent with what a componential analysis of a mature assistant to a driver involves. Studies on L2 idiom comprehension suggest that L2 speakers tend to take into account the meanings of the constituent parts of an idiom (Karlson 2013: 296–7). However, although this strategy may be more prevalent among L2 speakers, it has also been used by L1 speakers in their lexical innovations. For example, in American English, the compound *houseman* is used synonymously with *houseboy* (cf. Oxford Dictionary Online).

2.2.6 Conclusion

From the foregoing, we observe that UgE has different categories of lexical items constituting its lexicon, drawn from various languages both within and outside Uganda. Since some of its lexical items are contributed by endogenous as well as exogenous languages, it is natural for UgE to share some of its lexical items with other L2 varieties. However, generally speaking, the bulk of the UgE lexicon, especially with respect to borrowings, seems to be idiosyncratic to the variety, as the later sections will show, based on the results of their use in larger data sets.

8. *Stage* is a UgE word, meaning 'taxi rank' or 'bus stop'.

3. The lexicon in ICE Uganda

3.1 ICE-UG compilation and issues

The corpus analyses in Section 3 for lexical particularities use two different corpora. On the one hand, we have examined the written part of the Uganda component (ICE-UG) of the *International Corpus of English* (ICE), which has strictly followed the design laid out in Nelson (1996). Collection of ICE-UG began in 2013, at a time when the notion of UgE seemed to be somewhat alien to Ugandan linguists and when the phrase was, in the country, typically used to ridicule the variety of English that develops through the process of nativisation. Once contact with Makerere University (based in Kampala) and Gulu University had been established, data collection commenced and all data was then edited, annotated and saved at Ruhr University Bochum. The written component of ICE-UG was completed and released to the scientific community in February 2018. Its word count, including extra corpus passages, amounts to 431,776 words.

The data is somewhat biased in that it does not representatively reflect the different ethnicities and first languages found in Uganda. The written data has a considerable bias towards English used by speakers who have Luganda as their L1, which is particularly the case for public genres such as newspaper editorials or administrative writing. Furthermore, male users of English are overrepresented. The reason is that writings in Ugandan English are, predominantly, by L1 speakers of Luganda, which is due to the special position of this ethnic group in the Ugandan society: having been the first ethnicity or region to be under British protectorate, the Baganda were also the first to receive formal education in English and were favoured in the British administration of the protectorate. As a result, the region is better developed, and individuals have a better chance of higher education and public jobs.

However, the bias in the corpus might reflect the uses of English in that the Central region relies more on English and male users are more likely to produce writings in English.⁹ Nevertheless, statements regarding the influence of Luganda on UgE must be taken with a pinch of salt as they might be a result of this bias (further research into English as used by speakers of different L1s is necessary to tease out how intense Luganda influence is in the L2 Englishes).

The following section discusses the lexical particularities described in Section 2 in the data as well as a number of corpus linguistics delineators.

9. This assessment is based on the fact that, in Uganda, there are still more males than females pursuing education (cf. Uganda Bureau of Statistics 2016: 26). Also, males attain higher levels of education. As a result, they are more likely to occupy professional positions which involve writing than females.

3.2 Lexical properties of ICE-UG

The written part of ICE-UG has 33,364 different word types. Related to its 431,776 tokens, the data has a type-token ratio of 7.73%. This ratio allows for a first assessment of the richness of a vocabulary or of how “lexically varied” the data is. The written component of ICE-SL (Sri Lanka) has 34,582 types, and that of ICE-CAN (Canada) 33,972, so there does not seem to exist a crucial difference in terms of lexical richness between L2 corpora of different regions nor between the two L2 and the L1 corpus.

3.2.1 Borrowings in ICE-UG

As has been discussed in Section 2, UgE borrows from the nation’s indigenous and endogenous (Kiswahili) as well as exogenous languages. The written component of ICE-UG contains a total of 972 (i.e. 0.23%) indigenous words, excluding personal, place and street names, so that indigenous words make up 2.25 per 1,000 tokens. In comparison to other corpora within the ICE family, this is not unusually high or low, but considerably higher than the figures for ICE-EA (Kenya) and ICE-SL, as Table 7 shows.

Table 7. Indigenous words in selected L2 ICE corpora¹⁰

	Indigenous wordstock	Total word count	Indigenous words per 1 million words
ICE-UG	972	431,776	2,251
ICE-EA (Kenya)	552	403,380	1,368
ICE-EA (Tanzania)	765	393,791	1,942
ICE-SL	845	645,566	1,308
ICE-IND	1,762	588,855	2,992
ICE-PHI	1,245	591,793	2,103

The 972 tokens relate to 153 types. Not surprisingly the majority of these are broadly related to culture in the wider sense of the word, as Figure 1 demonstrates: 25% are labels for the various Ugandan ethnicities, 14% refer to culture and customs, and 13% to languages. Another large portion is made up by those types that describe food and drinks (12%), kinship (7%), and flora and fauna (6%). This is very much in line with what Schmied (2012) describes for what may be perceived

10. The various corpora were assembled at different points in time, which might have affected the results. Whilst data for ICE-UG was collected between 2013 and 2017, ICE-EA, ICE-IND, ICE-PHI contain texts from the 1990s. ICE-SL was compiled in two phases, with some of the data dating back to the early 1990s and others to the early 2000s.

as a standardised East African English.¹¹ An interesting 5% of all types ($n = 19$, 6 types) constitute function words: the data comprises a number of conjunctions as well as the diminutive prefix *ka-*.

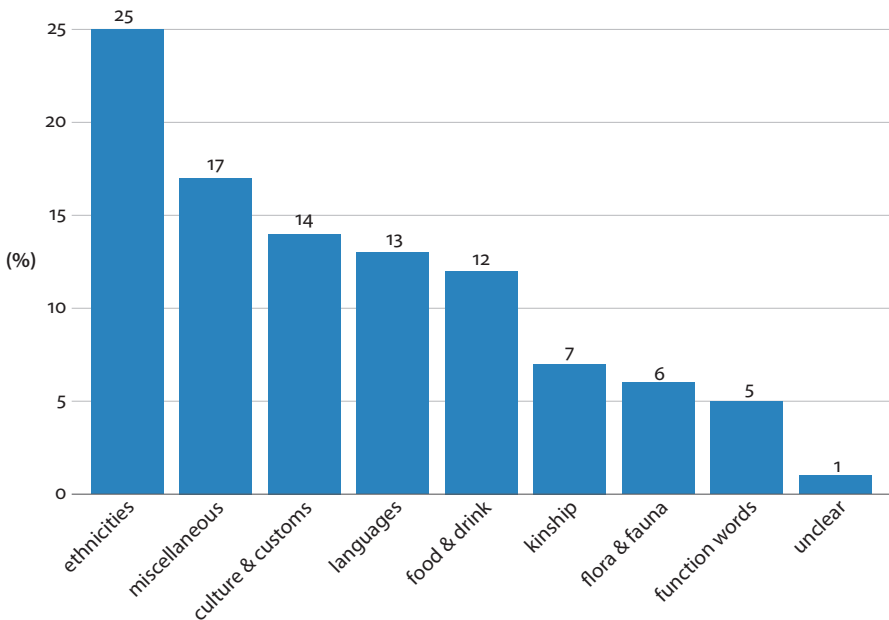


Figure 1. Borrowings by semantic field

Relatedly, 80% of the words are nouns. Other word classes are nouns/adjectives (7%), conjunctions (4%) and adjectives (3%), whilst verbs and prefixes make up only 1% each, and an additional 4% are unclear cases. Thomason (2001: 70) explains that the nature of borrowing allows for an assessment of the intensity and quality of language contact, of which borrowing is a result. Function words, such as conjunctions, indicate “slightly more intense contact” with their users being “reasonably fluent bilinguals”, who are, however, a minority. Derivational affixes, represented in UgE through *ka-*, on the other hand, are said to reflect “more intense contact” and a growing number of bilinguals. Given the current spread of English proficiency, resulting from free primary education (which in Uganda spans the

11. As Schmied (2012) explains, *standardised* does, in the case of East Africa, not suggest codification, as there are no official dictionaries, nor unofficial ones or precise and comprehensive descriptions at the level of grammar. In the light of a lack of systematic data collection, which would be necessary to serve as a basis for codification and eventual official publication, he states that “the standardisation debate has hardly begun” (2012: 235).

first seven years), in which English is the medium of instruction from year 5, the results reflect this development towards increasing uses of English by L2 speakers.

In terms of their L1 origin, 69% of the borrowed items originate from Uganda’s Bantu languages. Given the fact that most of the Bantu languages share a vast part of their word stock, it is not evident from which language precisely individual words originate. However, 3% could be clearly identified as Lugisu (a dialect of Lumasaaba) and 1% each of Lumasaaba, Rukonzo and Runyankore (in these cases, the precise origin was evident from the fact that the word referred to the ethnicity or language or objects unique to the ethnicity). The second most frequent donor language is Kiswahili with 13% and Acholi (the most frequently spoken Nilotic language in the North of Uganda) follows third with 5%. A number of other languages contribute to the lexicon of UgE at a ratio of 1% in ICE-UG. These comprise the indigenous language Ateso (another Nilotic language) but also three endogenous languages: Arabic, Hebrew and Hindi. A further 1% is hybrid (involving the prefix *ka-*) and another 7% could not be allocated.

The picture that emerges from this investigation is one of a clear dominance of Luganda, an artifact of the language’s overrepresentation in ICE-UG, but also potentially reflective of the socio-economic make-up of Uganda’s society. The Baganda, the ethnic group speaking Luganda, were the first to be ‘protected’ by the British and, as a result, also the first to encounter missionaries and missionary teaching. As a result, they were also the ones to be employed in the British administration and, thus, obtained influential positions. Their socio-economic dominance is also reflected, as discussed above, in their presence in politics and the media.

Not surprisingly, the individual borrowings are attested with highly heterogeneous frequencies. This is also the case for those items, which were discussed in Section 2, as indicated in Table 8.

Table 8. Selected borrowings in ICE-UG

Item	Attestation in ICE-UG (per 1 million words)
bodaboda (reborrowed Bantu)	43.98 (19)
chapati (Hindi)	34.72 (15)
posho (reborrowed Bantu)	16.20 (7)
mailo (reborrowed Bantu)	13.89 (6)
matooke (Bantu)	9.26 (4)
sim sim (Arabic)	4.63 (2)
oyee (Lingala)	2.31 (1)

Similarly, of those items discussed in Isingoma (2016), only very few items were attested with frequencies suggesting that they were not nonce words. *Kabaka* occurs 39 times, but of these 32 are from seven texts in the academic and popular

humanities writing sections in which the traditional system of rule was discussed. Similarly, *kwanjula* occurs 25 times, but of these, 22 are in one text only, and *gomesi* has twelve tokens, but 12 of them in just one W2B humanities text. Those items that occur somewhat frequently and spread across texts are *bodaboda* (19), *kaveera* (11), *posho* (7) and *matooke* (4). Items that occurred only once are *katikikiro*, *lukiiko*, *rwot*, *saza*, *bushera*, and *katogo*.

Genre effects:

The borrowings found were extremely unevenly spread across the various written genres captured in ICE-UG,¹² as summarised in Table 9. As this demonstrates, the vast majority are in popular writing humanities, social letters, academic writing from the humanities, and in fiction. As regards the humanities, this signifies that items were discussed for which there was no English equivalent, for example, cultural concepts. The categories of social letters and fiction, however, seem to indicate that more informal genres contain more borrowings.

Table 9. Distribution of borrowings (tokens) across ICE genres

ICE-category	Percentage
W2B humanities	31.48
W1B social letters	23.66
W2A humanities	21.40
W2F	15.33
W2D skills	2.57
W1A untimed essays	1.65
W2E	1.65
W2B social sciences	1.34
W2A social sciences	0.82
W1A exam papers	0.10

By contrast, no borrowings were found in business letters, administrative writing, popular and academic writing in the natural sciences and technology and in newspaper editorials. This result is also of relevance when it comes to interpreting the fact that many of the items listed in Section 2 are not attested in the data. It is likely that they are representative of a genre more informal than the ones captured in ICE-UG.

It does seem, however, that borrowings are not necessarily what renders the lexicon of one variety different from another. This is not really surprising, given the

12. The text genres captured in the ICE corpora are as follows: W1A – student writing, W1B – letters, W2A – printed academic writing, W2B – printed popular writing, W2C printed reportage, W2D – instructional writing, W2E – persuasive writing (press editorials) and W2F – creative writing. For details see Nelson (1996).

fact that they constitute a minor share of the variety's word-stock. The top-11 borrowings attested most frequently in ICE-UG are *empaako* ('praise name', 108.8 pmw, 47), *kanzu* (a kind of robe/tunic, 50.93 pmw, 22), *kwanjula* ('betrothal/engagement function', 46.3 pmw, 20), *gomesi* (female ceremonial wear, 37.04 pmw, 16) and *kabaka* ('king of Buganda kingdom', 37.04 pmw, 16), *eshabwe* ('ghee sauce', 32.41 pmw, 14), *kaveera* ('plastic bag', 25.46 pmw, 11), *bodaboda* ('motorcycle/bicycle taxi' or 'taxi man', 23.15 pmw, 10) and *ssenga* ('paternal aunt'/'cultural trainer', 23.15 pmw, 10).

A comparison of the wordlists for ICE-UG and ICE-EA finds these highly frequent borrowings ranked at 211 (for *bodaboda*) and lower in the keyword list. Keyness generally indicates an unusual (in)frequency of a particular lexical item in a corpus, in comparison to a (larger) reference corpus (see e.g. Weisser 2016: 173, 273). At the top of the keyword list, that is at higher levels of keyness, words are English word stock that differentiates the two varieties. Additionally, several of the borrowings listed above may be artefacts of the data collection process informing the ICE corpora. For example, *empaako* is overrepresented due to a scholarly article dealing with it, and the same is the case with *eshabwe*, which occurs in a paper discussing marriage ceremonies. More revealing results require larger corpora, as will be discussed in Section 4.

3.2.2 *English word stock in ICE-UG*

A closer look at the English words that have been identified as Ugandanisms in Section 2, due to their particular semantics, presents a similar picture as discussed above. They are captured in ICE-UG, but with low frequency. Whilst *to eat big* was not found in the data, there was a single item each for *went to the bush*, *shamba boy* and the archaism *thrice*.

Similarly infrequent in the data are those calques discussed in Isingoma (2016): *Ok please* (an expression of agreement and goodwill) and *to be lost* (in *you are lost*, meaning 'long time no see') do not occur at all, and the greeting *Well done* just twice, in reported speech in a fictional text. However, *introduction*, which refers to a betrothal/engagement ceremony, is attested 15 times, in various texts spread across the three genres of social letters, popular humanities writing and fiction.

Those items that had been discussed in Section 2 as having undergone semantic change as well as innovations could all not be found in ICE-UG. Given the fact that they were identified in major national newspapers in Uganda in the first place, however, suggests that it might be the comparatively small size of ICE-UG, together with the genres it represents, that causes their absence in the data. Section 4 suggests, therefore, complementing ICE-UG with evidence from a larger, web-based, corpus and from elicited data.

4. Limitations of ICE-UG and suggested supplements

As we have indicated in the introduction to this chapter, the nature of the ICE corpus design has an impact on studies at the level of the lexicon. Due to the small size of the carefully balanced ICE corpora individual lexical items do not occur in the data, whilst others occur to extents that are frequently artefacts of the data collection process. Mair (2007) has discussed the potential and usefulness of combining controlled corpora with web-based data, and the potential of the latter has further been emphasised by Davies (2010). Besides this, scholars such as Rosenbach (2013) have advocated multi-method approaches to arrive at a fuller understanding of the nature of second language English varieties. Whilst controlled as well as very large corpora enable us to arrive at a precise description of the use of particular items in these data, both lack explanatory power when it comes to assessing the acceptability of and familiarity with these words by those users of English apart from the authors represented in the data. Over the last couple of years, multi-method approaches have been suggested in various sub-disciplines of linguistics.

As we have discussed above, many of the lexical items considered characteristic of UgE do not show up in the written component for ICE-UG. We also mentioned above that the small size of just 431,776 words might account for this fact, since studies at the level of the lexicon typically warrant a larger corpus. Given the comparatively small size of ICE-UG, which is somewhat unfortunate for analyses at the level of the lexicon (cf. e.g. Meyer 2002: 14ff.), we also collected data of web-based UgE (Web-UG). Using SketchEngine's (Kilgarrieff et al. 2014) WebBootCat function, we followed the procedures of Davies and Fuchs (2015) used to compile GloWbE. Highly frequent 3-grams were employed as seeds to retrieve Ugandan websites. In a second step, websites that were not genuinely Ugandan, for example, mirror sites of Google, were excluded from the further procedure, and data was downloaded between May and July 2017 until the corpus had reached a size of approximately 12,300,000 words. Admittedly, this is a 'quick and dirty' corpus, with data being less controlled than the ones in ICE-UG. At the same time, however, the data is reflective of recent UgE.

As we have explained, unfortunately the spoken component of ICE-UG is still too small (currently comprising just approximately 100,000 words in the categories face-to-face conversations, telephone conversations, scripted and unscripted speeches) to allow for a useful analysis. However, we assume that the fact that a word is not attested in either ICE-UG or Web-UG does not imply that it is not used by Ugandan speakers of English. This section explores how low attestation in the corpus data correlates with acceptability and passive knowledge. To this end, two sets of data were collected.

Acceptability tests are generally held to provide insights into informants' prescriptive attitude towards individual grammatical structures or lexical items (e.g. Loewen & Reinders 2011). They have typically been employed widely in L1 communities. Studies in L2 varieties have documented that, frequently, responses do not indicate a clear picture of prescriptive attitudes but rather point towards insecurity and rather tap into the informants' interlanguage (e.g. Mackey & Gass 2016: 59). Additionally, a production test was administered to probe even more carefully into informants' factual familiarity with individual lexical items (cf. Taguchi 2013). This methodological triangulation aimed at ensuring that the strengths of either test type would neutralise the lacunae posed by the weaknesses of the other (cf. Hussein 2009). 41 speakers of English were involved, all of whom had completed at least secondary education (13 years of schooling). All informants are undergraduate students of English at Gulu University and teachers at both primary and secondary level. The respondents were selected purposively and by convenience in order to ensure that each of the three main language groups in Uganda (Bantu, Nilotic and Central Sudanic languages) was at least represented and only speakers who were readily available were contacted. The motivation to involve mainly students of English was to ensure that respondents were "experts" in English. Teachers were included due to the fact that they use English on a daily basis, as it is the language of instruction from the middle primary cycle onwards. As daily users of English in the country, the teachers' linguistic behaviour portrays entrenched forms which they pass on to their students, thereby providing further factual evidence of the nativisation of English in Uganda. The informants were aged between 20 and 40 years and were a mix of male (11) and female (30) (see details in Appendix 3). The tests were supervised to ensure that informants answered the questions as naturally as possible, without consulting dictionaries or other materials that would distract them from their UgE usage.

4.1 Ugandan English on the Internet: Web-UG

The Web-based UgE corpus (Web-UG) was scrutinised for the different items already discussed in Sections 2 and 3. As Table 10 demonstrates, the figures (attested instances per 1 million words) vary considerably across the individual lexical items. The most frequent borrowing is *bodaboda*, whilst others do not occur even in Web-UG. For those items that could be attested in Web-UG, corpus linguistics furthermore allows for an assessment of whether these are characteristic of UgE or rather (East) African English or L2 Englishes in more general terms or spread world-wide across both L1 and L2 varieties. Since 2015, GloWbE (Davis & Fuchs 2015) allows for comparisons across a vast number of varieties of English and thus an assessment of the uniqueness of an item to an individual variety, the results of which are also summarised in Table 10. GloWbE's national sub-corpora are of

different sizes as outlined in Davies and Fuchs (2015: 6).¹³ Whilst the web-interface of GloWbE indicates relative frequency or keyness across the corpora through shading the search results, word-per-million ratios were calculated to incorporate the different sizes of the corpora and to allow for a precise comparison with Web-UG. In any case, this enables an assessment of whether individual lexical items are uniquely Ugandan or whether, which is more likely, differences are more a matter of preference and frequencies. Such an assessment furthermore allows for an identification of individual terms that are preferred regionally, throughout East Africa (cf. e.g. Schmied 2012 for similarities and differences in the lexicons of East African Englishes) or Africa, and of items that are L1- or L2-specific.

Table 10. Borrowings in Web-UG and GloWbE¹⁴

Item	Attestation in Web-UG	Attestation in GloWbE
bodaboda	27.0 (333)	KE 2.92 (120); TZ 1.22 (43)
matooke	15.9 (185)	GB 0.00 (3), IE 0.02 (2), AU 0.00 (1), NG 0.05 (2), GH 0.15 (6), KE 0.41 (17), TZ 0.17 (6), JM 0.8 (7)
mailo	9.18 (113)	KE 0.17 (7), TZ 0.03 (1)
posho	6.42 (79)	US 0.00 (1), AU 0.00 (1), BD 0.03 (1), KE 0.56 (23), TZ 1.25 (44)
chapati	4.72 (58)	US 0.02 (7), CA 0.04 (6), GB 0.03 (11), AU 0.04 (6), NZ 0.01 (1), IN 0.47 (45), LK 0.09 (4), PK 0.58 (30), BD 0.43 (17), SG 0.4 (17), MY 0.35 (15), ZA 0.04 (2), NG 0.05 (2), GH 0.03 (1), KE 2.9 (119), TZ 2.87 (101)
kyeyo	3.98 (49)	GH 0.05 (2)
yaka	3.25 (50)	0
balokole	3.09 (38)	KE 0.02 (1)
bwana	3.01 (37)	KE 6.28 (258), TZ 4.49 (158)
simsim	1.87 (23)	KE 0.07 (3), TZ 0.31 (11)
kadongo kamu	1.63 (20)	0
kisanja hakuna	1.06 (13)	0
mchezo		
entandikwa	0.81 (10)	KE 0.02 (1)
bayaye	0.73 (9)	0
muzigo	0.73 (9)	0
sukuma wiki	0.73 (9)	AU 0.00 (1), IN 0.04 (4), KE 2.63 (108), TZ 0.17 (6)

(continued)

13. The sub-corpora relevant for our further discussions have the following sizes: Ghana 38,768,231 words, Kenya 41,069,085 words, Nigeria: 42,646,098 words, and Tanzania 35,169,042 words (cf. Davies & Fuchs 2015: 6).

14. Here, as in Table 11, figures are per 1 million words with raw frequencies given in brackets, sorted by frequency of occurrence in Web-UG.

Table 10. (continued)

Item	Attestation in Web-UG	Attestation in GloWbE
pakalast	0.49 (6)	TZ 0.03 (1)
shamba boy	0.49 (6)	AU 0.01 (2), KE 0.12 (5), TZ 0.09 (3)
boo	0.33 (4)	0
oyee	0.33 (4)	KE 0.07 (3), GH 0.03 (1)
malakwang	0.08 (1)	0

Whilst *kabalagala* is unattested even in the 12.3 million Web-UG,¹⁵ several items are not exclusive to the lexicon of UgE but are shared with other varieties, particularly East African ones; for example, they are attested in data obtained from Kenyan and/or Tanzanian websites. In these cases, however, there are differences in terms of frequency and preference. For most items, frequency was considerably higher in Web-UG than in the GloWbE subcorpora. However, and not surprisingly, the situation is different for borrowings that originate from Kiswahili, which is considerably more widespread in Tanzania and Kenya. In fact, both *bwana* and *sukuma wiki* are significantly (not in the statistical sense) more frequent in these two countries than in Uganda. Both *matooke* and *posho* seem to have spread considerably beyond Uganda, indicating that they are now part of the international lexicon of English. Schmied (2012: 245) documents a similar spread for “very old borrowings” such as *askari* (‘soldier/guard’) or *bwana* (‘Mister’). As he points out, they are “restricted to African contexts”, and in fact the items *matooke* and *posho* are only used when authors refer to eating in Uganda. Interestingly, *chapati* is highly frequent as well, attesting how much the originally Indian bread has been adapted in the local food culture. Despite the fact that Uganda lost almost all of its Indian community in the 1970s, *chapati* is more frequent than in Kenya and Tanzania. Surprisingly, in all three corpora the item occurs much more often than in those representing countries on the Indian subcontinent.

A check for those items mentioned in Section 3 as top-11 borrowings in ICE-UG reveals that, as has been indicated in Section 3, several of the high frequencies are in fact artefacts of the data collection process underlying ICE-UG. In the less controlled but potentially more everyday data assembled in Web-UG: the figures are as follows: *empaako* (5 of which 4 refer to the Empaako Monument in Fort Portal, 0.08 pmw, 1), *kanzu* (2.36 pmw, 29), *kwanjula* (, 2.44 pmw, 30), *gomesi* (3.17 pmw, 39), *kabaka* (119.59 pmw, 1,471), *eshabwe* (0.16 pmw, 2), *kaveera* (1.87 pmw, 23),

15. *Kabalagala* may be a lexical item that is on its way out of the UgE lexicon. In fact, the text in the Daily Monitor from which the example was taken mentions that the snack today has several other names. This seems to document that the lexicon of UgE, like that of any other variety, is dynamic and constantly changing.

bodaboda (27 pmw, 333) and *ssenga* (4.07 pmw, 50). In sum, *empaako* and *eshabwe* especially drop considerably in terms of their attestation in the data. Most other items still have high frequencies, which is not surprising given the fact that they are cultural borrowings that do not have an English equivalent and, at the same time, are regularly used when it comes to engagement and wedding ceremonies, but significantly smaller frequencies than in ICE-UG. A case in point, however, is *kabaka*, for which the frequency has risen considerably in Web-UG as compared to ICE-UG.

Of the three discourse markers *mbu*, *nga* and *nti*, only *mbu* occurs somewhat regularly, attested 40 times (30.77 pmw), whilst *nga* and *nti* are rather marginal with three (0.24 pmw) and one (0.08 pmw) occurrence respectively.

For the analyses of calques, the hits returned from the GloWbE web-interface needed to be manually searched to examine whether they had the same semantics as the ones identified as characteristically UgE in Section 2. For example, a search for “GO to the bush” (which covers *go*, *goes*, *gone*, *went*) returned 590 hits, many of which were attested in the US subcorpus. However, these were unrelated to the meaning ‘to join a rebel army’.

Table 11. Calques in Web-UG and GloWbE

Item	Attestation in Web-UG (per 1 million words)	Attestation in GloWbE
EAT big	2	0
GO to the bush	7.23 (89)	GH: 0.01; KE: 0.22; NG 0.02; TZ: 0.09
introduction	2.93 (36)	

The calques are, however, attested in several other African varieties of English, as a search in GloWbE revealed. A search for the collocate *eat* and *big* in GloWbE, however, did not yield any results for the semantics discussed in Section 2. Quite to the contrary, the phrase *go to the bush* seems to be a general African expression. The 89 occurrences in Web-UG are matched by four in the Ghanaian, nine in the Kenyan, 1 in the Nigerian and 3 in the Tanzanian subcorpus of GloWbE.

A similar strategy was adopted for those English items that have undergone semantic change in the examples discussed in Section 2. In the case of *local* and *Good Evening!*, searches were not meaningful, since too often the context (i.e. whether *local* was intended to signify the meaning discussed in Section 2, and whether *Good Evening* related to a time before or after 6pm) was lacking. Furthermore, a search for *local* returned a large number of hits that do not necessarily reveal whether the item is used with a negative or neutral connotation. The item has 5,160 hits. The Ugandan usage of *local* with a derogatory connotation is an example of uses escaping corpus linguistic methodology. The items that are attested beyond the level of nonce-words are *double-decker bed*, *fallen colleague*, and *to be grassing*.

Table 12. Lexical items having undergone semantic change

Item	Attestation in Web-UG	Attestation in GloWbE
double-decker bed	0.24 (3)	SG 0.05 (2), TZ 0.03 (1)
fallen colleague	0.24 (3)	0
to be grassing	0.08 (1)	0 in this meaning

As Table 12 shows, the items occur more frequently in Web-UG than in either of the subcorpora of GloWbE, if they occur there at all. However, there seems to be a crucial difference in the usage of these items, compared to the borrowings discussed above. Whilst for the latter considerable numbers were attested with high frequencies, both in ICE-UG and in Web-UG, this is not the case with the items discussed here.

A slightly different picture emerges from the analyses of Web-UG and GloWbE as regards innovative uses of English word stock. Whilst *cold power* and *give someone a push* were not attested in Web-UG, several other expressions occurred, albeit with very diverse frequencies, as Table 13 indicates. From the corpora, it seems that *good enough* and *special hire* are Ugandanisms in that they occur with considerable higher frequencies than in any other variety (as represented in GloWbE).

Table 13. Innovations using English word stock

Item	Attestation in Web-UG	Attestation in GloWbE
cold power	0	0
give someone a push	0	KE 0.02 (1)
good enough	0.89 (11)	NG 0.14(6)
pay in one's own currency	0.08 (1)	0
special hire	1.46 (18)	TZ 0.03 (1), JM 0.03 (1)
turn boy / turnboy	0.65 (8)	Turn()boy 2 KE, 1TZ Turn(man) 2KE
to mug	0	0

Finally, UgE behaves like other postcolonial Englishes (PCEs) as regards the use of lexical items that “have become archaisms in metropolitan Englishes” (Mesthrie & Bhatt 2008: 115). For example, *thrice* occurs once in ICE-UG and 65 times (5.28 pmw) in Web-UG, but is also widely attested in all the various subcorpora of GloWbE, with Pakistan standing out.

In sum, the searches reveal that a large number of items that were discussed in Section 2 do not even occur in Web-UG. There are several interpretations for this observation. On the one hand, the items may be idiosyncratic nonce words and innovations evinced by the authors of the individual newspaper articles from

which WebBootCat retrieved the data. In fact, as Mesthrie and Bhatt (2008: 115) emphasise, “as living rather than fossilised systems, New Englishes generate new colloquialisms too.” At the same time, there is a strong likelihood that individual items are not represented in Web-UG as they are more characteristic of spoken UgE, which is not captured in the data, but which is where colloquialisms have traditionally developed more easily than in written genres.

4.2 Familiarity and attitudes: Results from acceptability ratings

In the acceptability test, informants were asked to rate the following test items on a 5-point Likert scale, ranging from ‘good’ (1) via ‘ok’ (2), ‘neutral’ (3) and ‘odd’ (4) to ‘bad’ (5). Additionally, they had the option of indicating ‘unclear’ for items they did not understand, to avoid responses unrelated to factual assessments. In total, the test originally comprised 43 items, which included a number of distractor items that were standard British English sentences and one containing a South African lexical item, robot, ‘traffic lights’. The lexical items under scrutiny¹⁶ have been italicised in the list below:

- a. Since she is *below* 18, she had to ask her parents.
- b. Ministries are *eating big* on the new budget.
- c. The president seems to rule *pakalast*.
- d. The room in the hostel had six *double-decker* beds.
- e. After having lost his job, Samuel is *grassing* seriously.
- f. You can harvest *sukuma wiki* all year round.
- g. In A level, he *stroked* Divinity with English.
- h. Livingstone Kawuulu is a *special hire* driver working in Kampala.
- i. I have been to this place *thrice*.
- j. *Kabalagala* is served in many eateries, with tasty ingredients.
- k. A ceremony will be held in honour of our *fallen* colleague.
- l. Oh it is 4pm and I have not greeted you; *Good Evening!*
- m. *Good enough*, they had locked the door, so the thief could not enter.
- n. The minister insisted on death penalty; now he is being *paid in his own currency*, since his own brother has been sentenced to death.
- o. As I was *giving* my friend a *push* in the evening, I was bitten by a snake.
- p. He is a strong boy, so he can *mug* this box.
- q. Do you enjoy eating *cold power*?

16. The distractor items were not analysed. Furthermore, we exclude in our discussion those items for which the corpus analyses as well as the test results yielded no or uninterpretable results.

The results, visualised in Figure 2 (see Appendix 1 for the raw data), indicate that results fall into two groups. For some items, there was a clear (de)cline indicating that respondents had a precise understanding of the context as well as an attitude regarding their acceptability. This is the case with the items *below 18*, *sukuma wiki*, *special hire*, *thrice*, and *good enough*, which were all largely considered appropriate.

In other cases, the lines indicate complex and unclear results. Frequently, respondents seemed split as regards the acceptability of individual lexical items, for example, *to eat big*, *pakalast* etc.

In total, the enormous variability of the results in the acceptability test point towards the fact that in Uganda, as in many postcolonial nations, users of English represent a highly heterogeneous community. Whilst some have completed language shift to English as their first language or are balanced bilinguals equally proficient in English and an indigenous languages, many others, particularly in rural areas (similar to what Michieka 2009 discusses for rural Kisii in Kenya) rarely use English and use it as a foreign language at best. The majority of users will most likely be somewhere between these two extremes. Together with the highly exonerative orientation of English language teaching, test items representing indigenised uses may be rated unacceptable or less acceptable for two reasons: (a) because the informant has not developed a stance towards nativised Ugandan use of English and (b) because s/he is torn between what s/he encounters in daily use and what is prescribed in textbooks.

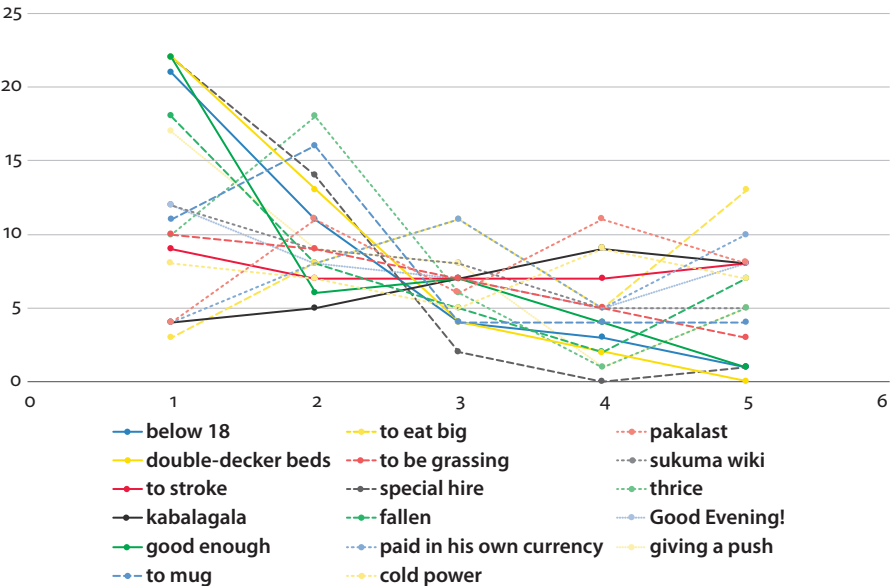


Figure 2. Acceptability ratings for selected items

We next looked at how rating correlates with frequency in the corpora, that is, whether more frequent expressions are rated more clearly than less frequent ones, to tease out whether frequency of exposure correlates with higher acceptability. In fact, acceptability ratios indicating that the item was rated acceptable, but also that there was a clear cline in responses, correlated with relative higher occurrence in Web-UG. Thus, the items with the highest acceptability ratio, *special hire* (1.51), *thrice* (1.66), *below 18* (1.76) and *good enough* (1.9) also occurred relatively frequently, albeit not in the same order (cf. Figure 2 above). With items that are attested less frequently, informants either showed a clear cline but rated them as less acceptable (e.g. *double-decker bed*) or the ratings were distributed across the scale in a chaotic way (e.g. *to be grassing*).

Since the ratings alone do not allow for an assessment as to how familiar respondents are with the test items, a production test was also carried out.

4.3 Factual knowledge: Results of the production test

The production test required the speakers to answer the following questions, which required them mainly to provide meanings of the lexical items as well construct sentences thereof.

- i. What does *She is a local girl* mean?
- ii. At what time of the day do people start using the greeting *good evening*?
- iii. a. Is this sentence grammatically correct? *Good enough, I had locked my door.*
b. If incorrect, please correct it.
- iv. a. Have you ever heard of the following expressions?
to pay one in his/her own currency
to give someone a push
to stroke (e.g. In A Level, I stroked Divinity.)
to mug (e.g. He is a strong boy, so he can mug this box.)
cold power
b. If yes, what does each expression in (a) mean? Construct a sentence using each of the expressions.

The results of the test are summarised in Tables 14 and 15, whilst examples of the sentences constructed by the informants are presented in Appendix 2.

Table 14. Results for questions (i) to (iii)

Expression	Meaning as discussed in Section 2	Other meaning	Invalid/blank
local	87% (36/41)	12.2% (5/41)	00.0% (0/41)
good enough	78% (32/41)	22% (9/41)	00.0% (0/41)
good evening	78% (32/41)	17.1% (7/41)	4.9% (2/41)

While the expression *good enough* ‘luckily’ is not attested in ICE-UG, it is attested in Web-UG, with 11 entries, and the acceptability ratings show that a sizable majority of the informants accepted its UgE usage (cf. (8a)), with a ratio of 1.9. These findings are confirmed by the production test results, where, despite the fact that only 78% agreed that the sentence in which the expression was used was correct, only 1 out of the 9 informants, who said that the sentence was incorrect, was able to “correct” the sentence, as shown in (8b). The rest who tried to “correct” it, provided sentences (some of which are given in (8c)) that are akin to the UgE usage.

- (8) a. Good enough, I had locked my door.
- b. I was lucky that I had locked my door.
- c. i. Good enough, I locked my door.
- ii. Good enough, I have locked my door.
- iii. Good enough! I had locked my door.

Thus, the attestation of *good enough* in Web-UG, its overwhelming acceptability and elicitation in the tests synergistically show that it is a well-established expression in UgE. The other two expressions (*local*, *good evening*) are generally deemed appropriate UgE lexical items, as the results in the table show. Table 15 provides results in response to question (iv) above.

Table 15. Responses to question (iv)

Expression	Heard of it	Not heard of it	Invalid	Meaning as discussed in Section 2	Other meaning
cold power	80.5% (33/41)	17.1% (7/41)	2.4% (1/41)	97.0% (32/33)	3.0% (1/33)
to give someone a push	100% (41/41)	00.0% (0/41)	00.0% (0/41)	95.1% (39/41)	4.9% (2/41)
to mug	85.4% (35/41)	9.7% (4/41)	2.4% (1/41)	97.1% (34/35)	2.9% (1/35)
to pay someone in his/ her own currency	60.9% (25/41)	36.5% (15/41)	2.4% (1/41)	52.0% (13/25)	48.0% (12/25)
to stroke	87.8% (36/41)	9.7% (4/41)	2.4% (1/41)	86.1% (31/36)	13.9% (5/36)

Interestingly, the expression *give someone a push* ‘walk/accompany someone’ is unattested in either corpus, and in the acceptability test, it received a ratio of 2.2, but was recognised by all the informants as an expression they have heard. As for its meaning, 95.1% of the informants gave a meaning that matches the UgE meaning and 4.9% gave a different meaning. Examples of sentences produced by the informants are in Appendix 2. A similar trend, albeit to a lesser extent, is observed for the expressions *mug*, *cold power* and *stroke*. Remarkably, the expression *pay one*

in his/her own currency ‘get a taste of one’s own medicine’, with a 2.47 ratio in the acceptability test, was only recognised by 60.9% and out of these only 52% provided a meaning that is consonant with the UgE usage. One reason could be that the expression is used idiomatically and idioms are well known as infrequent lexical items both among L1 and L2 speakers of English (cf. Karlson 2013; Biber et al. 1999). As an infrequent lexical item, many of the respondents may not have had it in their lexical repertoire. In fact, in some cases, a literal interpretation was provided by some informants, which is not surprising, since L2 speakers tend to disregard the holistic meaning encoded by an idiom and opt for retrieval of meaning based on individual meanings of the constituents of the idiom (Karlson 2013).

5. Conclusion

The various results indicate the power of multi-method approaches that add to pure corpus analyses. ICE-UG allows for investigations of uses in particular genres and thus insights that a lot of the characteristic borrowings, calques and uses of English words with adapted semantics occur in more informal genres. However, its small size is clearly an obstacle in that certain lexical items are not captured whilst others are overrepresented due to the data collection process and the corpus design principles. Web-based complements allow us to put into perspective such biases and to compare findings across corpora reflecting other varieties around the world. However, some of the neologisms and the creative language use that are characteristic of L2 varieties at the nativisation stage, as reflected in current language use in Uganda’s newspapers, particularly in their online versions, do not occur in either of the two corpora. How these are spreading can, and in our view should, be assessed from experimental data, which here revealed that uses in newspapers may not be necessarily shared by all L2 English speakers in a country.

Whilst the corpora point towards bilinguals’ creative use of their full linguistic repertoire, the tests present a more precise picture of the current stage of the use of English in Uganda’s society at large and of the processes characterising this. Semantic change and lexical innovations in Ugandan English seem to follow a pattern that has been found with most instances of language change and which Schneider (2007: 85) describes for innovations in the nativization phase: they spread following an S curve, with “variant forms used by some while at the same time a majority of others will stick to the old patterns,” and linguistic innovations in fact “begin with occasional uses, perhaps preferred uses, in the speech of some individuals”. In fact, the low frequencies of attestation in the corpora may be indicative of early stages of innovation at the level of the lexicon in Uganda.

References

- Algeo, John. 2010. *The Origin and Development of the English Language*, 6th edn. Boston MA: Wadsworth.
- Batibo, Herman M. 2005. *Language Decline and Death in Africa: Causes, Consequences and Challenges*. Clevedon: Multilingual Matters.
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.
- Bible Hub. <<http://biblehub.com/ecclesiastes/3-9.htm>> (16 July 2017).
- Bokamba, Eyamba G. 2012. A polylectal grammar of Lingála and its theoretical implications. In *Selected Proceedings of the 42nd Annual Conference on African Linguistics: African Languages in Context*, Michael R. Marlo, Nikki B. Adams, Christopher R. Green, Michelle Morrison & Tristian M. Purvis (eds), 291–307. Somerville, MA: Cascadilla Proceedings Project.
- Branford, Jean & Branford, William. 1991. *A Dictionary of South African English*. Cape Town: OUP.
- Buregeya, Alfred. 2007. Aspects of the vocabulary of Kenyan English: An overview. *University of Nairobi Occasional Papers in Language and Linguistics* 3: 1–31.
- Byabazaire, Deogratias M. 1979. *The Contribution of the Christian Churches to the Development of Western Uganda, 1894–1974*. Frankfurt: Peter Lang.
- Collins Dictionary. <<https://www.collinsdictionary.com/dictionary/english/stroke>> (30 July 2017).
- Corpus of Global Web-Based English (GLOWbE). <<http://corpus.byu.edu/glowbe/>> (16 July 2016).
- Crystal, David. 2003. *English as a Global Language*. Cambridge: Cambridge University Press.
- Dako, Kari. 2003. *Ghanaianisms: A Glossary*. Accra: Ghana Universities Press.
- Davies, Mark. 2010. More than a peephole: Using large and diverse online corpora. *International Journal of Computer Linguistics* 15: 405–411.
- Davies, Mark & Fuchs, Robert. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GLOWbE). *English World-Wide* 36(1): 1–28. <https://doi.org/10.1075/eww.36.1.01dav>
- Fisher, Allestree E. C. 2000. Assessing the state of Ugandan English. *English Today* 16: 57–61. <https://doi.org/10.1017/S0266078400011470>
- Haspelmath, Martin. 2009. Lexical borrowing: Concepts and issues. In *Loanwords in the World's Languages: A Comparative Handbook*, Martin Haspelmath, & Uri Tadmor (eds), 35–54. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110218442>
- Huber, Magnus. 2014. Stylistic and sociolinguistic variation in Schneider's nativization phase. T-affrication and relativization in Ghanaian English. In *The Evolution of Englishes. The Dynamic Model and Beyond* [Varieties of English around the World G49], Sarah Buschfeld, Thomas Hoffmann, Magnus Huber, & Alexander Kautzsch (eds), 86–106. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.g49.06hub>
- Hussein, Ashatu. 2009. The use of triangulation in social sciences research: Can qualitative and quantitative methods be combined? *Journal of Comparative Social Work* 1: 1–12.
- Igboanusi, Herbert S. 2010. *A Dictionary of Nigerian English Usage*, 2nd edn. Berlin: LIT.
- Isingoma, Bebwa. 2016. Borrowings and calques in Ugandan English. In Meierkord, Isingoma & Namyalo (eds), 149–172.
- Isingoma, Bebwa & Meierkord, Christiane. 2016. Introduction: Ugandan English – Challenges to, and food for, current theories. In Meierkord, Isingoma & Namyalo (eds), 1–15.

- Kabanankye, Kaban & Kwagala, Dorothy. 2007. Culture, minorities and linguistic rights in Uganda: The case of the Batwa and the IK. *Working Paper 11*. Kampala: Human Rights and Peace Centre (HURIPPEC).
- Karlson, Monica. 2013. Quantitative and qualitative aspects of L1 (Swedish) and L2 (English) idiom comprehension. *Studies in Second Language Learning and Teaching* 3(2): 293–319. <https://doi.org/10.14746/ssllt.2013.3.2.7>
- Kilgarrieff, Adam, Baisa, Vít, Bušta, Jan, Jakubíček, Miloš, Kovvář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel & Suchomel, Vít. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Kouega, Jean-Paul. 2006. *Aspects of Cameroon English Usage: A Lexical Appraisal*. Munich: Lincom.
- Kouega, Jean-Paul. 2007. *A Dictionary of Cameroon English Usage*. Frankfurt: Peter Lang.
- Lewis, Paul M., Simons, Gary F. & Fennig, Charles D. (eds). 2015. *Ethnologue: Languages of the World*, 18th edn. Dallas TX: SIL International.
- Loewen, Shawn & Reinders, Hayo. 2011. *Key Concepts in Second Language Acquisition*. Houndmills: Palgrave Macmillan. <https://doi.org/10.1007/978-0-230-34627-7>
- Mackey, Alison & Gass, Susan M. 2016. *Second Language Research. Methodology and Design*, 2nd edn. New York NY: Routledge.
- Mair, Christian. 2007. Change and variation in present-day English: Integrating the analysis of closed corpora and web-based monitoring. In *Corpus Linguistics and the Web*, Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds), 233–247. Amsterdam: Rodopi. https://doi.org/10.1163/9789401203791_014
- Meierkord, Christiane. 2016a. A social history of English in Uganda. In Meierkord, Isingoma & Namyalo (eds), 51–71.
- Meierkord, Christiane. 2016b. Diphthongs in Ugandan English: Evidence for and against variety status and Interactions across Englishes. In Meierkord, Isingoma & Namyalo (eds), 121–148.
- Meierkord, Christiane, Isingoma, Bebwa & Namyalo, Saudah (eds). 2016. *Ugandan English: Its Sociolinguistics, Structure and Uses in a Globalizing Post-protectorate* [Varieties of English Around the World 59]. Amsterdam: John Benjamins. <https://doi.org/10.1075/veaw.959>
- Melchers, Gunnel & Shaw, Philip. 2013. *World Englishes*, 2nd edn. Abingdon: Routledge. <https://doi.org/10.4324/9780203785416>
- Mesthrie, Rajend. & Bhatt, Rakesh M. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511791321>
- Meyer, Charles. 2002. *English Corpus Linguistics. An Introduction*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511606311>
- Michieka, Martha M. 2009. Expanding Circles within the Outer Circle: The rural Kisii in Kenya. *World Englishes* 28(3): 352–364.
- Monitor. <www.monitor.co.ug> (30 July 2016).
- Nakayiza, Judith. 2016. The sociolinguistic situation of English in Uganda: A case of language attitudes and beliefs. In Meierkord, Isingoma & Namyalo (eds), 75–94.
- Namyalo, Saudah, Bebwa Isingoma & Christiane Meierkord. 2016. Towards assessing the space of English in Uganda's linguistic ecology: Facts and issues. In Meierkord, Isingoma & Namyalo (eds), 19–49.
- Nassenstein, Nico. 2016. A preliminary description of Ugandan English. *World Englishes* 35.3: 396–420. <https://doi.org/10.1111/weng.12205>
- Nelson, Gerald. 1996. The design of the corpus. In *Comparing English Worldwide. The International Corpus of English*, Sydney Greenbaum (ed.), 27–35. Oxford: Clarendon Press.

- New Vision*. <www.newvision.co.ug> (August 2017).
- Observer*. <www.observer.ug> (30 July 2017).
- Oxford Advanced Learner's Dictionary* (OALD), 8th edn. 2010. Oxford: OUP.
- Oxford Dictionary Online*. 2017. <<https://en.oxforddictionaries.com/definition/houseman>> (2 July 2017).
- Okazaki, Sumie, E. J. R. David & Nancy Abelmann. 2008. Colonialism and psychology of culture. *Social & Personality Psychology Compass* 2(1): 90–106.
<https://doi.org/10.1111/j.1751-9004.2007.00046.x>
- Rosenbach, Anette. 2013. Combining elicitation data with corpus data. In *Research Methods in Language Variation and Change*, Manfred Krug & Julia Schlüter (eds), 278–294. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511792519.019>
- Scanlon, David. 1964. Education in Uganda. U.S. Department of Health, Education, and Welfare. Office of Education. Bulletin 1964, No. 32. Washington DC: US Department of Health, Education, and Welfare (now Health and Human Services).
- Schmied, Josef. 2012. Standards of English in East Africa. In *Standards of English. Codified Varieties Around the World*, Raymond Hickey (ed.), 229–255. Cambridge: CUP.
<https://doi.org/10.1017/CBO9781139023832.013>
- Schneider, Edgar. 2007. *Postcolonial English. Varieties of English Around the World*. Cambridge: CUP.
- Silva, Penny et al. 1996. *A Dictionary of South African English on Historical Principles*. Oxford: OUP.
- Simons, Gary F. & Fennig, Charles D. (eds). 2017. *Ethnologue: Languages of the World*, 20th edn. Dallas TX: SIL International. <<http://www.ethnologue.com>> (16 July 2017).
- Skandera, Paul. 2003. *Drawing a Map of Africa. Idiom in Kenyan English*. Tübingen: Narr.
- Ssekamwa, J. C. 1997. *History and Development of Education in Uganda*. Kampala: Fountain Publishers.
- Sempuuma, Jude, Isingoma, Bebwa & Meierkord, Christiane. 2016. The use of the progressive in Ugandan English. In Meierkord, Isingoma & Namyalo (eds), 173–199.
- Ssentanda, Medadi E. 2016. Tensions between English medium and mother tongue education in rural Ugandan primary schools: An ethnographic investigation. In Meierkord, Isingoma & Namyalo (eds), 95–117.
- Suruma, Ezra S. 2014. *Advancing the Ugandan Economy: A Personal Account*. Washington DC: Brookings Institution Press.
- Taguchi, Naoko. 2013. Prompted Responses. In *The Encyclopedia of Applied Linguistics: Quantitative and Mixed Methods*, Carol Chapelle (ed.), Malden MA: Wiley-Blackwell.
- Thomason, Sarah G. 2001. *Language Contact. An Introduction*. Edinburgh: Edinburgh University Press.
- Trudgill, Peter & Hannah, Jean. 2017. *International English. A Guide to Varieties of English Around the World*, 6th edn. London: Routledge. <https://doi.org/10.4324/9781315192932>
- Udo-Udo, Jacob. 2017. *Convincing Rebel Fighters to Disarm*. Berlin: Walter de Gruyter.
- Uganda Bureau of Statistics. 2010. *Uganda National Household Survey Report 2009/10*. Kampala: Uganda Bureau of Statistics. <http://www.ubos.org/UNHS0910/chapter3_literacy.html> (5 July 2017).
- Uganda Bureau of Statistics. 2016. National population and housing census 2014. <<http://www.ubos.org/onlinefiles/uploads/ubos/NPHC/2014%20National%20Census%20Main%20Report.pdf>> (30 June 2017).
- Weisser, Martin. 2016. *Practical Corpus Linguistics. An Introduction to Corpus-Based Language Analysis*. Oxford: Wiley. <https://doi.org/10.1002/9781119180180>
- Wolf, Hans-Georg & Polzenhagen, Frank. 2009. *World Englishes. A Cognitive Sociolinguistic Approach*. Berlin: Walter de Gruyter. <https://doi.org/10.1515/9783110199222>

Appendix 1. Results of acceptability test

	good	ok	neutral	odd	bad	Not clear	No valid reply	ratio
Since she is below 18, she had to ask her parents.	21	11	4	3	1	1		1.76
Ministries are eating big on the new budget.	3	8	11	5	13	1		3.43
The president seems to rule pakalast.	4	11	6	11	8	1		3.2
The room in the hostel had six double-decker beds.	10	18	6	1	5	1		2.33
After having lost his job, Samuel is grassing seriously.	4	8	11	5	10	3		3.24
They offered meat pie and sausage to escort the tea.	8	9	7	9	3	4	1	2.72
I am wondering whether food is already over.	6	8	13	8	6			3.0
You can harvest sukuma wiki all year round.	12	9	8	5	5	2		2.54
In A level, he stroked Divinity with English.	9	7	7	7	8	2	1	2.95
Livingstone Kawuulu is a special hire driver working in Kampala.	22	14	2	0	1	2		1.51
I have been to this place thrice.	22	13	4	2	0	0		1.66
With his bad behaviour, Moses ashamed his whole family.	20	12	2	2	5			2.02
Kabalagala is served in many eateries, with tasty ingredients.	4	5	7	9	8	9	1	3.36
A ceremony will be held in honour of our fallen colleague.	18	8	5	2	7	1		2.3
MTN offices are found in expensive flats in Kampala.	15	9	11	2	1	3		2.08
Oh it is 4pm and I have not greeted you; Good Evening!	12	8	7	5	8		1	2.73
Good enough, they had locked the door, so the thief could not enter.	22	6	7	4	1	1		1.9
The minister insisted on death penalty; now he is being paid in his own currency, since his own brother has been sentenced to death.	10	9	7	5	3	7		2.47
As I was giving my friend a push in the evening, I was bitten by a snake.	17	9	8	1	5	1		2.2
He is a strong boy, so he can mug this box	11	16	4	4	4	3		2.33
Do you enjoy eating cold power?	8	7	5	9	7	5		3.0

Appendix 2. Examples of sentences, depicting UgE usage, constructed in the production test

1. To pay someone in his/her own currency
 - i. You killed my brother, I will pay you in your own currency.
 - ii. Amama Mbabazi was paid in his own currency by the Public Management Bill.
 - iii. The madman was paid in his currency by stoning.
 - iv. A town mayor passed a certain law that is currently being used to pay him, I guess he is being paid in his own currency.
 - v. Because of cheating examination, Aloyo was paid in her own currency.
2. To give someone a push
 - i. Don't mind I will give you a push after lunch.
 - ii. I gave my girlfriend a push up to the bodaboda stage.
 - iii. I gave him a push after lunch.
 - iv. Considering the insecurity in Gulu, I rarely give my visitors a push without carrying my ID.
 - v. I will give you a push if you come to my place.
3. To stroke
 - i. I stroked Literature, Divinity and Entrepreneurship with Information technology.
 - ii. He stroked rice with matooke.
 - iii. I stroked posho with potatoes.
 - iv. In year three, I wish to stroke English language with Mass Communication.
 - v. Students prefer stroking Literature in English with History.
4. To mug
 - i. They will mug a heavy stone tomorrow.
 - ii. John is old enough to mug the box.
 - iii. I can mug that table for you.
 - iv. She mugged the suitcase alone.
 - v. I mugged the elephant boy during our fight at the park.
5. Cold power
 - i. Sarah ate cold power last evening.
 - ii. Street children usually take cold power.
 - iii. I eat cold power in the morning.
 - iv. I used to eat cold power when I was in a boarding school.
 - v. She ate her cold power which she left yesterday.

Appendix 3. Informants' demographic details

S/N	Nickname	Gender	Age	Occupation	L1	Language family	Other languages used daily
1	Junto	M	21	Student	Ateso	Nilotic	English
2	Truman	M	–	Stutent	Lugbara	Central Sudanic	English
3	Aunt	F	22	Student	Acholi	Nilotic	English
4	Ocan	M	23	Student	Acholi	Nilotic	English, Luganda, Kiswahili
5	Ruthian	F	21	Student	Ateso	Nilotic	English
6	Ashley	F	22	Student	Madi	Central Sudanic	English
7	Pastor's Kid	F	–	Student	Lusoga	Bantu	English
8	Inno	F	25	Student	Ateso	Nilotic	English
9	Latabu	F	22	Student	Acholi	Nilotic	English
10	Micke	F	21	Student	Acholi	Nilotic	English
11	Nyaps	F	23	Student	Dhopadhola	Nilotic	English, Luganda
12	Kitamuliko	F	40	Teacher	Rutooro	Bantu	Rutooro
13	Manukanika	F	21	Student	Lango	Nilotic	Luganda
14	Kanyonyi	M	27	Teacher	Runyankole	Bantu	English
15	Patra Carol	F	28	Student	Lango	Nilotic	English
16	Sukuma Wili	F	38	Teacher	Luganda	Bantu	English
17	Chimanzee	F	21	Student	Lugisu (Lumasaaba)	Bantu	Luganda, Lugwere, Lunyole Kiswahili
18	Kash	M	24	Teacher	Rutooro	Bantu	English
19	Jack-Dowkins	M	23	Student	Lango	Nilotic	English
20	Fadiya	F	23	Student	Madi	Central Sudanic	English, Arabic
21	Bakaswara	F	37	Teacher	Rutooro	Bantu	Runyankole
22	Pumpkin	F	30	Student	Ateso	Nilotic	English
23	Ssedunga	M	39	Teacher	Luganda	Bantu	Luganda
24	Flower	F	21	Student	Acholi	Nilotic	English
25	Dani	F	21	Student	Acholi	Nilotic	English
26	Beautiful	F	21	Student	Kupsabiny	Nilotic	English
27	Nickiminaj	F	21	Student	Rukonzo	Bantu	English, Alur, Acholi
28	Mr O. T.	M	22	Student	Lango	Nilotic	English
29	Glane	F	25	Student	Acholi	Nilotic	English
30	Paty	F	21	Student	Acholi	Nilotic	English
31	Jeff Kibs	F	21	Student	Kuman	Nilotic	Lango, Luganda
32	Kabalanja	F	37	Teacher	Rukonzo	Bantu	Lubwisi
33	Theopista	F	26	Student	Lango	Nilotic	Luganda
34	Gift	F	21	Student	Lugisu (Lumasaaba)	Bantu	English, Kiswahili, Luganda

(continued)

S/N	Nickname	Gender	Age	Occupation	L1	Language family	Other languages used daily
35	Fifi	F	20	Student	Madi	Central Sudanic	English, Luganda
36	Hage	M	32	Teacher	Rutooro	Bantu	Rukonzo, Lubwisi
37	Katuma	M	31	Teacher	Runyoro	Bantu	English, Acholi, Rutooro, Luganda
38	Milkah	F	21	Student	Dhopadhola	Nilotic	English, Luganda
39	Kadawino	M	22	Student	Reli	Cetral Sudanic	Kuku, Madi, Kakwa
40	Acako	F	20	Teacher	Rutooro	Bantu	English
41	Rachael	F	20	Student	Acholi	Nilotic	English

PART III

Applications of corpora in English language teaching and learning

A corpus-based analysis of conjunctive cohesion in English essays of Nigerian university learners

Adeyemi Iyabo
North-West University

The study examines the use of conjunctions in written texts by Nigerian university learners. It uses the corpus-based method and instruments to compare the Nigerian Learner English Corpus (NLEC)'s use of conjunctions to their native counterparts in Louvain Corpus of Native English Essays (LOCNESS). The analysis shows additive conjunctions have the highest frequency of use by the two learner groups, while causal, adversative and temporal conjunctions have below average usage. The learners repeatedly used particular conjunctions and underused others within the same category. The study concludes that the advanced learners do not display the optimal awareness of the various alternative conjunctive items available within the different groups to create stylistic variation in their texts for enhanced cohesion and overall coherence.

1. Introduction

Conjunctions are devices used to organise discourse in academic writing. They help to signal logical or semantic relations which are the most visible clues for restricting the interpretation and understanding of relations that exist in texts (Dooley & Levinsohn 2001). Zamel (1983) opines that it would be difficult, without conjunctions, to make sense of ideas, since these conjunctions prepare the readers to anticipate the ideas which follow. Scholars like Cook (1989) and McCarthy (1991) have asserted the importance of conjunctions to discourse and writing pedagogy. In spite of the critical role of conjunctions, their use in academic writing has been found to be challenging for second/foreign language learners of English (Ting 2003; Ong 2011) and even for native speakers (Norrish 1983, cited in Hamed 2014).

At the tertiary level, academic writing requires the ability of students to construct grammatically correct sentences and cohesive texts. When writers write, irrespective of whether it is in the L1 or L2, they bring to the task their knowledge of

the writing process and the strategies they will use in composing the text. Academic writing requires some skills and conventions like organisation in the development of ideas and information, a high degree of accuracy in choosing the right words so that there is no ambiguity of meaning, and also the correct use of grammatical devices to focus, link and emphasise ideas. This multidimensional skill can create problems for students, especially when writing in specific academic contexts (Matsuda et al. 2009: 461).

According to Hyland (2009: vii), communication is crucial to the work of academic communities. Students and researchers must gain fluency in the conventions of academic discourse to understand their disciplines, establish their careers and to successfully navigate their learning. In the case of students, the literacy demands of the academic and the disciplinary-specific communication requirements have to be met. These criteria are attainable by learning to use language in new ways (Hyland 2009: ix). For most second language writers, coming to terms with these literacy demands of academic writing might be more complex than for their L1 counterparts. Conjunctions are crucial discourse structure markers and they function as the link between the various text segments. Hence, the need for the learners to consciously use the appropriate conjunctions to ensure the comprehension and interpretation of their written academic texts.

This study examines the use of conjunctions to enhance cohesion and overall coherence in English essays of Nigerian university learners. The English language holds a significant position as a second language in Nigeria. It is socio-linguistically significant in Nigeria, due to its functionality in fulfilling the crucial role of communication in Nigeria's multifarious linguistic settings (Ogunsanya 2009). It is the language of sectors such as government, communication (within and outside the country), mass media, business and commerce, religion, and so forth. It is also the medium of instruction in the Nigerian education system. Several scholars have examined the introduction of the English language, its survival, and nurture in Nigeria. The summaries of their studies all agreed on the remarkable role the language plays in the complex Nigerian language situation (see Akere 2004; Awonusi 2004; Babatunde 2001; Bamgbose 1971; Jowitt 1991; Ogunsanya 2009).

The English language is operating in an extremely multilingual context; as such, the language serves as a useful tool for effective cross-cultural interaction in Nigeria. Afolayan (1991: 8) indicates the significance of the English language in Nigeria in its functions as 'an instrument of the total system of education, an instrument of individual and overall societal development'. The language has been firmly and intensely incorporated into the national life of Nigeria. It enjoys a pride of place in the Nigerian education system, and the need for learners to achieve a reasonable level of communicative competence cannot be over-emphasised. This includes linguistic, grammatical and pragmatic competence for them to be able to

use expressions to accomplish the desired communicative effect. There is, therefore, the need for the learners to have a conscious mastery of the mechanics of English language usage as much as possible for them to be fluent, accurate and coherent in their use of the language.

In this chapter, a corpus-based analysis of conjunctive cohesion in Nigerian university learners' and their native English counterparts' writing is carried out. The native speaker data was sourced from the Louvain Corpus of Native English Essays (LOCNESS), which is a corpus of essays written by 189 British and American university students. The Nigerian Learner English Corpus (NLEC) was compiled as an ESL corpus. This study addresses the following research questions:

1. How do Nigerian learners use conjunctions to achieve cohesion in written texts?
2. Is there a significant difference in Nigerian learners' choice and usage of conjunctions to achieve cohesion and coherence in writing compared to their native speaker counterparts?

In Section 2, the concept of conjunction is examined. Section 3 discusses studies on the use of conjunctions in ESL/EFL's writing. Section 4 focuses on the methodology and research tools used in the study. The analysis is presented in Section 5, followed by the discussion of findings in Section 6. Lastly, in Section 7, the conclusion together with the implications of the results for language teaching are presented.

2. Conjunctions

Conjunctions are used to describe the cohesive ties between clauses or sections of text and demonstrate a meaningful relationship between them. They function as connectors between two independent clauses or sentences (or paragraphs) in the text. Halliday and Hasan (1976: 227) explained that cohesive elements are cohesive not in themselves but indirectly, by virtue of their specific meanings. They are not primary devices for reaching out into the preceding or succeeding the part of a text, but they express particular meanings which presuppose the presence of other components in the discourse. They conjoin linguistic units such as sentences, paragraphs and even more substantial sections of the text (Quirk et al. 1985). They signal logical relations in texts and help the reader to connect different units and paragraphs to make sense of the text (Heino 2010); hence, they are essential cohesive devices in discourse. Conjunctions can be one-word items, for example *and* or fixed-word combinations, e.g., *as a result*. They contribute to the clarity and comprehensibility of a text.

Conjunctions assist readers in their interpretation of the pragmatic connections in a text in relation to the writer's thinking as well as express the semantic relationships between clauses (Hyland 1998). The use of conjunctions has been shown to help maintain text cohesion and coherence in academic discourse; they contribute to the creation of semantic unity. The importance of conjunctions for the understanding of inter-sentence relations and the construction of cohesive devices for text coherence has been highlighted by scholars such as Rahimi and Qannadzadeh (2010), Tseng & Liou (2006), and Yang & Sun (2012), among others. They operate as logical connections between clauses, sentences, and paragraphs in discourse, adding them together, comparing them, sequencing them in time, or explaining their causes, purpose or conditions.

Various names have been used to refer to conjunctions based on the role they perform in discourse, for example, connectors, logical connectives, discourse connectors, linking adverbials, etc. It is, however, important to note that irrespective of the names conjunctions are called, they all perform the same function of enhancing semantic relations in texts. Conjunction has been discussed and classified by the following scholars, among others: Carston (1993), Celcee-Murcia & Larsen-Freeman (1999), Good (2002), Halliday & Hasan (1976) Hoey, (1991), Martin & Rose (2003), Quirk et al. (1985), Vande Kopple (1985). Since the purpose of this study is to analyse conjunctive cohesion in L2 English writing with a focus on the implications for language teaching, Halliday & Hasan's (1976) taxonomy of conjunction is adopted. Halliday and Hasan (1976: 239) used four categories to distinguish between the several types of conjunctive elements, namely additive, adversative, causal and temporal conjunctions. A brief description of the four categories is given below as a basis for the data analysis.

- Additive conjunctions connect phrases, clauses, and sentences in writing. They are used to further the discourse topic. Additive conjunctions are commonly used to connect phrases, clauses, and sentences in writing. They are used as expressions to introduce an item as either a specific or a more general claim. They mark transitions from a more general point to a more specific supporting point or argument or mark transitions to a particular supporting example. Additive conjunctions are also used to link phrases, clauses in sentences as well as join compound and complex sentences. Examples of additive conjunctions include: *and, or, as well as, also, besides, by the way, for example, for instance, furthermore, in addition*, etc. (Halliday & Hassan 1976: 244).
- Causal conjunctions mark the relationships of reason, consequence, and purposes. Causal conjunctions are used in discourse to indicate a relation of cause and effect. They are used to explain the relationship between events or things, where one is the result of the other or others. They indicate the 'what' and 'why'

something happened in discourse. Examples of causal conjunctions include: *as a result, because, consequently, hence, in order to, in that case, otherwise, so, so as, unless*, etc. (Halliday & Hassan 1976: 243).

- Adversative conjunctions operate to introduce an item of information which is ‘contrary to expectation’ in discourse. Adversative conjunctions moderate or qualify the information in the first half of the sentence, paragraph or text. Adversative conjunctions perform various sub-functions which all play significant roles in the interpretation of discourse. They operate as elaborative devices to provide focus on the content of discourse. They also act as transitional markers in discourse. Examples of adversative conjunctions include: *although, but, however, in any case, instead, nevertheless, on the contrary, still, though, yet*, etc. (Halliday & Hassan 1976: 242).
- Temporal conjunctions specify the time sequence relationship which exists between propositions. Temporal conjunctions are used to express relations in time. They are employed when events in texts are linked with regards to the time of occurrence. They express happenings in connection with the simultaneity, anteriority or posteriority. These connectives answer the questions ‘when’ regarding event occurrence in discourse. Examples of temporal conjunctions include: *after, before, finally, in conclusion, in short, meanwhile, next, previously, soon, subsequently*, etc. (Halliday & Hassan 1976: 243).

The logical relationships between ideas are indicated by the use of appropriate conjunctions so that the reader can understand the relationship between the parts of a text. The four categories mentioned above reflect the semantic relations between sentences in texts. It is thus critical for ESL learners to understand the role of conjunctions in composing academic texts.

3. Studies on the use of conjunctions in ESL/EFL learners’ academic writing

It is essential that language learners use a variety of conjunctive devices to link the previous and following sentence(s) to make their writing explicit, orderly, and logical. Their usage also contributes to making their writing semantically, pragmatically and grammatically well-formed. A significant number of studies about conjunctions in ESL/EFL learners’ academic writing has been carried out. The results from these studies reveal the difficulties faced by learners in using conjunctions correctly in their writing. Some of these studies have indicated the overuse, underuse, and misuse of conjunctions by ESL and EFL learners.

The overuse of additive conjunctions in learners writing has been reported in the following studies: Zhang (2000) by Chinese learners, Ishikawa (2009) by Asian learners, Shen (2007) by Chinese and Japanese learners, Fakhra (2009) by Syrian learners, and Martinez (2015) by Spanish learners. The overuse of causal conjunctions has also been indicated by Fakhra (2009) in Syrian learners' English essays, Martinez (2015) by Spanish learners, Mo (2015) by Chinese learners, and Shen (2007) by Chinese and Japanese learners. Fakhra (2009) indicated the overuse of *but* and *however* as adversative conjunction in Syrian learners' writing.

The underuse of adversative conjunctions has been found by Chen (2006), Martinez (2015), and Muddhi & Hussein (2014) in Taiwanese, Spanish and Kuwaiti learners' writing respectively. However, Muddhi and Hussein (2014) observed that despite the general underuse of adversatives, *but* is significantly overused by the Kuwaiti learners in their study. Abusharkh (2012) and Mo (2015) also reported the underuse of causal conjunctions in Palestinian and Chinese learners' writing.

The misuse or erroneous use of some conjunction types has also been established by scholars. Crewe (1990) and Hammed (2014) indicated the misuse of adversative conjunctions by Singaporean and Libyan learners, while Fakhra (2009) and Yoon and Yoo (2011) identified the misuse of the causal conjunctions by Syrian and Korean learners. These studies and others have also established that learners sometimes have problems with the application of particular conjunctions in academic writing to achieve cohesion. The inappropriate use of these conjunctions often has negative impact on the overall quality of the learners' written texts.

In the Nigerian context, various studies have been carried out on the different forms of cohesion in student writing in general (Ojetunde & Okanlawan 2011; Malah 2015; Malgwi 2016; Mohammed 2015; Olateju 2006; Shittu 2015; Usman 2015). However, only a few of these studies have focused on investigating conjunctions in students' writing. Focusing on the investigation of conjunctions as a cohesive device in student writing, Ojetunde and Okanlawan (2011) examined the use of English conjuncts by General English Studies students in selected tertiary institutions in south-western Nigeria. They observed a high frequency and redundant use of conjunctions like *though*, *so*, *but*, *nevertheless*, and *so therefore*. They further indicate the misuse of *however* and *nevertheless*. The random use of additives *also*, *and*, *as well as* and *moreover* were also reported. They concluded that the students had poor mastery of conjuncts even at the tertiary level of education.

Mohammed (2015) analysed conjunctions as cohesive devices in General English Studies students' texts from Bauchi State University. She found the students used more additive conjunctions compared to the other conjunction types in their writing. She noted *and* is the most common conjunction in the students' texts, and further observed that in all the texts analysed in her study, there were only eleven cohesive conjunctions found in the students' texts. The conclusion drawn from

her findings is that the students are yet to master the mechanics of text connection through the use of conjunctions.

Based on the conclusions of some of these recent studies, this study intends to carry out quantitative and qualitative analysis on individual conjunctive items in the four categories to identify the pattern of usage, as well as particular conjunctions the learners have problems using in their texts. It will also compare the Nigerian university's texts to L1 counterparts to determine significant similarities or differences in Nigerian learners' choice and usage of conjunctions to achieve cohesion and coherence in writing compared to the L1 learners' texts. The comparative analysis will enhance a better interpretation of the results to determine whether the documented problems with the use of conjunction in the Nigerian advanced learners' writing is unique to them or not.

The study used the corpus research instruments WordSmith Tools 5.0 developed by Scott (2008) to search and extract the various conjunctive items from the corpus. This helped to account for the number of frequency in the corpus. The concordance was used to observe the pattern of usage and to disambiguate items that were ambiguous in use (see Section 4.3). The Paul Rayson's log-likelihood calculator was used as a pairwise comparison statistical instrument to determine the significance of the similarity and difference between two individual frequency scores. It also helped identify potentially significant differences that required closer analysis. The use of the quantitative and qualitative analysis helped to determine which particular conjunctions the learners struggle with and why.

The study also intends to fill the data-size gap, due to the observation that the data sources of previous studies on learners' writing are randomly sampled essays, i.e., 75 students in Ojetunde & Okanlawan (2011) and 20 students in Mohammed (2015). This study provides increased data analysis for the description of Nigerian tertiary learners' writing. This study investigates the use of conjunctive cohesion in the writing of undergraduates in Nigeria using an increased data size (467 essays) and more detailed and verifiable analysis based on the corpus-based method. The study's use of corpus linguistics as the method will provide an improved description of these advanced learners' language use.

The compilation of ICE-Nigeria component of the International Corpus of English (ICE) began in 2007. However, the application of corpus linguistic theory and methods to Nigerian English usage started to appear in publications around the year 2012. Some of the corpus-based studies include Adagbonyin et al. (2016), Adelakun (2013), Daramola (2012), Esimaje (2012, 2013), Fuchs & Gut (2015), Fuchs et al. (2013), Gut (2012), Gut & Fuchs (2013), Ibrahim (2015), Jauro et al. (2014), Okpongette & Okon (2016), and others. Available literature indicates no evidence of corpus-based analysis of cohesion of student writing in Nigeria. Several of the studies have investigated Nigerian English usage using diverse corpora. However,

none of the studies have focused on Nigerian students' writing. Based on the ample benefits of the research instruments and method of analysis, it is hoped that this study will offer a more accurate analysis of Nigerian students' use of conjunctions than those currently available, further bridge the observed gaps and contribute to a better understanding of the nature of Nigerian university students' writing.

4. Methodology

4.1 Corpus linguistics approach

This study adopted a corpus linguistics approach. The corpus-based instruments and method used in the study are of paramount importance as they directly and significantly affect the quality of the findings of this study and consequently the insights into the Nigerian learners' use of conjunctions in comparison to their L1 counterparts. The study's data was gathered from three Nigerian universities in different parts of the country. The use of improved data volume due to corpus linguistics method becomes imperative as it allows for the processing of a large amount of data, is more objective and can easily be verified and shared by other researchers and language teachers.

4.2 Corpora

The data for this study was sourced from the Nigerian Learner English Corpus (NLEC) and the Louvain Corpus of Native English Essays (LOCNESS). The NLEC was compiled in 2012 using the ICLE (The International Corpus of Learner English) guidelines for a study in Nigerian tertiary learners' written English essays. At the time of NLEC's compilation, the Corpus of Nigerian English, ICE-Nigeria (which began in 2007) by the Centre for English Corpus Linguistics (CECL), was still in process and was not available for public use (The ICE-Nigeria is now available for public use). For the compilation of NLEC, first years students of General English Studies (an introduction to the use of English for academic purposes module) from three Nigerian universities from the east, north and western part of Nigeria participated in the essay writing process. The students were asked to choose a topic from six argumentative and expository essay topics provided and write about two to three pages on the selected topic. Overall, 467 essays were collected with the total number of words amounting to 188,008 words. All the essays were handwritten and time-limited. The essays were then typed as word documents and numbered accordingly.

The Louvain Corpus of Native English Essays (LOCNESS) is the native speaker reference corpus of the International Corpus of Learner English (ICLE). It is widely used as a standard to compare with the non-native data in the learner corpus research community. The section of LOCNESS used in this study contains 189 British and American university students' essays of 264,095 words. Due to the disparity in the size of the two corpora compared in the study, the Star Trek online Random Number Generator was used to remove 20 essays from LOCNESS corpus. Thus, 169 essays from the LOCNESS corpus with the total of 196,102 words were used for the comparative study.

4.3 Corpus search

Corpus searches were carried out using WordSmith Tools 5.0. The wordlist feature was used to extract 5,441 conjunctive cohesive device tokens from LOCNESS and 5,864 from NLEC. The frequency numbers were normalised to 100,000 words to ensure a notionally common scale for both data groups and to eliminate any similarity or disparity due to data size, thereby enhancing the result accuracy. The concordance tool was used to disambiguate words that were ambiguous in use. The corpora were searched for the following conjunctions:

- Additive conjunctions: *also, and, as well as, besides, by the way, for example, for instance, furthermore, in addition, in other words, in the same way, likewise, moreover, not only, or and thus* (Halliday & Hassan 1976: 242).
- Causal conjunctions: *as a result, because, consequently, hence, in order to, in that case, in this respect, otherwise, since, so, so as, to this end, unless, and therefore* (Halliday & Hassan 1976: 243). The concordance was used to disambiguate *so* as an adverb of degree and *so* as causal conjunction.
- Adversative conjunctions: *although, at least, but, contrary to this, however, in any case, in fact, instead, nevertheless, on the contrary, on the other hand, still, though, whereas and yet* (Halliday & Hassan 1976: 242).
- Temporal conjunctions: *and then, at last, at the same time, at this moment, at this point, finally, in conclusion, in short, meanwhile, next, previously, soon, subsequently, to conclude, until then, after, before and while* (Halliday & Hassan 1976: 243). The concordance was used to disambiguate the use of *next* as adjective in noun phrases and *next* as temporal conjunction in the corpora.

4.4 Statistical analysis

A log-likelihood test was carried out to examine the relationship or association in the use of each conjunctive cohesive item between two categorical variables of LOCNESS and NLEC in the study. The Web-based log-likelihood wizard is available at <http://ucrel.lancs.ac.uk/llwizard.html>. The calculator for corpus frequencies was developed by Paul Rayson (UCREL University Centre for Computer Corpus Research on Language, Lancaster University, UK). The calculator was used for pairwise statistical comparisons. It was instrumental in the interpretation of results and the selection of items discussed. The calculator was used to determine how significant the difference between two individual frequency scores is and to identify potentially significant differences that require closer analysis (Rayson & Garside 2000). The log-likelihood (LL) wizard shows a plus or minus symbol before the log-likelihood value to indicate overuse or underuse respectively in corpus 1 relative to corpus 2. The higher the log-likelihood value, the more significant the difference is between two frequency scores.

Rayson provides a scale on the bases of which significance of log-likelihood values can be interpreted:

- 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
- 99th percentile; 1% level; $p < 0.01$; critical value = 6.63
- 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
- 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13

The chosen scale for the interpretation of the significance of the log-likelihood value for the analysis was set at the 99.9th percentile; 0.1% level; $p < 0.001$ because multiple comparisons were undertaken. The single table with the highest number of simultaneous comparisons contains 19 items. Thus, a higher statistical threshold had to be set. Hence the critical value for interpretation is 10.83. The frequency results and log-likelihood values are presented in tables providing the frequency of occurrence at each level of the variable. The Log-likelihood test was performed on each item in the table. An asterisk * is used to indicate where the value meets the criterion level for statistical significance. The test served as a guide to determine overused and underused conjunctive items.

5. Analysis and results

5.1 Overall results

The frequencies of the four types of conjunctions in both corpora are presented in Figure 1.

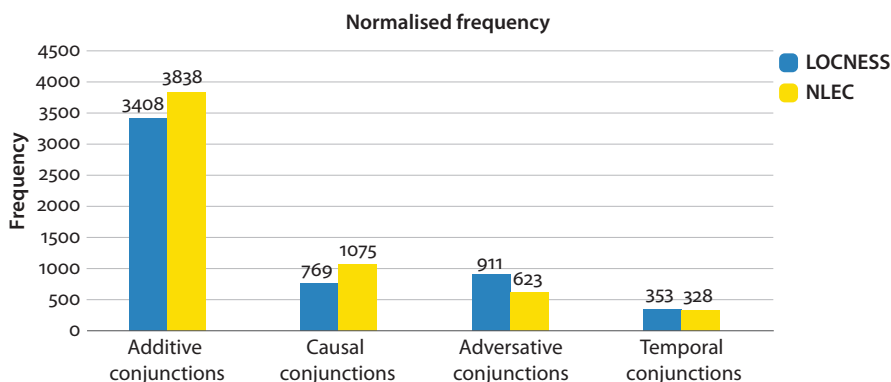


Figure 1. Frequencies of four types of conjunctions occurring in NLEC and LOCNESS (per 100,000 words)

Additive conjunctions are used most frequently in both corpora, followed by causal, adversative and temporal conjunctions respectively. The total frequency per 100,000 words for additive conjunctions is 3,838 in NLEC compared to 3,408 in LOCNESS (LL value of +25.43*) indicating significant overuse in NLEC, while causal conjunctions totalled 1,075 in NLEC and 769 in LOCNESS (LL value of +51.01*) indicating significant overuse in NLEC. Adversative conjunctions added up to 623 in NLEC compared to 911 in LOCNESS (LL value of -54.39*) indicating significant underuse in NLEC. The least frequent were temporal conjunctions at a total of 328 in NLEC and 353 in LOCNESS (LL value of -0.92) indicating a non-significant difference in both corpora.

5.2 Subcategories of conjunctive cohesion results

5.2.1 Additive conjunctions

The analysis indicates a significant overuse of additive conjunctions by the NLEC group with the LL value of +25.43*. Table 1 shows that the following additive conjunctions are overused in NLEC compared to LOCNESS: *or*, *furthermore*, *also*, and *for instance* and the underuse of *thus* by NLEC at -28.26*.

Table 1. Frequency of additive conjunctions in NLEC and LOCNESS

Additive conjunctions	LOCNESS normalised frequency/percentage	NLEC normalised frequency/percentage	Log-likelihood value
Also	233 (6.84)	382 (9.95)	+36.46*
And	2,650 (77.76)	2,503 (65.22)	-4.19
as well as	16 (0.47)	15 (0.39)	-0.03
Besides	6 (0.18)	3 (0.08)	-1.02
by the way	1 (0.03)	0 (0.00)	-1.39
for example	36 (1.06)	52 (1.35)	+2.93
for instance	5 (0.15)	43 (0.99)	+34.46*
Furthermore	5 (0.15)	38 (0.99)	+28.70*
in addition	5 (0.15)	18 (0.47)	+7.80
in other words	5 (0.15)	5 (0.13)	0.00*
in the same way	3 (0.09)	1 (0.03)	-1.05
Likewise	5 (0.15)	6 (0.16)	+0.09
Moreover	2 (0.06)	2 (0.05)	0.00
not only	35 (1.03)	14 (0.36)	-9.30
Or	354 (10.39)	747 (19.46)	+143.42*
Thus	47 (1.38)	9 (0.23)	-28.26*
Total	3,408	3,838	

Example (1) shows a Nigerian learners' attempt at presenting a cohesive essay using additive conjunctions:

- (1) Technology advancement has also caused sight problems for example, in the past, you seldom see people wearing spectacles whereas now, almost every teenager is wearing spectacles. The reason is because many teenagers now use the computers for so long to the extent that they spoil their own eyes. Also, it also serves as material for distraction for example blackberry, ipod, are used in the lecture room when lectures are going on, which can distract the lecture himself and also other students sitting beside them. Furthermore, recently in the United States the former health and human services secretary Tommy Thompson announced that he will be getting an implanted microchip under his skin, once it is under his skin he has no freedom again because with this microchip he can be easily tracked which can also damage the skin. Indeed, technology advancement is a curse in Nigeria rather than blessing because it is through technology advancement that we have crude oil which has caused economic breakdown. For instance, the government have abandon agriculture products and other mineral resources for crude oil which is unsustainable for the country.

<NLEC - 011>

Example (1) shows a general high-density of additive conjunctions. Moreover, it illustrates how Nigerian students tend to adopt overt connectives for coordination and subordination, rather than using other cohesive devices. These connections are made explicit by the use of different additive conjunctions. The overuse may also be an attempt by learners to impose surface logicity in their writing.

5.2.2 Causal conjunctions

The analysis indicates a significant overuse of causal conjunctions by the NLEC group with the LL value of + 51.01*. Table 2 shows that the following causal conjunctions are overused in NLEC: *as a result*, *because*, *so* and *so as*. However, the concordance display indicates the correct use of the causal conjunctions as a cohesive device in the learners' texts. *Because* and *so* are the two most frequently used causal conjunctions in the two corpora.

Table 2. Frequency of causal conjunctions in NLEC and LOCNESS

Causal conjunctions	LOCNESS normalised frequency/percentage	NLEC normalised frequency/percentage	Log-likelihood value
As a result	12 (1.56)	46 (4.28)	+21.27*
Because	307 (39.92)	441 (41.02)	+24.14*
Consequently	0 (0.00)	5 (0.47)	+6.93
Hence	8 (1.04)	12 (1.12)	0.81
In order to	43 (5.59)	54 (5.02)	+1.25
In that case	1 (0.13)	2 (0.19)	+0.34
In this respect	1 (0.13)	0 (0.00)	-1.39
Otherwise	9 (1.17)	3 (0.28)	-3.14
Since	55 (7.15)	41 (3.81)	-2.05
So	221 (28.74)	371 (34.51)	+38.42*
So as	6 (0.78)	31 (2.88)	+18.49*
To this end	1 (0.13)	1 (0.09)	0.00
Unless	6 (0.78)	1 (0.09)	-3.96
Therefore	99 (12.87)	67 (6.23)	-6.21
Total	769	1,075	

5.2.3 Adversative conjunctions

The analysis indicates the significant underuse of adversative conjunctions by the NLEC group with an LL value of -54.39*. Table 3 further shows the general underuse of the following adversative conjunctions in NLEC: *although*, *however*, *still*, *though*, and *yet*. The use of the adversatives signifies semantic relations necessary for the complete understanding and interpretation of meanings in the text parts.

Table 3. Frequency of adversative conjunctions in the NLEC and LOCNESS corpora

Adversative conjunctions	LOCNESS normalised frequency/percentage	NLEC normalised frequency/percentage	Log-likelihood value
Although	68 (7.46)	9 (1.44)	−51.20*
at least	1 (0.11)	7 (1.12)	+5.06
But	396 (43.47)	378 (60.58)	−0.42
However	168 (18.44)	58 (9.29)	−55.88*
in any case	1 (0.11)	2 (0.32)	+0.34
in fact	1 (0.11)	11 (1.76)	+9.75
Instead	42 (4.61)	45 (7.21)	+0.10
Nevertheless	3 (0.33)	0 (0.00)	−4.16
On the contrary	2 (0.22)	0 (0.00)	−2.77
On the other hand	20 (2.20)	12 (1.92)	−2.02
Still	88 (9.66)	51 (8.17)	−10.97*
Though	51 (5.60)	27 (4.33)	−10.91*
Whereas	16 (1.76)	11 (1.76)	−0.93
Yet	54 (5.93)	12 (1.92)	−28.91*
Total	911	623	

Though the quantitative analysis indicates the use of adversative conjunctions is low in both corpora, it is evident that the use of adversative is even more moderate in NLEC than in LOCNESS. Example (2) shows where adversative conjunction could have been used by the writer to introduce the contradictory part of the sentence:

- (2) While subsidy on the other hand, connotes ability to subsidize the price of commodity i.e., subsidizing any commodity purchased by a consumer. However, I want to disclose that the removal of fuel subsidy will not help Nigerian economy but will hinder the progress of the nation due to the following reasons: As we all aware that Nigeria is a nation endowed with tremendous mineral and natural resources, such as cocoa, rubber, timber, crude oil among others, the country is being faced with poverty and corruption that has remained a social malady to Nigerian leaders. This menace of corruption has enable Nigerians to have no trust on the Nigerian leaders. An appraisal of the jumbo salaries being earned by the political office holders in the country, neglecting Nigerians to suffer... <NLEC – 047>

The underlined space in line 6 in (2) is an instance where adversative conjunction such as *yet* or *nevertheless* could have been used to indicate or establish the conflicting component of the sentence. This is because the two claims (a) *Nigeria is a nation endowed with tremendous mineral and natural resources*, (b) *the country is being faced with poverty and corruption that has remained a social malady to Nigerian leaders*, are different components of the same sentence displaying adversative relationship. The non-usage of adversative conjunction gives a confusing notion of the

second claim (b) as an additive proposition to the first claim (a). Hence the need for an adversative conjunction which is crucial for creating the semantic relation needed for the understanding and interpretation of the text as a whole.

5.2.4 Temporal conjunctions

The overall result indicates a non-significant difference in the use of temporal conjunctions by the groups with the LL value of -0.92 . However, Table 4 shows the overuse of *before* in NLEC.

Table 4. Frequency of temporal conjunctions in the two corpora

Temporal conjunctions	LOCNESS normalised frequency/percentage	NLEC normalised frequency/percentage	Log-likelihood value
and then	19 (5.38)	5 (1.52)	-8.71
at last	1 (0.28)	1 (0.30)	0.00
at once	1 (0.28)	1 (0.30)	0.00
at the same time	7 (1.98)	3 (0.91)	-1.65
at this moment	1 (0.28)	0 (0.00)	-1.39
at this point	5 (1.42)	2 (0.61)	-1.33
finally	23 (6.52)	26 (7.93)	+0.18
in conclusion	3 (0.85)	0 (0.00)	-4.16
in short	1 (0.28)	2 (0.61)	+0.34
meanwhile	1 (0.28)	5 (1.52)	+2.91
next	22 (6.23)	15 (4.57)	-1.33
previously	7 (1.98)	1 (0.30)	-5.06
soon	14 (3.97)	5 (1.52)	-5.06
subsequently	4 (1.13)	0 (0.00)	-4.44
to conclude	3 (0.85)	1 (0.30)	-1.05
until then	1 (0.28)	0 (0.00)	-1.39
after	110 (31.16)	85 (25.91)	-3.21
before	62 (17.56)	129 (39.33)	+24.01*
while	68 (19.26)	47 (14.33)	-3.86
Total	353	328	

The concordance search for the use of *before* in NLEC points to the repetitive use of the conjunction in some learners' essays. Example (3) is an instance of the repeated use of *before* resulting in its overuse in NLEC.

- (3) Examinations are not true measures of students' ability and should be abolished yes! This is true because nowadays examination is not like before. Before I proceed will like to define examination... To see a situation where by student's have already see the exam before they are tested on it... If this bribery and corruption can stop it may even help in the examination malpractice furthermore, examination are not true measures of students ability through

the bribery and corruption, the society is worth to the extent that students will have seen the questions even months before the examination can we now say that examination is a true measure? No. However the only main point that the question usually leak before it is been written. Some students do not write it. Stay in class just because they already know that after all the question will be seen before the examination. Before I drop my pen I will like to say if there are any means of testing student ability rather than examination it could have been better.

<NLEC – 010>

The temporal conjunction *before* is used to connect an action or an event to a point in time. It signals events happening prior in discourse. The overuse of *before*, as observed from Example (3), is a clear example of overuse resulting in the superfluity of the conjunction. The overuse affects the overall quality of the text.

6. Discussion of results

This study showed that Nigerian learners are able to use conjunctions to indicate the logical relationship between larger text chunks, which is vital for achieving cohesion in their written texts. The analysis revealed significant differences in Nigerian learners' choice and usage of conjunctions to achieve cohesion and coherence in writing compared to their native speaker counterparts. The category analysis indicated the significant overuse of additive and causal conjunctions and the underuse of adversative conjunctions by NLEC writers. Further pairwise statistical comparisons showed specific overuse and underuse of particular conjunctions by the advanced learners in the study.

The overuse of additive conjunctions (see Table 1) might be attributed to the learners' over-familiarity with certain additive conjunctions. It may also be an attempt by learners to impose surface logicity in their writing. This could be an indication of NLEC learners' explicitness in writing, as they attempt to express their points clearly, thereby leaving nothing implied. The explicitness is a positive attribute of the NLEC writing. However, the concentrated use of *also* and the sparing use of other additives such as; *in addition*, *as well as*, *likewise* and *not only* (see Table 1), which have similar additive meaning by the two groups of learners could be an indication that these advanced learners are writing from a limited linguistic repertoire. Thus, they overuse these habitual additives. The repetitive use of familiar lexical items has been labelled lexical teddy bears by Hasselgren (1994).

The analysis of causal connectors reveals that the density of casual conjunctions used by Nigerian learners is higher (see Table 2). However, their variety is more limited than that of their English native speaker counterparts. The analysis' results indicate a challenge in the use of causal conjunctions by the two groups of learners

in the study. These causal relationships, while necessary for illustrating reasons for why an argument is relevant and why it is supported, provide a better account of the logical reasoning used to support an argument. The study showed that the two causal conjunctions *because* and *so* are used more frequently by learners in the two corpora compared to other conjunctions in the group. These findings seem to point at an over-familiarity on the part of the learners with the two causal conjunctive devices resulting in the little or non-use of others in the same category. However, the insufficient lexical diversity of the causal conjunctions in the two groups implies that the advanced learners' use of causal connectives are limited and there is the need for improvement. The ability to do this can significantly improve the overall coherence of their texts.

As far as adversative conjunctions are concerned, the study also shows general underuse by NLEC learners (see Table 3). The use of adversative conjunctions in argumentative writing is essential for semantic and logical consequence. Consequently, the underuse of adversatives observed in NLEC should be of great concern to language teachers. However, it is necessary to state that the few learners who used the adversatives in the corpus used them correctly. The underuse may also be due to the learners' lower familiarity with the usage of these words as contrastive connectors. Hence the need for improved exposure to all the possible alternative conjunctions in this category. Table 3 further indicates the very dominant role of *but* as an adversative linker in the texts. This shows the learners' choice of the syntactically simple option of coordination *but* accounts for more than 60% of all markers of the adversative relations in the NLEC learners' texts.

The analysis indicates no significant difference in the use of temporal conjunctions in the two groups (see Table 4). Temporal conjunctions are the least used sub-category compared to the other three types in the two corpora. The results show a very low frequency of use in both corpora at a total of 353 in LOCNESS and 328 in NLEC. The corpora were searched for Nineteen temporal conjunctions, only three, i.e., *after*, *before* and *while* showed moderate frequency of use. Other temporal conjunctions such as *at last*, *at once*, *at this moment*, *at the same time*, *at this point*, *in short*, *meanwhile*, *previously*, etc. all have below ten frequency of use in the two corpora. *At this moment*, *in conclusion*, *subsequently* and *until then*, were not used at all by Nigerian learners.

The results of the study indicate both groups of learners make infrequent use of temporal conjunctions as cohesive devices in their texts. The knowledge of how and when to use temporal conjunctions in writing can further enhance the coherence of the learners' texts thereby improving the quality. The observed general underuse or non-use in some instance of temporal conjunctions in second language learners' writing, evident in this study and other recent studies like Martinez (2015) and Mohammed (2015), can be seen as a pointer to the advanced learners' underdevelopment in the

use of temporal conjunctions. The underuse or non-use of these conjunctions might be an indication that the learners are yet to master the mechanics of text connection regarding event occurrence through temporal conjunctions.

A general observation from all the four tables is the learners' reliance on the common or popular conjunctive cohesive devices in their writing. The concordance display also indicates the learner's preference for simple coordinating and subordinating conjunctions to link phrases and clauses within the sentence. This is evident in Table 1 with the dominance of *and*, *also* and *or* as additive conjunctions, *because* and *so* in Table 2 as causal conjunctions, *but* and *however* in Table 3 as adversative conjunctions and *before* and *after* in Table 4 as temporal conjunctions. The reliance on these common options may account for the overuse of the said conjunctions in these advanced learners' writings.

The observed low or non-employment of all the particular conjunctive devices in writing by the learners might be linked to Granger & Tyson's (1996) statement that even at a reasonably advanced level, connectors are difficult to master in L2 acquisition. The underuse and limited range of conjunctive items in the Nigerian learners' corpus compared to their L1 counterparts imply that the learners' acquisition and manipulation of connectors are somewhat limited. This reveals that for these advanced learners' the use of conjunctions remains at an elementary stage, which inevitably impacts the overall quality of their texts. These university learners, despite being at an advanced level of learning, still exhibit some difficulty in using the full range of possible conjunctive ties effectively in English writing. The difficulty may be due to the learners' possession of a restricted repertoire of linking words and knowledge of the relationships they signal.

Finally, the comparative study indicates that the challenge of using particular conjunctions and not using other alternatives available within the four categories in writing is not unique to the L2 advanced learners in this study. The study points to significant similarities and differences in the choice and use of conjunctive items in NLEC and LOCNESS respectively. Hence, the need for concerted efforts to expose the advanced learners to learning exercises that can further enhance their use of the full range of the different variety of conjunctive devices effectively in writing to avoid lexical monotony, enhance linguistic finesse and improve the cohesion and coherence of their texts.

7. Conclusion

Based on the findings of this study, and their implications for language teaching, the following recommendations and conclusions are presented. Teachers need to work with learners to improve their use of the various alternative conjunctions in each

category for enhanced linguistic and stylistic variation. Hubbard (1993) indicated that conjunctive cohesion regularly links the meaning of textual units to create a holistic text. This accounts for its importance to the overall coherence of texts. The overuse of particular additive and causal conjunctions, and the underuse of others within the same category points to the advanced learners' essays not showing optimal awareness of the diverse available structuring possibilities in the use of the said conjunctions to create stylistic variation in their written texts.

The underuse of temporal conjunctions observed in this study is an indication of learners' insufficient knowledge in this area of cohesion. Such insufficient knowledge, as the analysis reveals, is likely to be problematic to readers, and negatively affect textual coherence as seen earlier in Example (2), the learners' failure to make adversative propositions, which may confuse a reader who initially reads them as an additive proposition. It was also observed that the learners need to consciously improve their use of cause-effect text connectors to enhance the quality of their writing. With regards to adversative and temporal conjunctions, there is the need for rigorous teaching in as learners' development in these areas seems elementary and their usage shows a lacks variety.

There is the need for the continuous and improved teaching of the various types of conjunctions with the particular focus on the less frequently used ones. Language teachers need to raise learners' awareness of the stylistic impact of different conjunction types and the effect it has on text coherence. As observed in the study, the conjunctive items adopted by the NLEC writers are sometimes repetitive and humdrum. Although the NLEC essays display a fair level of cohesion and coherence, the better and improved use of the various conjunctive cohesive devices and not the monotonous use of particular ones will ultimately enhance the overall coherence of the Nigerian advanced learners' essays.

Acknowledgements

1. This chapter reports on a part of the results of a doctoral study entitled "Corpus-based analysis of cohesion in written English essays of Nigerian tertiary learners" that was completed at the North-West University, South Africa, under the supervision of Prof. Bertus van Rooy. I also want to acknowledge the financial support of the research focus area Understanding and Processing Language in Complex Settings (UPSET) towards the development of the Nigerian Learner English Corpus.
2. I am grateful to the anonymous reviewers for their comments and recommendations; also to Prof. Bertus van Rooy whose insight, expertise and helpful comments have significantly improved the final presentation of this chapter.

References

- Abusharkh, Bassem. 2012. Cohesion and Coherence in the Essay Writing of Palestinian College Students. MA thesis, Hebron University.
- Adagbonyin, Asomwan S., Aluya, Isaiah & Samuel Edem. 2016. A corpus-based approach to the linguistic features in Nigerian and American presidential speeches. *Journal of Linguistics, Language and Culture* 3(1): 1–27.
- Adelakun, Tahir O. 2013. A Corpus-based Critical Discourse Analysis of Election News in Nigerian Newspapers. MPhil. Dissertation, University of Birmingham.
- Adeyanju, Dele. 2009. The English language in the outer circle. In *The Sociolinguistics of English and Nigerian Languages*, Adeyanju Dele (ed.), 1–13. Munich: Lincom.
- Afolayan, Adebisi. 1991. The Concept of English as a Second Language as Applied Linguistic Policy for Development. *JESEL* 3: 6–21.
- Akere, Funsho. 2004. Nigerian English in sociolinguistic perspective: Uses, users and emerging varieties. In *Nigerian English: Influences and Characteristics*, Dadze, A. B. K. & Awonusi Segun (eds), 256–284. Lagos: Concept Publication.
- Awonusi, Victor O. 2004. The function of English in Relation to other Nigerian languages. In *Nigerian English: Influences and Characteristics*, Dadze, A. B. K. & Awonusi Segun (eds), 67–69. Lagos: Concept Publication.
- Babatunde, Sola T. 2001. The English as a second language phenomenon in Nigeria. In *Studies in Contemporary English Usage*, Babatunde, Sola T. (ed.), 105–150. Ilorin: Haytee Books.
- Bamgbose, Ayo. 1971. The English Language in Nigeria. In *The English Language in West Africa*, John Spencer (ed.), 34–48. London: Longman.
- Carston, Robyn. 1993. Conjunctions, explanation and relevance. *Lingua* 90: 27–48.
[https://doi.org/10.1016/0024-3841\(93\)90059-6](https://doi.org/10.1016/0024-3841(93)90059-6)
- Cook, Guy. 1989. *Discourse*. Oxford: OUP.
- Celce-Murcia, Marianne & Diane Larsen-Freeman. 1999. *The Grammar Book: An ESL/EFL Teacher's Course*, 2nd edn. Boston MA: Heinle & Heinle.
- Crewe, William J. 1990. The illogical of logic connectives. *ELT Journal* 44(4): 316–325.
<https://doi.org/10.1093/elt/44.4.316>
- Chen, Cheryl W. 2006. The use of conjunctive adverbials in the writing of Taiwanese EFL learners. *International Journal of Corpus Linguistics* 11(1): 113–130.
<https://doi.org/10.1075/ijcl.11.1.05che>
- Daramola, Adeyemi. 2012. A lexicogrammatical study of the uses of “like” in Nigerian English. *US-China Foreign Language* 10(9): 1523–1537.
- Dooley, Robert A. & Levinsohn, Stephen H. 2001. *Analyzing Discourse: A Manual of Basic Concepts*. Dallas TX: SIL International.
- Esimaje, Alexandra U. 2012. A corpus-based lexical study of sermons in Nigeria. *English Language Teaching* 5(9): 24–32.
- Esimaje, Alexandra U. 2013. Patterns of lexical collocations in sermonic texts. *Token: A Journal of English Linguistics* 2: 207–222.
- Fakhra, Amani. 2009. Relative Clauses and Conjunctive Adjuncts in Syrian University Student Writing in English. PhD dissertation, Warwick University.
- Fuchs, Robert & Gut, Ulrike. 2015. An apparent time study of the progressive in Nigerian English. In *Grammatical Change in English World-Wide* [Studies in Corpus Linguistics 67], Peter Collins (ed.), 373–387. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.67.15fuc>

- Fuchs, Robert, Gut, Ulrike & Soneye, Taiwo. 2013. "We just don't even know": The usage of the pragmatic focus particles *even* and *still* in Nigerian English. *English World-Wide* 34(2): 123–145. <https://doi.org/10.1075/eww.34.2.01fuc>
- Good, Edward C. 2002. *A Grammar Book For You and I-oops, Me!: All the Grammar You Need to Succeed in Life*. Sterling VA: Capital Books.
- Granger, Sylvaine. 1998. *Learner English on Computer*. London: Longman.
- Granger, Sylvaine & Tyson, Stephanie. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15(1): 17–27. <https://doi.org/10.1111/j.1467-971X.1996.tb00089.x>
- Gut, Ulrike. 2012. Towards a codification of Nigerian English: The ICE Nigerian Project. *Journal of the Nigeria English Studies Association (JNESA)* 15(1): 1–13.
- Gut, Ulrike & Fuchs, Robert. 2013. Progressive aspect in Nigerian English. *Journal of English Linguistics* 41(3): 243–267. <https://doi.org/10.1177/0075424213492799>
- Halliday, Michael A. K. & Hassan, Ruqaiya. 1976. *Cohesion in English*. London: Longman.
- Hamed, Muftah. 2014. Conjunctions in argumentative writing of Libyan tertiary students. *English Language Teaching* 7(3): 108–120. <https://doi.org/10.5539/elt.v7n3p108>
- Hasselgren, Angela. 1994. Lexical teddy bears and advanced learners: A study into the way Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4(2): 237–260. <https://doi.org/10.1111/j.1473-4192.1994.tb00065.x>
- Heino, Paula. 2010. Adverbial Connectors in Advanced EFL Learners' and Native Speakers' Student Writing. BA dissertation, Stockholm University.
- Hoey, Michael. 1991. *Patterns of Lexis in Text*. Oxford: OUP.
- Hubbard, Hilton. 1993. Some coherence correlates in exposition writing. *South Africa Journal of Linguistics Sup.* 15: 55–74. <https://doi.org/10.1080/10118063.1993.9724334>
- Hyland, Fiona. 1998. The impact of teacher-written feedback on individual writers. *Journal of Second Language Writing* 7: 255–286. [https://doi.org/10.1016/S1060-3743\(98\)90017-0](https://doi.org/10.1016/S1060-3743(98)90017-0)
- Hyland, Ken. 2009. *Academic Discourse*. London: Continuum.
- Ibrahim, Abidat A. 2015. The Past Perfect Tense in Nigerian and British English. A Comparative Corpus-based Study. MA thesis, International Islamic University, Malaysia.
- Ishikawa, Shin'ichiro. 2009. A corpus-based study on Asian learners' use of English linking adverbials. *Themes in Science and Technology Education* 3(1–2): 139–157.
- Jauro, Luka B., Ngama, John, Teneke, Gabril G. & Wada, Richard S. 2014. The use of clause relations in Nigerian Television Broadcast Media. *British Journal of Arts and Social Sciences* 11(1): 43–50.
- Jauro, Luka B., Teneke, Gabriel G., Bitrus, Ishaya B., & Moses, Ikper V. 2014. An evaluation of the use of discourse markers in Nigerian newspapers. *New Media and Mass Communication* 23: 25–29.
- Jowitt, David. 1991. *Nigerian English Usage: An Introduction*. Lagos: Longman.
- Kadri, Goodluck C., Igbokwe, Uche L., Okebalama, Udodirim N. & Egbe, Cajetan I. 2006. The use of lexical cohesion elements in the writing of ESL learners. *Research in Language* 3: 221–234.
- Malah, Zubair. 2015. Lexical cohesion in academic discourse: Exploring applied linguistics research articles abstracts. *Research Journal of English Language and Literature* 3(4): 291–299.
- Malgwi, Grace. 2016. A study of the character of lexical cohesion in ESL texts. *Literacy Information and Computer Education Journal* 7(1): 2211–2214. <https://doi.org/10.20533/licej.2040.2589.2016.0292>

- Martin, James R. & Rose, David. 2003. *Working with Discourse: Meaning Beyond the Clause*. London: Continuum.
- Martinez, Ana C. L. 2015. Use of conjunctions in the compositions of secondary education students. *Social and Behavioral Sciences* 212: 42–46.
- Matsuda, Paul K., Ortmeier-Hooper, Christina & Matsuda, Aya. 2009. The expansion of second language writing. In *The Sage Handbook of Writing Development*, Roger Beard, Debra Myhill, Jeni Riley & Martins Nysrand (eds), 457–471. Thousand Oaks CA: Sage.
<https://doi.org/10.4135/9780857021069.n32>
- McCarthy, Michael. 1991. *Discourse Analysis for Language Teachers*. Cambridge: CUP.
- Mo, Junhua. 2015. A contrastive study of the use of causal connectives by Chinese EFL learners and English native speakers in writing. *Theory and Practice in Language Studies* 5(11): 2426–2432. <https://doi.org/10.17507/tpls.0511.30>
- Mohammed, Asabe S. 2015. Conjunctions as cohesive devices in the writings of English as second language learners. *Procedia – Social and Behavioural Sciences* 208: 74–81.
<https://doi.org/10.1016/j.sbspro.2015.11.182>
- Muddhi, Saud K. & Hussein, Riyad F. 2014. Corpus-based study of conjunctive adjuncts in the writings of native and non-native speakers of English. *English Linguistics Research* 3(2): 18–32.
- Norrish, John. 1983. *Language learners and their errors*. Hong Kong: Macmillan.
- Ogunsanya, Ayoyinka O. 2009. The emergence of English as a first language. In *The Sociolinguistics of English and Nigerian Languages*, Adeyanju Dele (ed.), 129–156. Munich: Lincom.
- Ojetunde, Folashade & Okanlawan, Bidemi. 2011. The usage of English conjuncts by students in selected tertiary institutions in south western Nigeria. *The African Symposium* 11(1): 57–71.
- Okpongette, Alexander S. & Okon, Iniobong E.. 2016. Word formation processes in the Nigerian English-based pidgin. *Journal of Qualitative Education* 12(2): 232–239.
- Olateju, Mojisola A. 2006. Cohesion in ESL classroom written texts. *Nordic Journal of African Studies* 15(3): 314–33.
- Ong, Justina. 2011. Investigating the use of cohesive devices by Chinese EFL learners. *The Asian EFL Journal Quarterly* 11(3): 42–65.
- Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rahimi, Ali & Qannadzadeh, Jamal-Addin. 2010. Quantitative usage of logical connectors in Iranians' EFL Essay writing and logical and linguistic intelligence. *Procedia-Social and Behavioural Sciences* 5: 2012–2019. <https://doi.org/10.1016/j.sbspro.2010.07.406>
- Rayson, Paul & Garside, Roger. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora, held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL, October, 2000)*, 1–6. University of Science and Technology, Hong Kong.
- Scott, Mike. 2008. *WordSmith Tools version 5*, Liverpool: Lexical Analysis Software.
- Shen, Ying. 2007. *An Analysis of Discourse Markers Used by non-native English Learners: Its Implication for Teaching English as a Foreign Language*. Kanda University of International Studies. NII-Electronic Library Service. Web: <<http://www.kuis.ac.jp/icci/publica>>
- Shitu, Fatimoh M. 2015. Collocation errors in English as Second Language (ESL) essay writing. *International Scholarly and Scientific Research & Innovation* 9(9): 3270–3277.
- Sinclair, John McH. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

- Ting, Feng. 2003. An Investigation of Cohesive Errors in the Writing of PRC Tertiary EFL Students. MA thesis, National University of Singapore.
- Tseng, Yen-Chu & Liou, Hsien-Chin. 2006. The effects of online conjunction materials on College EFL students' writing. *System* 34: 270–283. <https://doi.org/10.1016/j.system.2006.01.006>
- Usman, Hafsat. 2015. The Relationship in the Use of Cohesive Patterns and the Written Discourse Performance of Hausa Students in Government Secondary Schools in Zaria. MA thesis, Ahmadu Bello University. <<http://kubanni.abu.edu.ng:8080/jspui/handle/123456789/7564>>
- Vande Kopple, William J. 1985. Some exploratory discourse on metadiscourse. *College Composition and Communication* 36(1): 82–93. <https://doi.org/10.2307/357609>
- Yang, Wenxing & Sun, Ying. 2012. The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education* 23: 31–48. <https://doi.org/10.1016/j.linged.2011.09.004>
- Yoon, Jung Wan & Yoo, Isaiah W. 2011. An error analysis of English conjunctive adjuncts in Korean College students' writing. *English Teaching* 66(1): 225–244. <https://doi.org/10.15858/engtea.66.1.201103.225>
- Zamel, Vivian. 1983. The composing processes of advanced ESL students: Six case studies. *TESOL Quarterly* 17: 165–187. <https://doi.org/10.2307/3586647>
- Zhang, Meisou. 2000. Cohesive features in the expository writing of undergraduates in two Chinese universities. *REL C Journal* 30(1): 61–95.

African corpora for standards in African academic English

Case studies on prepositions

Josef Schmied

Chemnitz University of Technology

This contribution proposes to replace the traditional native-speaker model at African universities with a sophisticated and stratified corpus model of nation-, university- and department-specific usage. It illustrates that the long-discussed realistic “national standard” may be possible – at least for restricted domains at advanced levels such as postgraduate studies. A qualitative and quantitative analysis of prepositions and their variation in three case studies shows that there is some flexibility in preposition usage: cases where different preposition choices can be explained by equally valid cognitive principles, cases where the addition of prepositions may be acceptable because this adds explicitness (which may be preferred in non-native contexts) and cases where prepositions may appear redundant since there is no choice and no semantic opposition. Although sociolinguistic values and attitudes may be necessary complements to linguistic frequency analyses, a careful corpus-linguistic study of prepositional choices irrespective of standardised native conventions is the basis for all discussions of new functional standards for African English.

1. Introduction

For the first descriptions of African English (cf. Schmied 1991) or national varieties in Africa (e.g. Jowitt 1991), no solid empirical database was available.¹ The first comprehensive attempt to include African English in a worldwide framework was the International Corpus of English (ICE; cf. Greenbaum 1996 or Schmied 2011),

1. Previous research, especially in Nigeria, started the discussion of usage and standards usually in unpublished PhD theses or university journals, only few are accessible internationally in edited volumes; they illustrate Nigerian scholarship as well as many usages before the corpus-linguistic beginnings in the 1990 (e.g. Bamgbose 1971, 1982).

where East Africa (i.e. Kenya/Tanzania), South Africa and Nigeria were included in the early stages; but only East Africa was completed in the mid-1990s, and ICE Nigeria has become available only in 2014. Today, more specialised corpora are available in some countries, such as the Corpus of Cameroon MA theses, which is analysed in this project. It is therefore not surprising that questions of corpus exploitation and application have hardly been asked in Africa in the context of empirical norms in higher education.² This is particularly surprising, since the question of standards of English has never been more urgent than in Africa today, where university exams cannot be sent to Cambridge for correction any more as in colonial days, but the expansion of university education has made the internal and external debate of language standards a national issue in many countries.

Since prepositions are particularly flexible (as illustrated in the three case studies below), they have been identified as problem areas in many usage descriptions of English in Africa. Thus, Jowitt (1991: 107) writes:

One is the area of phrasal and prepositional verbs. In PNE [Popular Nigerian English], but not in SBE [Standard British English], *cope* is followed by *up* and *discuss* by *about*, and these PNE features might be indicated in a PNE lexicon, especially as they are features peculiar to certain lexical items and involve no general rules. On the other hand, the structure of these lexical items might be properly regarded as a matter of syntax or morphology, since they are instances of verbs which may, or alternatively, may not, be followed by a preposition or an adverb.

But “standards” does not necessarily mean (British) Standard English (StE), a flexible national standard may be codified in usage dictionaries, for instance. A good example of such national usage standards is *A Dictionary of Nigerian English Usage* (Igboanusi 2002, s.v.), which includes a long list of verbs with non-standard prepositions (most are intrusive, a few replacive): *advocate* (+for), *air* (+out), *appoint* (always +as), *ask* (from), *charge with* (+for), *cope with* (+up with), *congratulate on* (for), *demand* (+for), *descend* (+down), *emphasize* (+on), *lend* (+out), *list* (+down), *pool* (+together), *refer* (+back), *regret* (+for), *repatriate* (+back), *repeat* (+again), *reply* (+back), *request* (+for), *return* (+back), *reverse* (+back), *substitute for* (with) *succeed in* (with), etc. – a few are special transitive cases like *pay* (+for) *the fees*, *rear* (+up) *a family*, *solicit* (+for) *news*, etc.

In this corpus-based analysis, I include three prototypical case studies of “replacement”, “addition” and “omission”, although these terms suggest too much normative

2. This is also why this volume fills a major research gap in providing a comprehensive overview of the challenges and opportunities of modern empirical work on Englishes in Africa. The small case studies in this and other articles can be used as models for further studies based on other data collected in other contexts.

learner perspectives for a “functional view”; hence, “alternative”, “explicit” and “implicit” may be more neutral appropriate terms to propagate non-native usages:

In cases like *fill in/out/up* the choice seems functionally redundant, as can be seen from the different standardisations in UK and US idiomaticity (*in* vs. *out*, respectively). From a cognitive perspective, all three prepositions are possible. A corpus-linguistic analysis reveals preferences depending on following noun semantics or the usual socio-textual variables, such as educational level and teaching-culture background, etc.

In cases like *discuss* [+*about*] the meaning difference may not be important, and it is still unclear whether a semantic difference can be detected in the collocates or in socio-textual variables. In cases like *suffer a dictatorship*, the preposition may be dropped by non-native speakers because in the context the meaning is always unambiguous, so that the preposition can be considered redundant.

Thus, this qualitative and quantitative analysis of preposition variations in different but comparable corpora of academic writing may not only help us to identify verbs after which more flexibility in prepositions is current usage, but also where this usage may be accepted as a functional standard, irrespective of native-speaker intuition or traditional codification.

2. A corpus-based functional view on standards for Africa

The issue of standards of African English has been discussed for a long time, usually in the context of “falling standards” in education (cf. Schmied 1991: 108ff.; Schmied 2012). But since Kachru’s concept of three concentric circles, including second-language countries as norm-developing (Kachru 1990), the theoretical discussion of standards has gained momentum in many African nations, where it can be seen as a case of covert codification (Hickey 2012: 20). The practical work on standardisation is extremely limited. Seidlhofer (2007: 316) explains the issue:

It is extremely desirable and reasonable to describe the norms of native speaker use as is now extensively done in corpus linguistics, but it is entirely undesirable and unreasonable to suppose that those norms should automatically serve as the default standards for pedagogical purposes. Of course, there need to be standards in language pedagogy, and this must involve prescribing some model of use that calls for conformity, otherwise both teachers and learners would be in a limbo and no teaching or learning would be possible, let alone measurement of learning outcomes. But the question arises as to what criteria such a model needs to meet.

2.1 Five criteria for accepting non-standard features

Non-native standards of English have the advantage that they can be scrutinised and analysed according to logical and processing usability. This has to be included in the five criteria that acceptable “non-standard” features have to fulfil to differentiate them from common learner errors: frequent usage even at advanced level, international usage, variational acceptability, functional appropriateness, transparency and processability. All these terms have to be discussed in detail:

Non-native standards have to be frequent, meaning that they should occur in different genres, different disciplines, on educational levels, especially on higher levels, and should not be related to individual writers or individual classes if they are taught in particular classes or schools. Non-native standards should be international whereby the same usages occur in different countries, for instance in the European Union, in Asia or in different varieties in South Africa, for instance.

When we say that non-native standards should be acceptable, this means that there should be no strong negative connotation in variational (stylistic, dialectal, or socio-linguistic) terms, whereby users of such a variety risk being classified as uneducated, which would go against the academic reputation sought in MA theses (the main database for this study). As Mair and Mollin (2007: 347) emphasise (in European contexts):

New standards need to be standards in the mind, too. Ideally, the speakers sampled in the corpus should thus be asked whether they consider features which have emerged in the corpus to be potential markers of a new variety as correct, or whether they would use these themselves. Under-reporting of new forms then implies a lack of institutionalization of the new variety, while answers to the correctness may indicate the degree to which the native-speaker standard is still the prestige model in the community.

When we say non-native standards should be functionally appropriate, this means that they should be justifiable in logical semantic terms. This also means that irregularities may be regularized, and sub-rules may be neglected without undue emphasis on what appears as conventions without function or unnecessary distinctions and sub-rules. This may be seen in the context of the processability hypothesis or theory that states that learners restructure their L2 knowledge systems in an order of which they are capable at their stage of development (Pienemann 1998).

Non-native standards should be transparent in the sense that the resulting texts should be intelligible in terms of their sense and pragmatic effect. This also means that they are processable, i.e. the forms should go unnoticed, without intentional cognitive salience according to formal criteria, if they are not, of course, intended as awareness-raising style features.

In conclusion, we can say that if we combine corpus-, cognitive and sociolinguistic methodologies, we may be able to identify new standards because of their frequencies in a wide range of styles and academic usages, their acceptability in online questionnaires, and their familiarity in eye-tracking experiments with the appropriate user groups.

2.2 New focus on “problem cases” in real usage

Most handbooks on academic non-native writing are useful for studying the principles and key concepts, but they tend to neglect many problematic or even controversial lexical and grammatical items in common usage, even at advanced (university) levels. In the following examples, we have deliberately chosen such “problem cases” and propose “neglecting the rules”, arguing that the functional necessity of choosing one correct form and deciding against a “deviant” form may not be necessary. However, I also try to indicate where such “generosity” has its functional limits because useful differences, even subtleties, may have such a high functional value in academic discourse that they should be maintained.

Most examples can be classified as problems of “under-differentiation”, which can be found in grammatical and lexical choices, not only in register and stylistic choices (cf. Schmied 1991: 129). If the choice is correlated with the choice of other clause elements (such as conjunctions or prepositions), the functional load is low, and the choice is functionally redundant (e.g. *fill in/out/up* and *discuss* +/- *about* below) – and can thus be accepted as an option, not an error. In some bases, under-differentiation may make a difference, but the meaning difference may not be important (e.g. *discuss about/on* below). This may even contradict the principle of clarity, because more explicit usually means clearer to the specialist and, in lexical cases, more concise.

In the following, I discuss examples of what is an error according to Standard English usage books and even modern grammars. In a comparative analysis (cf. Schmied 2015a, 2015b), however, all these deviant usages occur in academic writing by German, Chinese and African writers, so they can neither be mother-tongue- nor teaching-induced. This raises the issue of whether such usages simply contradict (often unsystematic) conventions and could be seen as acceptable, or at least not marked as explicitly incorrect, in academic *lingua franca* English (cf. Seidlhofer 2007).

3. Corpus data contexts

The contribution analyses prepositions in a (recent) corpus of Cameroon English MA theses. It puts these detailed analyses in two contexts: Firstly, two corpora of Cameroon writing are studied to show that prepositional usage in the corpus from the 2010s is similar to that in the 1990s,³ hence CCE90 and CamMA10 in the following analyses. The second context entails using selected valuable online resources to compare the wider international usage: the Electronic World Atlas of Varieties of English (eWAVE) (Kortmann & Luckenheimer 2013), the News on the Web corpus (NOW) and the Global Web-based English corpus (GlowBE) (Davies & Fuchs 2015).

The detailed qualitative and quantitative analysis of two types of prepositional usage is based on the Corpus of Cameroon MA theses (CamMA10) collected from the English Department at the University of Yaoundé I between 2010 and 2015, which contains more than 2.7 million words – a minimum to make it suitable for a quantitative analysis of preposition choice after verbs, as the following analysis shows.

The file names in the CamMA10 are more detailed than usual, because they contain information on the year and the gender of writers, so CAMDIPESII2011FMLF is a female writer from 2011, etc. This additional information is useful for sociolinguistic analyses but cannot be exploited in our case studies below, since the frequencies for the specific cases are only high enough for high-frequency linguistic variables like modal auxiliaries or adverbs.

4. Prepositions as a special case for variability in English grammar

4.1 Quantitative corpus comparisons

This section focuses on the frequency of simple prepositions in the two Cameroon corpora compared with two others included in the International Corpus of English (cf. Schmied 2011), from Kenya and Britain, prototypical English as a Second- and First-language nations respectively.

Due to the high frequency of simple prepositions,⁴ the analysis in this section is centred only on the ten most frequent forms, a common practice in corpus-based

3. Nkemleke (2019) explains the background and set-up of the CCE90, which is a very useful equivalent to the Brown/LOB family of corpora of written English.

4. A more detailed analysis of compound and complex prepositions in CCE90 can be found in Schmied & Nkemleke (2010), a similar approach in a contrastive English – German perspective in Schmied (2002), and other case studies based on the same corpus collections in Nkemleke (2011).

studies of prepositions (e.g. Mindt & Weber 1989; Mwangi 2003). A frequency profile of these prepositions in the Cameroonian and Kenyan English corpora is presented in Table 1, with British English frequencies included for comparison. To compensate for the unevenness of the corpora sizes, the figures for CCE90, CamMA10, ICE-K (401,813 words) and ICE-GB (400,000 words) have been normalized (per 1 million words).

Table 1. Frequency of simple prepositions in CCE90, CamMA01, ICE-K and ICE-GB

Preposition	CCE90	CCE90 (per 1 M)	CamMA10	CamMA10 (per 1 MA)	ICE-K (per 1 M)	ICE-GB (per 1 M)
of	28,842	35,149	103,347	37,802	27,416	27,314
to	24,328	29,648	80,469	29,434	25,478	9,004
in	17,273	21,050	68,998	25,238	17,556	15,714
for	8,323	10,143	20,234	7,401	8,480	7,328
on	4,877	5,943	15,257	5,580	5,596	5,246
with	4,875	5,941	15,088	5,518	5,148	5,068
by	4,356	5,308	13,760	5,033	4,626	4,714
at	3,370	4,106	7,706	2,818	3,858	3,634
from	3,632	4,426	12,305	4,500	3,542	4,032
about	1,284	1,564	3,839	1,404	1,496	1,116
Total	101,160	123,278	341,003	124,728	103,196	83,170

Overall, simple prepositions are more frequent in the Cameroonian and Kenyan corpus. This means that the most frequent prepositions are too frequent in the CCE and ICE-K. The preposition *to* is marked because of its very high frequency in the CCE and ICE-K. A close look at some of the first few hundred examples indicates one general trend, namely that *to* appears to be used repeatedly in a series of *to*-infinitive and other constructions in a single sentence as in (1)–(3).

- (1) ...through those grades is a method used by the executive to reward docile and obedient Judges who obey the executive and **to** punish Judges that dare **to** exhibit any degree of independence or refuse **to** do the bidding of the executive.
(CCE/mi019.txt)
- (2) Their condition of service should be revised to give them adequate pay and other facilities such as lodging to put them in a position to resist temptation such as bribery and corruption.
(CCE/mi019.txt)
- (3) <#/>You are requested **to** study the document and make your proposal **to** the employer **to** be submitted before 14.02.96 mid-day <#/>
(W1B-BK10)

What this actually demonstrates is a particular preference for the use of *to*, probably owing to the fact that *to* is a common structuring preposition which users of English can use for clear sequences in argumentation. However, when such users are not exposed to, or learn to practice other ways of expanding and projecting propositions (cf. Halliday & Matthiessen 2004), they resort to shorter sentences with the result that many turns in the discourse involve *to*-infinitives.

A more detailed analysis of compound and complex prepositions in CCE90 can be found in Schmied & Nkemleke (2010). Prepositions have been used particularly for cognitive approaches since the inspiring work by Lindstromberg (1998). This analysis focuses on the cognitive processes of “Simplification through incorporation”, which overlaps with “omission” in this analysis, and “Explicitness: from simple to complex forms”, which overlaps with “addition” in this analysis.

4.2 Three case studies on alternative, explicit and implicit prepositions

This section concentrates on usages of prepositions in Cameroon, which do not occur pervasively nor exclusively, i.e. they are not always used in this marked form in Cameroon, nor is this usage restricted to Cameroon; but they are common and frequent in many African countries (see Simo Bobda 2002 esp. for Cameroon). We distinguish three types of usages and try to illustrate them by giving detailed analyses of what we identify as prototypical but not easy cases. They are interesting linguistically in that they can be viewed alternatively in cognitive linguistic perspectives, and psycholinguistic explanations may be found for several alternatives.

4.2.1 *Replacement or (free variation) alternative prepositions: The case of fill in/out/up*

Variation means that different prepositions are alternatives without differences in meaning, i.e. the choice is functionally redundant, as in the case of *fill in/out/up*, which are all cognitively “logical” concepts. The concept of *in* (interiority) can be seen in this case as a hole or gap in the reference object (the landmark) that the located object is moved in (the trajectory); the concept of *out* focuses on the trajectory as completing the hole so that it no longer exists; the concept of *up* sees this filling in a vertical dimension.

The AntConc analysis of the CamMA Corpus shows 93 hits for *fill in*, mostly *fill in+the [form/gap]* as in (4), but also other more special constructions, such as related alternatives for *gaps* (as in Examples (5)); unusual collocates (with special semantics in (6)); unusual syntax (with a noun clause in (7)) and an unusual repetition (with the first *in* in (8) dependent on *fill*, the second on *class*):

- (4) The Authors should equally endeavour to use more meaning-based activities to practice the grammar points treated in the student's book because in real life situations, language is not used to *fill in* gaps per se, but to convey ideas and to fill in information gaps. (CAMDIPESII2010FECN)
- (5) It took over one month to convince teachers to *fill in* questionnaire and return. (CAMDIPESII2011FMLF)
- (6) Depending on the results obtained from such a test or assignment as the case may be, the teacher might likely decide to organize follow-up activities to *fill in* students on aspects of the text they missed out. (CAMDIPESII2010FALC)
- (7) Please a tick () in the box against the answer of your choice or *fill in* where necessary. (CAMDIPESII2011FCAM)
- (8) The students were given the questionnaires to *fill in* in class and the researcher collected them immediately after. (CAMDIPESII2012FAGH)

Whereas four cases of *fill out* in our corpora are clear metalanguage examples (where the American forms are discussed) and only two are used with *questionnaires*, four (special) cases of *fill up* need to be analysed carefully. (9) can be seen as a convenient metaphorical usage of a container of place to become full (AOLD, s.v.), the other cases of *up* in front of *gap* (10) or *hollow/hole* (11) and (12) seem redundant:

- (9) It is just something temporal to *fill up* her hours of tedium and idleness while her man is away at work. (CAMDIPESII2011MDTD)
- (10) The study revealed that boys can be more polite than girls when it comes to euphemistic usage because girls tend to *fill up* the gap with their mimic and general comportment. (CAMDIPESII2010MAN)
- (11) At Coetzee's greatest moment of need, it is Anya, a white Filipino woman who is there to *fill up* the gaping hollow in the old man's life. (CAMDIPESII2011MDTD)
- (12) In attempting to convert the darkman to the White man's way of life, the novel lectures us, the white man has fallen helplessly down the hole he wanted to *fill up* which is Kurtz's fate exactly. (CAMDIPESII2012FENE)

To compare prepositional choice in all Anglophone countries worldwide, we can put the occurrences of *fill in*, *out* and *up* in one figure (Figure 1) and see that all forms with *in* predominate in British-oriented nations (including the non-native Hong Kong and South Africa, for instance); all forms with *out* can be found in North American and Jamaica and partly in Ireland and New Zealand; and *up* forms

seem to be Asian, since they appear prominently in India, Bangladesh, Singapore/ Malaysia and even the Philippines (despite the American heritage there).

Corpus of Global Web-Based English (GloWbE)

SEARCH

FREQUENCY

CONTEXT

HELP

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP...]

COMPARE

	☒ CONTEXT	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	☐ FILL IN	14316	2499	1011	3650	860	1102	789	587	211	181	274	319	381	327	484	451	239	207	281	284	179
2	☐ FILLED IN	4273	790	314	956	277	305	262	265	73	68	115	65	70	84	115	152	51	48	91	101	71
3	☐ FILLING IN	3866	698	307	1089	250	323	213	152	44	38	62	93	70	58	95	86	48	44	82	62	52
4	☐ FILLS IN	761	216	55	184	41	57	33	34	14	2	7	26	8	10	15	16	11	6	8	10	8
1	☐ FILL OUT	8867	2429	1321	1131	582	632	463	294	99	112	198	118	148	233	171	148	102	138	190	145	213
2	☐ FILLING OUT	2713	783	355	354	179	206	144	93	31	28	59	46	37	65	47	48	40	21	61	46	70
3	☐ FILLED OUT	2511	795	326	314	174	185	124	58	20	28	38	45	40	56	48	30	21	24	65	44	76
4	☐ FILLS OUT	273	88	20	54	11	24	9	16	4	2		1	5	6	4	6	4	5	5	3	6
1	☐ FILL UP	6405	1185	437	1092	246	434	229	499	122	128	249	307	332	311	159	141	109	103	120	118	84
2	☐ FILLED UP	2956	507	195	447	141	217	113	279	71	82	102	102	108	156	69	56	84	53	54	86	34
3	☐ FILLING UP	2710	493	183	549	106	206	111	232	64	39	107	88	102	98	62	70	29	28	59	44	40
4	☐ FILLS UP	761	170	60	152	39	54	30	52	13	18	21	29	22	20	15	21	11	5	16	7	6

Figure 1. The prepositional choice after *fill** in the GloWbE corpus

4.2.2 Addition or explicit preposition: The case of *discuss about*

Addition means that a preposition can be seen as redundant (but also as explicit), as in the case of *discuss about* (although it can be seen as an analogy of *talk about*) because *about* is (among others) the topic marker in English, since topic is the default noun phrase after *discuss*.

Discuss about* provides 8 hits in our CamMA concordance (*discussion about* many more, although it is not mentioned in the standard learner dictionaries LDOCE or OALD) – relatively few, compared to the 300 other cases of *discuss*, which are mostly immediately followed by a noun. Whereas (13) shows how deeply the need from the preposition is felt, since it seems unnecessarily completed, the following (14)–(16) are instructions, and thus probably all teacher induced. The rest show the syntactic options prepositional object (17) and even clause (18):

- (13) The plethora of themes in this text gives students the opportunity to have a wide range of topics to *discuss about*. (CAMDIPESII2012MSAV)
- (14) In groups of five *discuss about* the characters in the text and fill in the table below. (CAMDIPESII2010FGAS)
- (15) *Discuss about* events in the students' own country during the period of the story. (CAMDIPESII2011FMMN)
- (16) Lesson Objectives: By the end of this lesson, students should be able to: Respond to questions on the scenes which reflect death in the play. *Discuss about* the causes and consequences of death in the play. (CAMDIPESII2010MMNB)

- (17) Unlike the conventional plays which *discuss about* society and its politics Theatre of the Absurd focuses more on the metaphysical aspect of life than the physical.
(CAMDIPESII2010MGDT)
- (18) The teacher can, at certain levels of the projection, pause it and together with the students *discuss about* what is going on at each stage of the projection.
(CAMDIPESII2012MJM)

To check whether the form *discussed about* is used in other varieties of English today, we consulted the NOW corpus and retrieved the result in Figure 2, which shows that it occurs quite frequently and the first instances all come from India. This was confirmed by a search in GloWbE, which shows the result in Figure 3, where Australian, Canadian and British examples were retrieved. Finally, the country-specific display in Figure 4 clearly shows that this is first and foremost an Asian usage (Sri Lanka, Bangladesh and Malaysia feature relatively more prominently than India), although Nigeria and Tanzania indicate that it does occur regularly in Africa, maybe at a lower level.

NOW Corpus (News on the Web)				SEARCH	FREQUENCY	CONTEXT	CONTEXT +
FIND SAMPLE: 100 200 500 1000 PAGE: << < 1 / 15 > >>							
CLICK FOR MORE CONTEXT				☐ ☒	SAVE LIST	CHOOSE LIST	CREATE NEW LIST
1	18-01-20 IN	The New Indian Express	A B C	he had no idea of it being such a serious matter. " No one discussed about it in my family when I was growing up. It was only when			
2	18-01-20 IN	Telangana Today	A B C	Prof. A Venugopal Reddy on implementing Model Curriculum in the university. The meeting also discussed about surprise inspections and i			
3	18-01-20 IN	Morning Express	A B C	22850945 # During the technical session, the resource person discussed about the various management techniques involved in successful			
4	18-01-20 IN	The Asian Age	A B C	as the party's stand. " he said. # The core committee also discussed about the party's preparations for the bypoll. It decided that both leve			
5	18-01-19 IN	Mathrubhumi English	A B C	emotions. Jithu's grandfather Johnny Kutty and grandmother Ammini John said they had never discussed about property or sharing of the pr			
6	18-01-18 IN	Outlook India	A B C	chance meeting with the Prime Minister in December 2015, wherein we discussed about how Khadi and Village Industries Commission (KVI			
7	18-01-17 IN	The New Indian Express	A B C	the earlier summits had fructified. # The Chief Minister said: " Whatever we discussed about the business summit, you can not say that we			
8	18-01-17 IN	Business Standard	A B C	and those that had fructified. # The Chief Minister said: " Whatever we discussed about the business summit... you can not say that we only			
9	18-01-17 IN	Morning Express	A B C	mischief. " It appealed. # It further clarified that the COB meeting never discussed about ticket issue for the ensuing elections. The NPF said			
10	18-01-16 IN	DailyO	A B C	Kashmir does not show any sign of ending anytime soon. A lot is being discussed about the role of Pakistan in supporting this insurgency g			
11	18-01-12 IN	NYOOOZ	A B C	to be cordial and sympathetic with poor patients and treat them right.On the occasion he discussed about the shortage of staff, medical eq			
13	18-01-12 IN	Morning Express (1)	A B C	The constraints and possibilities of a settlement # Much has been either speculated or discussed about the likely Political Settlement even i			
14	18-01-11 IN	The Shillong Times	A B C	Thursday for bringing about revival and stimulating growth of traditional industry. # The meeting discussed about various ways to preserve			
15	18-01-11 IN	The Shillong Times	A B C	Industry through different measures keeping with the modern industrial practices. # The meeting also discussed about methods of encour			
16	18-01-11 IN	Odisha Television Ltd.	A B C	patients under Niramaya Yojana. # Charing a meeting of the CDMOs, the Minister discussed about the government schemes for patients. >			
17	18-01-11 IN	Business Standard	A B C	Ben Wallace, UK Minister for Security & Economic Crime was fruitful. # We discussed about issues of cooperation in cyber security, radicaliz			
18	18-01-11 MY	malaysiandigest.com	A B C	# Meanwhile his lawyer Josefine revealed that Ahok and his wife Veronica Tan had actually discussed about his intention to divorce her late			
19	18-01-08 IN	Firstpost	A B C	with a collection of over Rs 130 crore. # While a lot has been discussed about Padmavat, Bal's Padman is a significant film that showcases			
20	18-01-06 IN	DailyO	A B C	the politically aware Dalits elsewhere, is now a national news. # Many have discussed about what happened in 1818 at Bhima Koregaon, w			
21	18-01-06 MY	Malay Mail	A B C	have 17 seats there. A lot of capable candidates, but we have not discussed about these candidates, because I think we should win first. " #			
22	18-01-06 NG	Ripples Nigeria	A B C	Association of Nigeria nor has he ever been a member. He has also never discussed about that group with Chief Paul Unongo. # " Furtherm			
23	18-01-05 IN	Free Press Journal	A B C	security professionals. # Offered a platform to discuss technologies & its impact. # Discussed about attackers getting smarter. # " Need to ke			
24	18-01-05 NG	The Punch	A B C	Association of Nigeria nor has he ever been a member. He has also never discussed about that group with Chief Paul Unongo. # " Furtherm			

Figure 2. The use of *discussed about* in the NOW corpus texts (on 18/01/20)

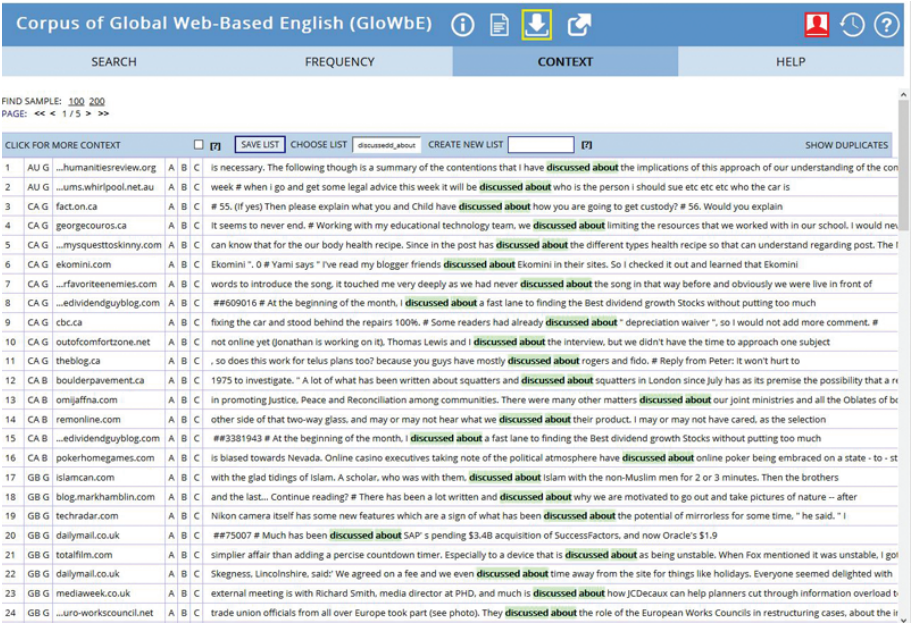


Figure 3. The use of *discussed about* in the GloWbE corpus texts (on 18/01/20)

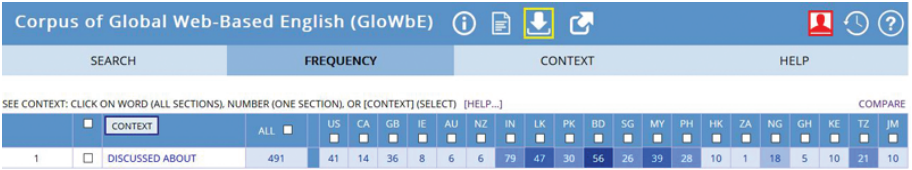


Figure 4. The use of *discussed about* in the GloWbE corpus in all Anglophone nations (on 18/01/20)

4.2.3 Omission or implicit preposition: The case of suffer

Omission means that a preposition that is considered obligatory in Standard English (StE) is not used to add a noun phrase. The preposition can be seen as integrated or implicit in the intransitive or transitive verb, as in the case of *suffer* [under/from] a dictatorship/a lot of hunger, where the meaning difference between *suffer from an illness* and *suffer under the system* is disregarded. The 211 cases of *suffer* in the CamMA corpus do not yield many unusual examples, except: (19) clearly demands a *from* in StE, (20) possibly an *under*, (21) and (22) also a possessive *their*.

- (19) The latter constantly and persistently undergo physical suffering, and more importantly, they *suffer* __ psychological trauma in their own societies.
(CAMDIPESII2014MBB)
- (20) Enitan says that the “only place where they *suffered* a dictatorship was in their pockets” (235). Enitan feels heartbroken and pitiful for this poor woman selling sugar cane by the roadside.
(CAMDIPESII2010FEAE)
- (21) ... should make it no very special cause for surprise that women develop good characteristics common to those who *suffer* minority status and a marginal existence.
(CAMDIPESII2012FNNN)
- (22) As a result of ethnic discrimination, African Americans *suffer* __ unemployment problems.
(CAMDIPESII2010FCDN)

The most frequent prepositions are *suffer** *from*, *in*, and *under*; whereas *suffer in Africa* and *suffer in silence* are normal intransitive usage, *people suffer in abject poverty* can be seen as a case of preposition substitution for *under* (if not as adverbial), just as in the pattern *Citizens suffer in the hands of neo-colonial leaders/of their own brothers/of authorities* and even *of bad laws*. Out of 816 AntConc cases of *suffer**, 160 are followed by *from*, 51 by *in* and only 11 by *under*. This suggests rather a shift from *under* to 0 or to *from* and *in*, but only few of these prepositional phrases depend on the verb; most are adverbial expansions. These cases can also be seen as under-differentiation, where the choice plus or minus preposition is not important.

The following examples with the verbs *arrive*, esp. when immediately following a place (name) as in (23)–(25)) and *reply* (usually before *letter* (26) and (27)) were found in the Cameroon corpora. Here, the analogy is with the related words *reach* and *answer*, which do not require a following preposition.

- (23) After three hour’s drive we *arrive*__ “Cardiff” the capital of “P Wales”. It is the third biggest town in the “UK”.
(CCE/pl133).
- (24) Even though Max registers him again for the late entrance, he is not the least optimistic that the document will *arrive*__ Cameroon. (CAMDIPESII2012MJM)
- (25) Our school resume on the 15th of September after enjoying a long and boring holiday. When I *arrive*__ the campus on the 15th I found things a little abnormal until the first week was over, then I came back to a new stand.
(CCE/pl038)
- (26) Tell her to *reply*__ my letter if she knows she is my friend (CCE/pl.114)
- (27) <I><#/> Hello <name/>, <#/>How are you? <#/>How are things? <#/>Us guys down here are doing fine, am really sorry I could not *reply* __ your letter soonest, <slang/>coz I had <-/alot> of jobs to finish. (WIB-SK31).

Explanations for this type of simplification, integration or implicitness have been discussed from the L1 interference perspective and complexity of prepositional meaning in English by many scholars including Schmied (1991: 67), who argues that “the use of fixed prepositions with many verbs is more expansive in English than in African languages, and the idiomatic prepositions may follow semantic, morphological or even traditional Latin rules, which are often difficult for [ESL] learners”. Despite efforts at teaching prepositions in ESL schools, some of these features persist across virtually all text types, esp. in Cameroon. As this type of modification becomes entrenched in these countries (cf. Sey 1973; Angogo & Hancock 1980; Platt, Weber & Ho 1984; Jibril 1991), recognition of such features as stable Africanisms is becoming widely acknowledged.

For a worldwide comparison of omission of prepositions, we can refer to eW-AVE, the electronic World Atlas of Varieties of English. This is the only preposition feature (#216) included in the world-map (Figure 5) and it shows that the omission of StE prepositions exists not only in pidgins and creoles, but also in several African Englishes. The examples from the GloWbE corpus confirm this (Figure 6).

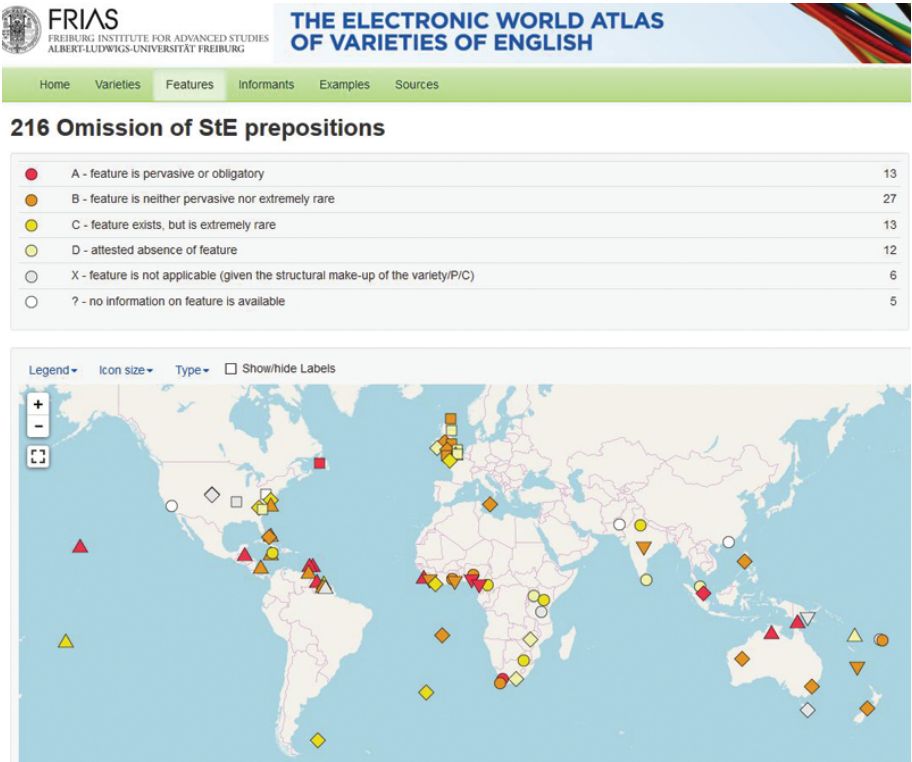


Figure 5. Omission of StE prepositions in eWAVE

Corpus of Global Web-Based English (GloWbE)																										
SEARCH			FREQUENCY			CONTEXT															ACCOUNT					
SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP...]																										
			CONTEXT			ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
1	☐	SUFFERING	103415	18175	5955	20984	5444	7638	3926	5253	7895	3248	2141	1946	3675	2044	2501	2094	2485	2136	2542	1745	1588			
2	☐	SUFFER	76118	14607	4622	16778	3969	5845	2932	4152	2219	2464	1683	1612	1992	1757	1527	1848	2095	1537	1847	1373	1267			
3	☐	SUFFERED	71410	11879	4495	18219	4547	5233	2866	3303	2377	2364	1382	1159	1446	1302	1306	1639	1773	1683	1615	1267	1555			
4	☐	SUFFERS	14861	2962	1018	3333	692	1108	537	893	400	554	370	328	290	289	278	346	332	279	334	278	240			
1	☐	SUFFERING FROM	24204	3345	1430	5439	1274	1550	886	1828	626	1019	803	555	621	574	642	513	613	678	671	586	551			
2	☐	SUFFER FROM	19850	3287	1248	4420	1057	1491	734	1334	476	609	602	492	472	436	531	605	324	389	511	486	346			
3	☐	SUFFERED FROM	9533	1519	550	2728	544	634	321	433	296	275	222	166	191	190	269	195	167	202	208	211	212			
4	☐	SUFFERS FROM	6615	1290	483	1571	298	426	215	408	142	271	181	170	122	109	135	172	104	110	143	135	130			
1	☐	SUFFERED UNDER	550	97	30	90	21	41	21	18	39	21	10	10	13	16	5	23	13	20	28	20	14			
2	☐	SUFFERING UNDER	458	99	27	87	17	38	17	11	30	8	14	8	12	9	8	12	11	22	12	11	5			
3	☐	SUFFER UNDER	348	104	24	50	7	23	16	14	6	7	7	10	6	6	6	10	16	10	7	8	11			
4	☐	SUFFERS UNDER	47	8	3	7	2	6	1	3	2	1	2	3	1	1	2		1	3			1			

Figure 6. Word choice after *suffer** in the GloWbE corpus

5. Conclusions

This comparative analysis of African usages⁵ has argued that the “African English” preposition choice after verbs that is generally considered frequent for African English may be explained cognitively. Even advanced users of English may process the language differently from the StE models, which does not necessarily contradict a general standard usage pattern. Such a new focus on functional standards may make it easier for learners, even those of advanced proficiency (i.e. level C by the standards of the Common European Framework of Reference for Languages), but also relatively young academic writers at African universities, to accept a modified usage standard as a general guideline.

At national level, it remains to be seen whether professional decision makers take this as a decisive step in simplifying their work or whether they see this as a dangerous path towards promoting incomprehensible academic writing, a new variant of the permanent “falling standards”. At international level, publishers may have to be more generous in accepting new functional standards for English. The three case studies on alternative, explicit and implicit prepositions put forward in this article may be considered helpful for the discussion of concrete examples at African universities and African examination centres. This modest contribution

5. An analysis of the cases discussed here in the much bigger corresponding Chinese MA corpus confirms most of the results obtained from the Cameroon corpora, but clearly less extreme: The preposition after *fill** is predominantly *in*, less often than expected (because of the American Standard model) *out*, but also *up*. The preposition *about* occurs relatively less often after *discuss**. The standard preposition after *suffer* is *from* and only rarely *in* or *under*, but that may, of course, depend on the meanings in context.

can only illustrate how an empirical corpus-based study can help to decide on such important issues. Such cases could be tolerated as functionally “acceptable” and not marked wrong if they are perceived as transparent and do not cause particular communication problems. For applied language teachers and researchers, this means that a detailed corpus-linguistic description of usages in their specific African learning and publishing environments (e.g. Kautzsch 2019) would make valuable contributions to language development in their immediate contexts in schools, and universities, and the nation as a whole.

Of course, such usage analyses have to be complemented by widely distributed (online) attitude questionnaires and digital perception studies, which use digital humanities technology like eye-tracking (Schmied & Hofmann *fc.*) that can prove the “naturalness” of such usages,⁶ not only from the production, but also from the reception side. If advanced non-native academic readers do not visibly hesitate in their reading when they come across the usages described above, non-standard usages cannot be rejected as “disturbing”, to say the least. Thus, we can triangulate investigation methodologies and substantiate recommendations to accept new functional standards, which are until today rarely accepted by native speakers. From the examples above, *discuss about* seems most controversial, stigmatized as education-based substandard, although communicative problems cannot occur. Many more studies may be necessary to illustrate that such social attitudes are not supported by linguistic facts. Even if this is a long way in coming, in the end, it may be a shorter way for young African academics to find scholarly recognition of their work.

Acknowledgements

This is a continuation of my long-term collaboration with Daniel Nkemleke in Yaoundé. I am grateful to him for his collaboration in the joint publications (Schmied & Nkemleke (eds) 2016; Schmied & Nkemleke 2010) and for providing the data for the Cameroon MA Corpus, on which part of the following analyses and most of the following examples are based. I also thank Alexandra Esimaje for various discussions on the topic during our research meetings.

6. This is also supported by Mesthrie (2017: 344f.), who includes “Use of prepositions with verbs like *mention (about)*, *discuss (about)*” in his list of L2 SSE syntax and concludes: “The case for universality of L2 psycholinguistic processing is therefore strong”.

References

- Angogo, Rachel & Hancock, Ian. 1980. English in Africa: Emerging standards or diverging regionalisms? *English World-Wide* 1: 67–96. <https://doi.org/10.1075/eww.1.1.07ang>
- Bamgbose, Ayo. 1971. The English language in Nigeria. In *The English Language in West Africa*, John Walter Spencer (ed.), 35–48. London: Longman.
- Bamgbose, Ayo. 1982. Standard Nigerian English: Issues of identification. In *The Other Tongue: English across Cultures*, Braj Kachru (ed.), 99–111. Urbana IL: University of Illinois Press.
- Davies, Mark & Fuchs, Robert. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36: 1–28. <https://doi.org/10.1075/eww.36.1.01dav>
- Halliday, Michael A. K. & Matthiessen, Christian M. I. M. 2004. *An Introduction to Functional Grammar*, 3rd edn. London: Edward Arnold.
- Hickey, Raymond. 2012. Standard English and standards of English. In *Standards of English: Codified Varieties around the World*, Raymond Hickey (ed.), 1–33. Cambridge: CUP.
- Hickey, Raymond. 2012 (ed). *Standards of English: Codified Varieties around the World*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139023832>
- Igboanusi, Herbert. 2002. *A Dictionary of Nigerian English Usage*. Ibadan: Enicrownfit.
- Jibril, Munzali. 1991. The sociolinguistics of prepositional usage in Nigerian English. In *English around the World: Sociolinguistic Perspective*, Jenny Cheshire (ed.), 519–537. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511611889.036>
- Jowitt, David. 1991. *Nigerian English Usage: An Introduction*. Ibadan: Longman.
- Kachru, Braj B. 1990. *The Alchemy of English: The Spread, Functions, and Models of Non-native Englishes*. Urbana IL: University of Illinois Press.
- Kautzsch, Alexander. 2019. Lexical and morpho-syntactic features of Namibian English on the web: A corpus-based approach. In *Corpus Linguistics and African Englishes* [Studies in Corpus Linguistics 88], Alexandra Esimaje, Ulrike Gut & Bassey Antia (eds). Amsterdam: John Benjamins. (this volume)
- Kortmann, Bernd & Lunkenheimer, Kerstin (eds). 2013. *The Electronic World Atlas of Varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <<http://ewave-atlas.org>> (22 January 2018).
- Lindstromberg, Seth. 1998. *English Prepositions Explained*. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.88>
- Mair, Christian & Mollin, Sandra. 2007. Getting at the standards behind the standard ideology: What corpora can tell us about linguistic norms. In *Anglistentag 2006. Proceedings*, Sabine Volk-Birke & Julia Lippert (eds), 341–353. Trier: Wissenschaftlicher Verlag.
- Mesthrie, Rajend. 2017. English in Africa – a diachronic typology. In *The History of English: Varieties of English*, Alexander Berg & Laurel Brinton (eds), 330–348. Berlin: De Gruyter.
- Mindt, Dieter & Weber, Christel. 1989. Prepositions in American and British English. *World Englishes* 8(2): 229–238. <https://doi.org/10.1111/j.1467-971X.1989.tb00658.x>
- Mwangi, Serah. 2003. Prepositions in Kenyan English: A Corpus-based Study in Lexical-Grammatical Variation. PhD dissertation, Chemnitz University of Technology.
- Nkemleke, Daniel. 2011. *Exploring Academic Writing in Cameroon English: A Corpus-based Perspective*. Göttingen: Cuvillier.

- Nkemleke, Daniel. 2019. Practical corpus linguistics: Designing and exploiting a written corpus for research with special reference to Cameroon English. In *Corpus Linguistics and African Englishes* [Studies in Corpus Linguistics 88], Alexandra Esimaje, Ulrike Gut & Bassey Antia (eds). Amsterdam: John Benjamins. (this volume)
- Pienemann, Manfred. 1998. *Language Processing and Second Language Development: Processability Theory* [Studies in Bilingualism 15]. Amsterdam: John Benjamins.
<https://doi.org/10.1075/sibil.15>
- Platt, John Talbot, Weber, Heidi & Lian, Ho Mian. 1984. *The New Englishes*. London: Routledge & Kegan Paul.
- Schmied, Josef. 1991. *English in Africa: An Introduction*. London: Longman.
- Schmied, Josef. 2002. Prototypes, transfer and idiomaticity: An empirical analysis of local prepositions in English and German. In *Studies in Contrastive Linguistics*, Luis Iglesias Rabade & Susanna M. Doval Suarez (eds), 947–959. Santiago de Compostela: University of Santiago de Compostela.
- Schmied, Josef. 2011. Using corpora as an innovative tool to compare varieties of English around the world: The International Corpus of English. *Rassegna Italiana di Linguistica Applicata* 1–2: 21–37.
- Schmied, Josef. 2012. Standards of English in East Africa. In *Standards of English: Codified Varieties around the World*, Raymond Hickey (ed.), 229–255. Cambridge: CUP.
<https://doi.org/10.1017/CBO9781139023832.013>
- Schmied, Josef. 2015a. Graduate academic writing in Europe in comparison: A research-based approach to metalanguage and genre. In *Graduate Academic Writing for South Eastern Europe: Practical and Theoretical Perspectives* [REAL Studies 8], Josef Schmied (ed.), 1–24. Göttingen: Cuvillier.
- Schmied, Josef. 2015b. Academic writing world-wide: Comparing metadiscourse conventions, credibility, and new functional standards. In *Academic Writing and Research across Disciplines in Africa. From Students to Experts*, Josef Schmied & Daniel Nkemleke (eds), 9–34. Göttingen: Cuvillier.
- Schmied, Josef & Hofmann, Matthias. Forthcoming. New digital methodologies for old grammar problems: Corpus analyses and eye-tracking to discover non-native English article usage preferences. *Anglistentag 2017. Proceedings*. Triers: Wissenschaftlicher Verlag.
- Schmied, Josef & Nkemleke, Daniel. 2010. Prepositions in Cameroon and Kenyan English: Comparability on the basis of text corpora. *Syllabus Review* 1: 31–53.
- Schmied, Josef & Nkemleke, Daniel (eds). 2016. *Academic Writing and Research across Disciplines in Africa. From Students to Experts* [REAL Studies 10]. Göttingen: Cuvillier.
- Seidlhofer, Barbara. 2007. English as a lingua franca and communities of practice. In *Anglistentag 2006. Proceedings*, Sabine Volk-Birke & Julia Lippert (eds), 307–318. Trier: Wissenschaftlicher Verlag.
- Sey, Kofi A. 1973. *Ghanaian English: An Exploratory Survey*. London: Macmillan.
- Simo Bobda, Augustin. 2002. *Watch your English: A Collection of Remedial Lessons on English Usage*, 2nd edn. Yaoundé: B & K Language Institute.

Semiotic signature of transformation in a diachronic corpus of a South African political party

Bassey E. Antia and Tamsyn Hendricks

University of the Western Cape

Corpus analysis has become established as an approach to the study of language description or for applied pursuits in language teaching, terminology, and so on. However, because of the social indexicalities of language use, corpora can also inform studies of social phenomena. This chapter draws on social semiotics to argue that, in the analysis of social phenomena, meanings that are socially significant can be read not only from what is said in corpora, but also from a range of other resources, such as names of persons and places as well as language choices made in texts. This chapter thus uses two heuristics, onomastics and discursive mono-/multilingualism, to query a diachronic corpus associated with a South African political party for evidence of whether or not the party has over time become more inclusive, contrary to its discursive positioning by a rival party as an untransformed organisation. The analysis shows evidence of the party opening up to diversity in terms of race, gender, geography, and language choice, but the finding raises the question of the relationship between semiotic evidence and reality.

1. Introduction

Preceding chapters in this volume have illustrated how corpora have been created and/or used to study Englishes in Africa. A few disclaimers are, however, in order. Corpora in Africa are, obviously, not only about variational studies of English. They have supported the development, enhancement, or evaluation of a broad range of natural language processing tools in Africa: a speech synthesizer in Ibibio (Ekpenyong et al. 2014), part of speech tagging in Bantu languages (de Pauw et al. 2012), a statistical English-Luo machine translation system (de Pauw et al. 2010), spell checkers, as well as extended applications such as language identification in Gikuyu, isiZulu, Tiv, Swahili, Hausa and Igbo (Akosu & Selmat 2013; Chege et al.

2010; Bosch & Eiselen 2005). The use of corpora for the development of predictive texting and translation memory solutions in other environments has also been described (Antia 2009).

Corpora do not have to be only about spoken or written language use. They can involve other meaning-making resources, including images, which present a different set of challenges for established understandings of corpus and the corpus analysis process (Baker 2006). Similarly, corpora and the corpus analysis process are not limited to the study of language as an end in itself or for such established applied pursuits as language teaching (Aijmer 2009; Sinclair 2004) or lexicography and terminology (Benson, Benson & Ilson 1986; Antia 2000); because of the social indexicalities of language use, corpora can also inform studies of social phenomena. They have been used to study sentiments on financial markets (Ahmad 2011), public health messaging (Antia & Razum 2012), knowledge evolution (Ahmad 1996), and so on.

In this chapter, the focus is on the affordances of corpora for studying the identity politics of a political organisation in South Africa. Much work on political discourse analysis (van Dijk 1997; Fairclough & Fairclough 2012; Chilton 2004) has focused on the (contents of the) *sayings* of political actors within institutionalised contexts. Surely, political actors have other modalities for appending their signatures onto the semiotics of the political landscape, beyond the contents of what they say. In this chapter we attend to the *beings* or identities of these actors as may be gleaned from alternative datasets. As a way of underscoring even further the analytical and interpretive possibilities of corpora which we have seen in previous chapters, we analyse the semiotic potential (van Leeuwen 2005) of onomastics and discursive mono-/multilingualism in a diachronic corpus of texts that is associated with a South African political party and its predecessor organisations. We use two heuristics, onomastics and discursive mono/multilingualism, to query the corpus for an answer to the question: what semiotic evidence is there of the Democratic Alliance (and its predecessor organisations) having transformed beyond the way the party is interactively positioned by the rival African National Congress and others as an untransformed White party, one that merely rents Black faces and hires Black leaders? (Mottiar 2015; City Press 2014).¹

Subsequent sections of the chapter address the following in turn: a political history of South Africa, political discourse analysis and the negotiation of identities

1. In this chapter, the following labels common in contemporary South African discourses on population groups are used: Black, Coloured, Indian/Asian and White. These discourses do however sometimes use Black in a much broader sense that includes all non-White groups. This chapter uses each of these labels with an initial capital letter, irrespective of whether the label is used as a noun or an adjective.

of political actors, onomastics and discursive mono/multilingualism as heuristics for a corpus-based political discourse analysis, methods, results and findings, discussion, and conclusion.

2. South Africa: A historical précis of a political landscape

South Africa is a multicultural and multilingual country, with eleven, soon twelve, official languages. After some 80 odd years of minority British, then Afrikaner rule, the country came under Black majority rule in 1994 when Nelson Mandela was elected President in the country's first ever democratic elections. In 1948, the Afrikaners, who are descendants of Dutch settlers, acceded to power and introduced apartheid or the strict separation of races. The territory of the country was carved up into zones for the different races and ethnicities. The particular focus was on zoning Black Africans according to the languages they were believed to speak. Thus, Black Africans were placed in ten Bantustans or so-called homelands – isiXhosa-speaking Blacks were placed in two homelands, Ciskei and Transkei; isiZulu-speaking Blacks were enclaved in KwaZulu; Lebowa was created for the northern Sotho people and so on.² These homelands were created by the apartheid government as 'tribal' areas in which the Black population would have so-called rights (such as voting) which they were denied in White South Africa. White South Africa at the time consisted of the Cape Province and Orange Free State (Berry 1996). Even in many environments where a given population group was said to be dominant, this apartheid spatial architecture was very much visible and its legacies continue to be felt to date.

Under apartheid, Blacks had very little say in the national politics of the country, and Black political parties were banned and had to operate underground or in exile. As the major liberation struggle movement, the African National Congress (ANC), founded in 1912, quite naturally drew the majority of its support from Blacks (in the broader sense of all non-White groups), but also enjoyed support from liberal Whites. In 1959, the political formation which will be studied in this chapter was formed. Called Progressive Party at inception,³ it was a liberal party and, perhaps unsurprisingly given the times, also a party of predominantly White, English-speaking males who were ideologically ill at ease with the policy of the party (the National Party) that had implemented apartheid (Swart 2003). In 1975, it became the Progressive Reform party. Other stops in its metamorphosis were the Progressive Federal Party (1977–1989), the Democratic Party (1989–2000), which

2. <<http://www.sahistory.org.za/article/homelands>> (4 October 2017).

3. <<http://www.sahistory.org.za/organisations/democratic-alliance-da>>

then metamorphosed from 2000 into the current Democratic Alliance following a merger (short lived) with the National Party.⁴ The National Party had now been rebranded as the New National Party. The New National Party would later be dissolved and then merged with the African National Congress.

An important point to note from this account is that while the Democratic Alliance and its predecessors may have been White dominated, the political organisation saw itself, like a number of other parties, as having been in ideological opposition to the apartheid government. There has been some contention over the manner in which this political formation has demonstrated its solidarity with the Black majority, and some of these contentions date back to the organisation's formation in 1959. There was, for instance, the Progressive Party's proposal of replacing political franchise (i.e. the right to vote) that had been premised on race, with a new form of qualified franchise based on 'civilisation' as determined by educational qualification. This new qualified franchise was perceived as race in disguise at the time and as a ploy to have upper-class non-Whites become allies of the White ruling class.⁵ In sum, the foregoing account sets the stage for the contestation of identities described in the next section.

3. Negotiating organisational identities in political discourses

Political discourse analysis (PDA) may be used as a synonym of critical discourse analysis and understood as an approach to the analysis of discourses (of whatever kind and not necessarily in the party political or governmental arena) that seeks to uncover power asymmetries (van Dijk 1997). This acceptance is roughly equivalent to what may be referred to as the politics of language (Dunmire 2012; Chilton 2004). However, PDA is also increasingly understood as the language of politics, as the text and talk of politicians in contexts where achieving legitimacy, power, influence, credibility, authority, or making other political actors lose such capital is a goal (Chilton 2004; van Dijk 1997).

Specifically, van Dijk (1997) describes political discourse in part as the text and talk of professional politicians and political parties in the public sphere, and in relevant contexts describable as political; however, he cautions against too narrow a view of what constitutes political discourse, who its relevant actors are, and what the relevant settings are. Text and talk associated with "cabinet meetings,

4. <<http://www.sahistory.org.za/dated-event/democratic-party-now-democratic-alliance-formed>> (30 September 2017).

5. <<http://www.sahistory.org.za/archive/the-progressive-party>>

parliamentary sessions, election campaigns, rallies, interviews with the media,” among others, exemplify political discourse (van Dijk 1997: 14). When politicians and political parties are doing politics, or acting politically, they are attempting to achieve political goals, “such as making or influencing political decisions that pertain to joint action, the distribution of social resources, the establishment or change of official norms, regulations and laws, and so on” (van Dijk 1997: 14).

Central to Political Discourse Analysis (PDA) as an approach is the establishment of relations between the semiotic (or meaning-making) resources in discourses and the political contexts of the deployment of such discourses. Such resources must be “functional in the adequate accomplishment of political actions in political contexts” (van Dijk 1997: 25). Such actions may include: lobbying, criticizing the opponent, election campaign, party propaganda, seeking to broaden the membership base, governing, legislating, and so on.

Much writing on political discourse analysis has focused on the sayings of political actors. This emphasis on sayings is predicated on the widespread consensus that rhetoric is a “powerful weapon for shaping political belief and action” (Dunmire 2012: 735), without language there is no politics (Chilton 2004), etc. It is through these sayings that the doings and beings of these actors are constructed. Van Dijk’s approach, for instance, requires attention to topic (what politicians say about their policies, their opponents, themselves); textual schemata (argumentation structure, warrants for argument, opening and closing); local semantics (polarization, e.g. positive self-presentation and negative other-presentation, disclaimers: “I have nothing against them...but”); lexicon (words which are used to describe us vs. them); syntax (using deictic pronouns in a biased way, especially the first person plural subject or possessive pronouns (we/our)).

It is granted that meaning-making is central to the ‘construction of beliefs about events, policies, leaders, problems, and crises that rationalize or challenge existing inequalities’ (Edelman 1988: 104 cited in Dunmire 2012: 737). However, the challenge which social semiotics (Halliday 1978; van Leeuwen 2005) poses is whether verbalisations or sayings exhaust the meaning-making resources available for constructing beliefs about political actors in specific local contexts. Social semiotics alerts us to the fact that social communication often takes place through non-verbal modes (such as images, dressing), sometimes in combination with unique but socially motivated meanings that can be read from verbal language. Social semiotics sees semiotic potential (that is, possible meanings) that can be made with given resources in specific contexts, and these resources include language, gesture, names, music, artefacts, food, architecture, dance, photographs/images, colours, public spaces, and so on.

In the related field of political communication, which is at the intersection of communication studies and political science, there is keen awareness of this broader social semiotic view of political discourse. McNair (2001: 3) notes that in

the field, the notion of political language has been understood to comprise “not only rhetoric but paralinguistic signs such as body language, and political acts such as boycotts and protests”. He then goes on to define political communication as involving “not only verbal or written statements, but also visual means of signification such as dress, make-up, hairstyle, and logo design, i.e. all those elements of communication which might be said to constitute a political ‘image’ or identity” (McNair 2001: 4).

To view a political party through such prisms is, as McNair also makes obvious, to argue that there may be socially significant semiotic potential in the names of persons with whom the party is or is not associated (e.g. as its leaders or members); in places the party goes to or does not go to for electoral campaigns; in the manner the party approaches themes; in the languages it chooses to use or not to use; in the colours it identifies with; or in what entities are depicted in its publicity materials and how these entities are positioned. With these semiotic resources, political parties can express who they are, what they are doing, how they want to be related to, and so on. These alternative semiotic resources with which political parties may be associated (alternative to their sayings) mean that political parties can be seen as entities that are embodiments of semiotic signatures which can be ‘read’. The notion of semiotic potential suggests that these resources by themselves may not be meaningful (Kress & van Leeuwen 2006). These potentials are realised within a social framework of representation and interpretation; in other words, they must be “functional in the adequate accomplishment of political actions in political contexts” (van Dijk 1997: 25).

Within South Africa, one framework of interpretation and for accomplishing political goals is transformation. The very meaning of the term is contested by politicians, and the term has been likened to “motherhood and apple pie ... assumed to be an unqualified human good” (De Vos 2010: no pagination). Beyond perhaps consensus around a minimalist acceptance as righting the wrongs of the country’s colonial and apartheid past in order to evolve a cohesive nation with a secure future, there is hardly any agreement on the shape of this political cash cow that is milked by politicians and their respective organisations. Political organisations jostle to appropriate to themselves exclusively the credentials of commitment to this cow, and to actively deny competitors any or much stake in the fortunes of this cow.

The ANC defines its mission as

the fundamental transformation of the South African economy in order to empower black people, especially Africans, (collectively as well as communities and as individuals); eliminate poverty and the extreme inequalities generated by the apartheid system; generate productive employment opportunities for our people at a living wage and ensure balanced South African economic development.

(African National Congress 1997: no pagination)

Through then party leader Helen Zille, the Democratic Alliance issued a press statement in 2010 whose title read in part “The time is ripe for real transformation” and in which it sought to distinguish its understanding of transformation from that of the ANC. In part, the statement reads as follows:

If apartheid was characterised by racial division, one party dominance, power abuse and crony enrichment, it follows that transformation strategies should be centred on transcending race and gender, opening up the democratic space, putting limits on power and broadening opportunities for all. This is what we in the Democratic Alliance mean when we talk of “transformation”.

(Zille 2010: no pagination)

Both parties interestingly premise their respective notions of transformation on apartheid era injustices. The ANC views transformation as redress for Blacks for injustices of the past in the economic, political, and social realms, and considers the DA as a White party committed to the preservation of White privilege. The DA advances a notion of transformation which, in its view, is based on constitutional democracy and the rule of the law, and considers the ANC’s transformation agenda as being about racial representativity (“numbers game”) and ideological alignment with party rather than with the ideals of a democracy. To the DA’s self-positioning as an organisation that has come through a lineage of predecessors, and transformed from a White male party to “become the most diverse party in the history of South Africa” (Selfe 2015: no pagination), a party hopeful (early in 2014) of fielding a Black female presidential candidate, the ANC interactively positions it (the DA) as renting Black faces and renting Black leaders (City Press 2014; Mottiar 2015). Even Dr Ramphela, who the DA had hoped to field as presidential candidate early in 2014 but who later changed her mind, is claimed by the ANC to have once said as follows, presumably in respect of the same DA: “a white party remains white”.⁶

Obviously, in light of the foregoing, questions arise regarding how a perspective on social semiotics might elucidate the appropriateness of the identity conferred on the DA, or of the party’s contestation of that identity and its claim to a different one. In their seminal work on identities in multilingual contexts, Pavlenko and Blackledge (2004: 19) define identities as a range of options, discursive and otherwise, “offered by a particular society in a specific time and place to which individuals and groups of individuals appeal in an attempt to self-name, to self-characterize, and to claim social spaces and social prerogatives.” As one of several types of identity which Pavlenko and Blackledge identify, negotiable identities “refer to identity options that are/can be contested and resisted” (Pavlenko & Blackledge 2004: 21).

6. <<https://mg.co.za/article/2014-02-03-ramphela-da-agang-merger-a-joint-mistake>> (16 August 2017).

This is obviously the case with the Democratic Alliance which has sought to articulate its anti-apartheid credentials and commitment to a transformed society in South Africa, all of these in the face of criticisms by its rivals. Interestingly, for our envisaged diachronic reading of the DA in the context of the current contestation with the ANC, Pavlenko and Blackledge (2004: 21) note that “negotiation does not necessarily involve two or more physical parties – it may also take place within individuals, resulting in changes in selfrepresentation”. In the case of the DA, negotiation of identity is both synchronic (involving two entities, the DA and the ANC notably) and diachronic (concerning historical metamorphoses of the organisation today referred to as the DA). Pavlenko and Blackledge’s (2004: 14) view that “languages, or rather particular discourses within them, supply the terms and other linguistic means with which identities are constructed and negotiated” is strikingly relevant to our discussion.

The question for this chapter, then, is how a semiotic perspective that transcends the sayings in a corpus of texts associated with the DA and its predecessors can elucidate aspects of the identity of the party that are perhaps lost to other forms of political discourse analysis. We are interested in the party’s claim to a transformed identity, and we operationalise transformation in the single sense of opening up to diversity, that is, transcending racial and gender lines. Admittedly, transformation is the contested term that it is because of its multifaceted nature, and which no single treatment of the topic could hope to exhaust. Given our chosen angle to political discourse analysis in which we emphasise the use of alternative datasets to the study of political organisations, the heuristics we employ for the analysis of our corpus are names and languages. As will be seen subsequently, this choice is significant in light of apartheid-era marginalisation of Blacks and females in mainstream/national politics; the consignment of African languages to enclaves called homelands; and other aspects of apartheid-era spatial architecture.

4. What’s in a name and a language?: Theorising heuristics for a corpus-based political discourse analysis

According to Hough (2013), onomastics is the study of names of all kinds such as those of people, places, landscape features, buildings, organisations, and institutions. She notes that the names people bear are not only instructive in terms of what they say about the name bearers, but also what they communicate about those in the world of the name bearers. Thus, proper names of persons or anthroponyms “can reveal, among other things, a person’s cultural heritage, nationality, and economic standing” (Block & Onwuli 2010: 464). As the managers of presidential candidate Obama realised in 2008 with their principal’s middle name ‘Hussein’,

a “name can be a blessing or a burden if one’s life chances are partly determined by the background clues conveyed by it” (Block & Onwuli 2010: 464). Diverting attention away from Hussein became a campaign strategy to sell the candidate to an electorate constructed as ill at ease with the Arab and Muslim world.

Indeed, it has been observed that in public or political discourses, “onomastic material is widely used as a means of aligning oneself with preferred ideologies or sets of preferred representations of reality” (Galasiński & Skowronek 2001: 52). In the context of a political party, the names of members (synchronically) may have as much semiotic potential for identity analysis within a transformation framework as any observed changes in the diachronic patterns of membership names. Similarly, within the same transformation framework, there is semiotic potential in the toponyms with which a party is associated (e.g. places where the party goes to for its campaign rallies). The foregoing suggests, then, that potentially socially significant meanings reside in changes observed in the anthroponyms and toponyms in a diachronic corpus of texts of a political party.

Context is one of the determinants of language choice identified by Fishman in his seminal 1965 article ‘Who speaks What language to Whom and When?’ Fishman chose to understand context in terms of such binary pairs as intimacy vs. distance, formality vs. informality, and solidarity vs. non-solidarity; in other words, how in the assessment of interlocutors the choice of a particular language would be indexical of one or the other term in each binary pair (Antia 2017). Work on spoken and written code-switching has taken further the motivations for language choice, and has equally attended to the identity constitutive or relevant dimensions of code choice.⁷ Myers-Scotton’s (1993) markedness model of code-switching invites a conceptualisation of the choice of codes in interaction in terms of assumptions of what is normative within a community. A choice is unmarked when it conforms to what would be expected, and referential meaning is central to such choice. In contrast, a choice is marked when it is unexpected and serves a strategic objective beyond the mere referential meaning. Of course, defining what is expected in particular interactional spaces may not be unproblematic. At any rate, disarticulating from any specifics of definition, the model invites us to consider what the semiotic potential (e.g. in terms of identity construction) could be of a political party’s choice of codes (expected/unexpected of it). It also invites us to consider what the semiotic potential is of changes over time in the code choices.

One widely reported function of code-switching is that of communicating friendship, desire to identify and affiliate with. The use of the listener’s stronger language in parts of the conversation may indicate deference, a desire to belong

7. We do not distinguish between code-switching and code-mixing in this discussion.

or to be accepted. Tang's (2008) study of code-switching in political campaigning discourse in Taiwan shows how "in campaign situations, by converging to people's code(s), politicians attempt to draw potential supporters' attention and momentarily to create social solidarity, particularly the integration of a society or a particular ethnic group" (Tang 2008: 99). It obviously makes sense to take a scalar view of this identification function because switching, especially in the context of speeches at political rallies, could reflect no more than tokenism (greeting, solidarity call), or be substantive in a variety of ways (e.g. duplicating multilingualism, in other words, translating the same substantive information in different codes; through fragmentary multilingualism in which only a small part of the overall message is in a different language; via complementary multilingualism in which substantive and different parts of the overall message are communicated in different languages (Kallen & Dhnnacha 2010).

From an identity perspective, which as observed by Pavlenko and Blackledge is brought into sharp relief when there are contestations, there is semiotic potential in code-switching for solidarity purposes. From the perspective of transformation of identities over time, there is equally socially significant semiotic potential in any observed changes in overall (singular) language choices as well as in the use of code-switching. With respect to the latter, Bhatt (2008) argues that

code-switching, as linguistic hybridity, is a third space where speakers (readers/writers) (re-)position themselves with regard to new community practices of speaking and writing. It is in this space that (speakers/) writers, as well as (hearers/) readers, are presumed to have the capacity to synthesize, to transform: code-switching serves as a visible marker of this transformation. (Bhatt 2008: 199)

In sum, in a diachronic corpus of texts of a political organisation, changes observed in the code choices (singular, mixed) as well as the content the choices are used to communicate can be operationalised as semiotic signatures of identity construction, which can then be read from a transformation lens (in South African acceptations).

5. Method

Following van Dijk's (1997) description of examples of political discourse, the textual data collected consisted of DA's campaign speeches, parliamentary speeches, debates and parliamentary membership lists. The latter illustrate what Hoey (2001) refers to as colony texts (exemplified further by dictionaries and telephone books), in contrast to the former which qualify as mainstream texts. Colony texts are characterised, among others, by the absence of continuous prose and by components whose meanings are autonomous rather than deriving from a sequence. Although

a majority of the person names were contained in the colony texts, they did also occur in mainstream texts.

The diachronic corpus of data for this study covers a period of 26 years, divided as follows into two subcorpora:

1. Time frame 1: 1989–2007 (Democratic Party, DP)
2. Time frame 2: 2008–2015 (Democratic Alliance, DA)

This is a relatively short time span for a political formation dating back to 1959, but data for earlier periods were unavailable. It is nonetheless worth noting that it is within the 26 year period (1989–2015) that some of the most important changes in this political formation are claimed to have taken place (Selfe 2015).

For the broader data set on which this chapter is based, an overview of the different types of material obtained and the total number of word tokens in the respective time frames is presented in Table 1.

Table 1. Overview of document types in the corpus

Types of documents	Time Frame 1 number of documents (and number of word tokens)	Time Frame 2 number of documents (and number of word tokens)
Speeches	35 (32,500 tokens)	35 (52,211 tokens)
Parliamentary membership lists	1 (339 tokens)	1 (407 tokens)
Manifestos/reports	2 (19,909 tokens)	4 (38,453 tokens)
Photographs of billboards	0 (0 tokens)	6 (63 tokens)
Photographs of campaign posters	8 (71 tokens)	5 (40 tokens)
Logos	3 (3 tokens)	1 (5 tokens)
Total number of documents and tokens in time frame	49 (52,822 tokens)	52 (91,179 tokens)

Even though this chapter excludes photographs for reasons of space, its analysis of names of people and of places as well as of language choice in texts justifies the social semiotic framing. A corpus analysis tool, AntConc (version 3.2.4w), was used for the quantitative processing of the data. The concordance functionality of AntConc was used to obtain frequency information for search items (symbols, as will be seen below). Texts that were not obtained electronically were first typed up or scanned using Optical Character Recognition (OCR) software. There was post-editing of texts obtained via OCR. Before texts were imported in plain text format into AntConc (into two folders representing each time frame), they were manually annotated with symbols. Symbols were chosen that would not be in conflict with those for wildcard searches and that would reduce the spectre of noise in hits.

- { was placed before every toponym (place name) linked to the DA or its predecessor organisations. Toponyms referred to in speeches that related to the activities of other political organisations were not taken into account.
- > was placed before names of individuals associated with the DA. First and second names were rewritten as one word. Thus, ‘Helen Zille’ became >*HelenZille*. Here as well, only names associated with the DA and its predecessor organisations (as contextually determined) were taken into account, not references to names linked to other organisations.
- On a duplicate copy of the corpus, an < was placed before names that could be identified as female. Thus, ‘Annelie Lotriet’ became <*AnnelieLotriet*. Other names were considered either male or non-classifiable.
- Since language identification in multilingual corpora is a challenge for many natural language processing applications and AntConc does not identify languages in texts, the following procedure was adopted. As English was the dominant language in the corpora, it was decided that each non-English word token in both corpora (apart from proper nouns that were processed as described above) be manually assigned to its language using the following symbols: = (Afrikaans), \$ (isiXhosa), & (Northern Sotho), and ^ (isiZulu).

In the analysis of toponyms, the following procedure was adopted. After frequency information had been obtained, the geographical dispersion of toponyms in the corpora of the two time frames was compared. Toponyms identified were looked up on political/administrative maps of South Africa to allow for a determination of any changes that may have taken place in the party’s geographical reach over time. Data such as these would be read as meaning that, from one time frame to the other, the party’s geographical penetration had changed. Any observed changes would be read from the standpoint of transformation (in the sense of opening up to, or closing out on, non-traditional communities).

With respect to the analysis of anthroponyms, it is worth noting that although a majority of the names came from colony texts, names occurring in mainstream texts were also part of the analysis. An obvious problem with using names as heuristic is that the names people bear do not necessarily give away their ethnic or racial belonging and gender for a host of idiosyncratic naming practices, name changes or choices as a result of marriage, religion, socio-political expediency, etc. Nonetheless, personal/cultural knowledge and internet searches (especially for names of party functionaries on parliamentary membership lists) informed the allocation of identified names to ethnic/racial groups, specifically, Black, Coloured, White, Indian, and Other. The frequency of occurrence of anthroponyms per ethnic/racial category was determined, and the frequencies in the two time frames compared. The same process was conducted for names related to gender. The

results are then discussed in terms of any shifts that may have taken place with respect to demographic and gender groups with which the party is associated. A shift in the direction of greater inclusion would be regarded as evidence or signature of semiotic transformation.

Finally, with respect to language choice, the frequency of each of the non-English codes was determined, then compared across both time frames. In quantitative analytical terms, the semiotic potential of increased diversity in the languages of texts in the corpora, from one time frame to the other, would be that effort was being made to identify (more) with non-traditional audiences. In terms of a qualitative analysis, semiotic transformation towards greater inclusiveness would be said to have taken place under certain conditions. An example could be that in time frame one a majority of the texts were in one language, but in time frame two there were texts that made use of more than one language, or there were texts in languages that had not previously been used. For texts with code-switching, the interpretation would be that, semiotically, transformation towards greater inclusiveness had taken place if:

- in the texts of time frame one, there were few code-switching instances compared to time frame two;
- in the texts of time frame one, the languages involved in the code-switching were only limited to English and Afrikaans, but in time frame two indigenous African languages were included; and
- in the texts of time frame one, the most common function of code-switching were the symbolic one (greeting, etc.), but in time frame two there were various substantive types of information in different languages along with their translations.

6. Results

In this section, the results of the corpus queries are presented and analysed under several sub-sections. Data on onomastics are presented and analysed in a manner that seeks to unlock the semiotic potential of the following: ethnicity-indexical names, gender-indexical names, and toponyms. Data on language choice are presented and analysed quantitatively and qualitatively.

6.1 The representation of ethnicity-indexical names

Let us consider anthroponyms from the standpoint of ethnicity/race. Figure 1 allows us to see what changes the corpus suggests have taken place in the ‘ethnicity/race of person names’ associated with this political formation over the two time frames.

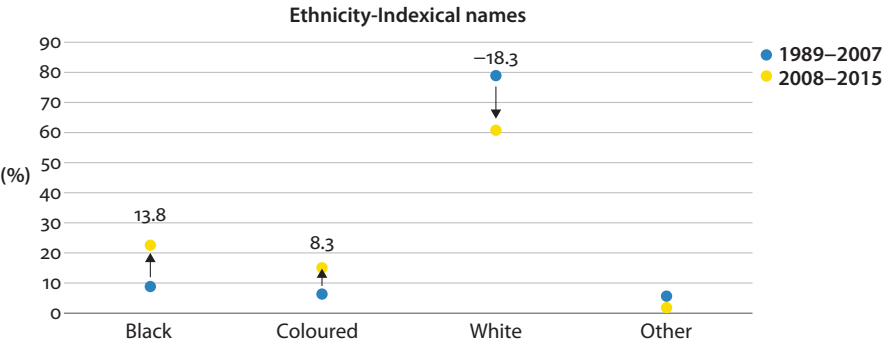


Figure 1. Ethnic distribution of anthroponyms associated with the DA and its predecessors over time

Figure 1 shows that, in time frame one, the majority (79.1%) of the party’s leadership/membership had White-indexical names. The number of Black African-indexical names stood at 8.8% and Coloured-indexical names came up to 6.6%. It may be recalled that in time frame one (1989–1994), non-Whites were only beginning to gain a measure of legitimate and de-criminalised presence in open national political spaces. The ANC, for instance, was unbanned in 1990. In some contrast, in the second time frame, the number of White-indexical names has dropped by 18.3%, from 79.1% to 60.9%. In other words, White sounding names have a negative growth percentage of –18.3%. The number of Black African-indexical names has increased by 13.8%, from 8.8% in time frame one to 22.6% in time frame two. The number of Coloured-indexical names has increased by 8.3%, from 6.6% in time frame one to 14.9% in time frame two. What these figures suggest is that, although still White dominated, the party is opening up to other ethnicities. We see this clearly in such names as *Lindiwe Mazibuko*, *Mmusi Maimane*, *Shahid Esau*, *Tsepo Mhlongo*, *Yusuf Cassim*, *SolomonMalatsi*, etc. It would be interesting to determine whether other datasets corroborate the change observed here. Let us consider the dataset on gender.

6.2 The representation of gender-indexical names

The semiotic potential of anthroponyms from a transformation perspective cannot only be seen from the standpoint of race/ethnicity. Gender is also important, considering that before the advent of democracy in South Africa in 1994, women membership of the National Parliament for instance amounted only to 2.7% (Pitamber 2016). Let us consider Figure 2.

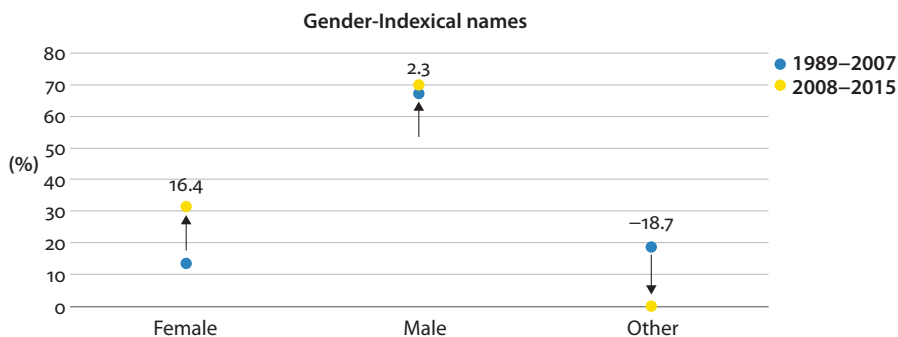


Figure 2. Gender in the DA and its predecessors in the respective time frames

Figure 2 shows that, even though the party is still male dominated, from time frame one to time frame two, the growth rate of (leading) women members in the party is more than that of men – increasing by 16.4%. We see names such as *Lindiwe Mazibuko*, *Helen Zille*, *Patricia de Lille*, *Annelie Lotriet*, *Phumzile van Damme*, *Natasha Mazzzone*, etc. At one point, the four most important offices of the party were (going to be) held by women: presidential candidate (Mamphela Ramphele), party leader and premier of the Western Cape (Helen Zille), parliamentary leader (Lindiwe Mazibuko) and mayor of Cape Town (Patricia de Lille). It can, therefore, be argued that these female names communicate an organisation that seeks to transform from a past of male-dominated leadership and membership to one which encourages and supports female leadership and membership. Such an interpretation is consistent with semiotic signatures of transformation read into ethnicity. Let us turn to toponyms in order to determine if there is corroboration of these other signatures.

6.3 The semiotic potential of toponyms

As intimated in an earlier section, under apartheid, some parts of the country were considered ‘White’ areas, whereas others were seen as ‘Black’ homelands. Figure 3 presents a political/administrative map of South Africa between 1910 and 1994 (left map) the country’s administrative map post 1994 (right map).



Figure 3. Left Map – Divisions in South Africa (1910–1994).⁸ Right Map – Divisions of South Africa (1994–present)⁹

As the left map shows, South Africa consisted of four provinces between the years 1910 and 1994. These provinces were known as the Cape Province, Natal, Orange Free State, and the Transvaal. Within the provinces of ‘old’ South Africa, there were ten ethnically based homelands, as mentioned earlier. The homelands (Venda, Lebowa, etc.) are depicted in the key in Figure 3 (left map). In the right map, the new, post-apartheid South Africa is divided into nine provinces. The census of 2011 indicated that the majority of the Black population resides in the following provinces: Eastern Cape, Kwa-Zulu Natal, Mpumalanga, Limpopo, Gauteng, and the North West (Statistics South Africa 2012). In the old South Africa represented on the left map, these provinces fell under the Cape Province, Transvaal, Natal, and minor areas of the Orange Free State. The census further indicates that other demographics (e.g. Whites, Coloureds) reside in the Western Cape, Northern Cape and the Free State – which were previously identified as the Cape Province and the Orange Free State on the left map in Figure 3 (Statistics South Africa 2012). It is against this backdrop that any remarkable changes in the distribution of toponyms in documents of this political organisation become of analytical interest.

In time frame one, of the 169 toponyms mentioned, 86 (50.8%) were in White areas, especially the Cape Province but also the Orange Free State (see Figure 4), and include *Cape Town* (21), *Parliament* (44), *Stellenbosch* (3), *Vredenburg* (3), *Kimberley* (4), *Bloemfontein* (3), *Upington* (2), etc. There is also mention of *Eastern Metropolitan* (2), *Port Elizabeth* (6), and *Mafikeng* (3), which were classified as Black areas.

8. Source: <<http://www.sahistory.org.za/article/homelands>>

9. Source: <http://www.places.co.za/maps/south_africa>

In time frame two, of the 192 toponyms, 105 (54%) were Black areas such as (or which fall under) *Limpopo*, *Port Elizabeth*, *KwaZulu Natal*, *Madibeng*, *Johannesburg*, *Soweto* and *Gauteng* – all located in provinces which had Black homelands (Transvaal and Natal on the right map in Figure 3). White areas, especially in the Cape Province of the right map, accounted for 87 (45.3%) of the toponyms.

In Figure 4, the toponyms are assigned to the post-1994 (or contemporary) provincial structure of South Africa seen earlier on the right map of Figure 3. This allows for a province by province account of shifts in toponyms across both corpora.

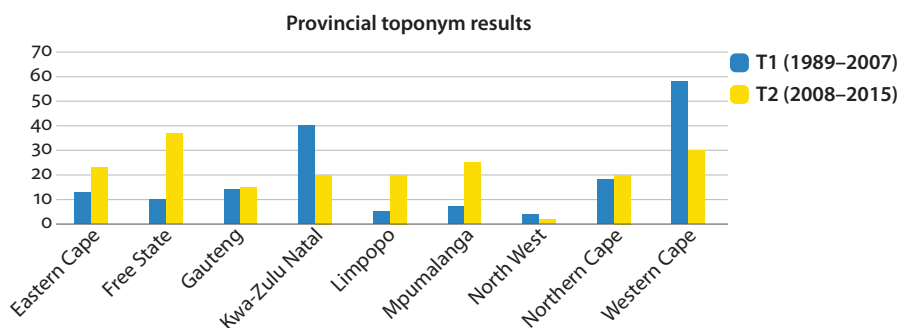


Figure 4. Depiction of frequency of toponyms in provinces (post-apartheid) for the respective time frames

Figure 4 shows that in time frame two, places in the Eastern Cape, Free State, Limpopo and Mpumalanga (all of which have majority Black African populations) seem to feature more in documents of the party than was the case in time frame one. The increase is 10% in the Eastern Cape, 15% in Limpopo, and 18% in Mpumalanga. Interestingly, in time frame two, toponyms associated with the Western Cape, the birthplace of the party, have dropped by 28% in time frame two compared to time frame one, and are even lower than the percentage in the Free State.

From the lens of transformation as greater inclusiveness, a semiotic potential of the shifts seen in the data on toponyms is that the party is reaching out to (or becoming present in) Black areas. Simultaneously, the shift may also be read as communicating some metamorphosis from a regional party to one with an increasingly national spread. So far, on the basis of race/ethnicity, gender and toponyms, the evidence has pointed to changes that are consistent with our operational definition of transformation as opening up towards diversity. The data show the party inscribing itself into the semiotics of the political landscape as a transformed/transforming organisation. In the next sub-section, we turn to another heuristic (language choice) to determine in what direction the evidence will be pointing.

6.4 The semiotic potential of language use and code-switching

As intimated earlier, English is the dominant language in the corpora of both time frames. What is of interest, therefore, is the incorporation of other languages in texts of the political organisation as well as possible shifts in the number of words per language. A first insight into the languages (besides English) occurring in both and the number of words in each is provided in Table 2.

Table 2. Language use according to counts of word tokens across corpora of both time frames. Counts for non-English words normalised per 10,000 words

Time frame one		Time frame two	
Total number of tokens 32,500		Total number of tokens 52,211	
Afrikaans	923	Afrikaans	129
isiXhosa	4.3	isiXhosa	50
		isiZulu	69
		Southern Sotho	27

Table 2 shows that from time frame one to two, there was a seven-fold drop in the relative frequency of Afrikaans word tokens (normalised per 10,000); besides an almost twelve-fold increase for isiXhosa in time frame two, in this time frame languages unattested in time frame one become visible. The semiotic potential of these changes is that the party is fashioning an identity in the construction of which Afrikaans is increasingly less important and indigenous African languages are considered of some importance. Given the limitations of many corpus analysis programs in processing multilingual texts, a manual analysis of the corpora is a lot more revealing.

Of the 35 texts in each time frame, there were texts produced: in English only; in Afrikaans only; that were English dominant but had ‘bits’ (such as greetings, and phrases) in Afrikaans and Bantu languages (such as isiXhosa, isiZulu, and Southern Sotho); that were English dominant but had ‘passages’ (paragraphs or other substantive content) in Afrikaans; and that were English dominant, but had ‘passages’ of Bantu languages. Figure 5 summarises this distribution. The x-axis refers to the number of texts that are described by the language choice options on the y-axis.

In Figure 5, it is remarkable that in time frame two there are no Afrikaans only texts. Also, we see a drop in English-Afrikaans texts. Besides the emergence of a category on English with ‘passages’ of Bantu languages in time frame two, there is also in this time frame a two-fold increase in the number of English texts with ‘bits’ of Afrikaans and Bantu languages. From the standpoint of transformation as inclusiveness, the semiotic potential of these differences is that the DA is attempting to reach new audiences (or to identify with new audiences), and is doing so

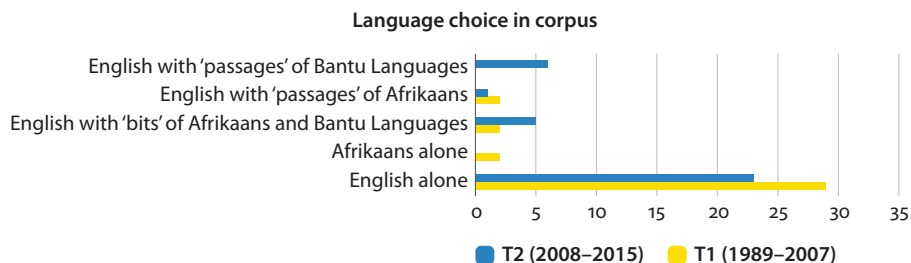


Figure 5. Categorisation of texts in corpus according to language choice

with increasingly more substantive messaging in a variety of languages. It is thus transcending some of the symbolic multilingualism in time frame one which we see in sloganeering, e.g. “Amandla” (solidarity call or rallying cry of Black South Africans, from isiXhosa and isiZulu and means ‘power’), or “beka abantu phambili” (isiXhosa for ‘Put the people first’). Compare the speech excerpts in Figures 6 and 7.

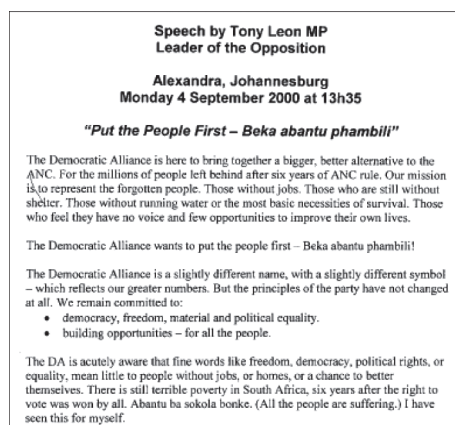


Figure 6. Bits of Bantu languages or fragmentary/token multilingualism in excerpt of English speech by Tony Leon in Alexandra (Johannesburg) in 2000 (time frame 1)

In Figure 6, party leader Tony Leon (White male) is giving a speech in an area where the majority speak two related Nguni (Bantu) languages, isiXhosa and isiZulu. Assuming Leon was hoping to address them either directly because they formed the majority of the audience at the venue, or indirectly through a differently constituted audience, it would seem their numbers (and politics is a game of numbers) justified using an unmarked code choice; in other words, using either or both mutually intelligible Nguni languages. Leon does this, but in a fragmentary way. The text of his speech translates the title “Put the people first” into isiXhosa “beka

Let me tell the ANC chairperson a truth here today, one that I am sure every culture in South Africa will agree with: running away when you should answer for what you have done wrong makes you a coward – not a leader. *ANC elekile ho thibela setjhaba ho tseba nnete. Nnete va hore mopresidente o tsotella tsohle tse tswelang vena molemo e sena setjhaba sa Afrika-Borwa.* [The ANC wanted to stop South Africa from learning the truth. The truth that our president is a man that acts for his own benefit and not in the benefit of the people of South Africa.] *Ena ha se ANC e lwanetseng tokoloho. Ena ha se ANC e lwanetseng boipuso. Ena ke ANC e bolavang le ho fedisa tokoloho e lwanetsweng ka thata ka lebaka le mehora ya bona.* [This is not the ANC that fought for liberation. This is not the ANC that fought for democracy. This is an ANC that is eroding the very freedoms that we fought for because of their own greed.] That is the great truth of the ANC under Jacob Zuma. It is no longer a party of the brave. It is a party of those who hide when they should stand up. Democrats; Right here in the Free State, the corruption of Zuma's ANC robs South Africans of opportunities. It keeps people in oppression - unable to enjoy the freedoms of our constitutional democracy. It deprives South Africans of their fundamental rights. Rights such as access to water, education and housing. *Daar is 'n water krisis oor die hele Suid-Afrika en die Vrystaat is een van die provinsies wat die ergste aeraak is. Van Bloemfontein tot Brandfort. Van Warden to hier in Kroonstad.* [There is a water crisis across South Africa and the Free State is one of the worst affected areas. From Bloemfontein to Brandfort. From Warden to here in Kroonstad.] Towns all across the Free State are withering because the ANC government cannot provide water. Last week, I

Figure 7. Duplicating and complementary multilingualism involving Afrikaans, English, Southern Sotho in excerpt of speech by Lindiwe Mazibuko in Kroonstad (Free State) in 2014 (time frame 2)

abantu phambili”, and does so elsewhere where the title is repeated. Elsewhere, he plugs in an occasional “Abantu sokola bonke” (All the people are suffering).

In some contrast to this is Figure 7, where Lindiwe Mazibuko (Black female) is giving a speech in a predominantly Afrikaans-speaking Kroonstad, but then decides to go for a marked choice of English, Afrikaans and Southern Sotho, construing both her audience and her organisation as multilingual and multicultural. Unlike the fragmentary multilingualism of Leon’s speech, her text exemplifies duplicating and complementary multilingualism, as there are passages in which substantive content is repeated in different languages or different substantive content is expressed in different languages. For the record, neither Afrikaans nor Southern Sotho is a language with which she was born. She clearly puts her reported interest in languages and qualification in political communication to good use.¹⁰

Of course, the increased multilingualism in time frame two which is communicated by the above datasets is clearly a reflection of new demographics becoming involved in leadership roles in the party. In the party’s changing language choices may be read a signature of transformation. The semiotic potential of more multilingual communications is diversity, an identity that in turn suggests the party is reaching out to new constituencies and simultaneously inviting these and other constituencies to redefine their own identities along the lines of multiculturalism and multilingualism.

10. <<http://www.sahistory.org.za/people/lindiwe-mazibuko>>

7. Discussion

This chapter was prompted by the assumption that a social semiotic approach to the analysis of political organisations could offer a novel lens for clarifying their contested (transformation) identities. Following Pavlenko and Blackledge (2004: 19), we viewed identity as semiotic resources leveraged by individuals or organisations “to self-name, to self-characterize, and to claim social spaces and social prerogatives.” We argued for the relevance of a social semiotic view in order to pave the way for analysing political identities on the basis of meaning-making resources other than their sayings around which much of the work on political discourse analysis has revolved.

Corpus evidence generated from several heuristics, *viz.* ethnicity-indexical names, gender-indexical names, toponyms and language choice/code-switching, suggested changes that were consistent with our definition of transformation as opening up to diversity, that is, transcending racial and gender lines. Such opening up to diversity was seen against the backdrop of the apartheid-era marginalisation of Blacks and females in mainstream politics; the consignment of African languages to enclaves called homelands; and other aspects of apartheid-era spatial architecture.

Context is critical to determining whether semiotic potential is indeed socially significant and corresponds to facts on the ground. There would seem to be a requirement that political discourse analysis justify its relevance by offering political scientists insights they would not have attained by other methods (van Dijk 1997). This rather tall order is made even taller when, writing on identity, Pavlenko and Blackledge (2004: 14) argue that a description of identities as “discursive constructions does not imply that they are not ‘real’ in the material world.”

What real world evidence, then, is there for the semiotic potential of the corpus analysed? Given well known dangers of causal relations, it seems safer to take a ‘coherentist’ view that claims that real world evidence is somewhat coherent with, rather than caused by, the semiotic signature of transformation in the corpus, in the sense in which we have operationalised transformation.

In the August 2016 local government (municipality) elections in South Africa, the DA had its strongest showing as yet, winning 26.90% of the popular vote. This represented an increase of over almost 3% on the preceding 2011 election results. Table 3 presents the performance of the ANC and the DA, the two leading parties, with respect to major or so-called metropolitan municipalities.

Table 3. Performance of two major parties in metropolitan municipalities in 2016 local government elections¹¹

Party	Type of control	Municipalities	Change
African National Congress	ANC Majority	3	 4
	ANC Coalition	1	 1
	<i>Total</i>	4	 3
Democratic Alliance	DA Majority	1	 0
	DA Coalition	1	 1
	DA Minority	2	 2
	<i>Total</i>	4	 3

We see from Table 3 that the total number of metropolitan municipalities in whose running the DA is involved as a consequence of its performance in the 2016 municipal elections stands at the same 4 as the ANC's. On its website, the DA brags about this victory in the following way:

On 3 August 2016, the DA contested the 2016 Local Government Elections. As a result of our performance, we now run 37 municipalities, including South Africa's legislative capital, Cape Town; the administrative capital, Tshwane, our economic centre, Johannesburg; and the city named after our democracy's founding father, Nelson Mandela Bay.¹²

This electoral victory builds on the party's electoral fortunes in the 2014 general elections at which it displaced "the Inkatha Freedom Party (IFP) as official opposition in KwaZulu-Natal and became the official opposition in the Eastern Cape, Mpumalanga, Northern Cape and Free State," while strengthening its position as the governing party in the Western Cape (Mottiar 2015: 106).

Such performance can either be seen as the outcome of a strategy, aspects of which this chapter has captured as semiotic signature/potential of transformation. Of course, the link between electoral fortunes and claimed semiotic signature of transformation can also be dismissed outright because, for some, the issue of a transformed party does not even arise. We can only hope to be able to give some insight into both perspectives rather come down on the side of one or the other.

The argument for a successful strategy or one in the making would, for some analysts, hinge on the observation that, in addition to whatever else explains the

11. Source: <https://en.wikipedia.org/wiki/South_African_municipal_elections,_2016>; upward pointing arrow indicates gain, while a downward pointing one indicates loss. A dash indicates absence of change.

12. <<https://www.da.org.za/campaign/local-government-elections-2016/>> (12 September 2017).

traction which the party is having on the electorate, its self-portrayal as a diverse institution also counts for something. In this chapter, we premised this argument around strategy on the point that names in public discourses constitute capital for “preferred representations of reality” (Galasiński & Skowronek 2001: 52) as well as on the suggestion that code choices (both marked and unmarked) in mixed-language writing/speech have the potential of repositioning speakers in relation to new audiences, allowing them to reinvent themselves (Bhatt 2008). In this regard, it is worthwhile noting that, from about 2006, part of the DA’s strategy to enhance its electoral fortunes was “to revamp its image to appeal to all South Africans, regardless of race and ethnicity. The revamp included using various patriotic symbols such as the South African flag, images of crossracial groups of people and employing all 11 national languages” (Mottiar 2015: 108). This strategy acknowledges very clearly the political capital which semiotics other than language narrowly defined constitutes for a political party.

Of course, as far as constituencies in the Black electorate are concerned, the DA might have travelled further along this road by having speeches and texts written completely in Bantu languages. Besides signalling semiotic transformation at a theoretical level, a potential consequence at a practical level might have been the enlargement of the party’s (Black) support base beyond urban areas in which it has traditionally had its strongest following (Mottiar 2015).

In dismissing as tenuous at best any links between semiotic signatures of transformation in the corpus and the electoral fortunes of the DA, some other analysts may argue that the issue of transformation does not even arise, given alleged racism within the party directed against Blacks;¹³ it may be claimed that the party was merely a beneficiary of a protest vote intended as a wake-up call to the ANC (Mottiar 2015); the semiotic potential of DA’s success being mainly in urban centres may be pointed out; it could be stressed that Bantu languages, in spite of inroads, remain marginal to the party’s communications; it could also be claimed that no evidence has been adduced that such semiotic signatures of transformation cannot be found in corpora associated with other political parties.

What this study has offered is a set of hypotheses on semiotic resources the DA is leveraging to build an identity of a transformed (that is, diverse) party. Only time will tell if these semiotic resources (anthroponyms, toponyms, and multilingualism) constitute baseless identity performances or if, in the imagination of the electorate, they are outward manifestations of an identity that is responsive to a new social order.

13. <<https://mg.co.za/article/2016-01-28-das-black-leaders-live-with-racism>> (10 September 2017); <http://www.huffingtonpost.co.za/sipho-hlongwane/it-s-about-time-the-democratic-alliance-black-caucus-spoke-out_a_21897537/> (10 September 2017).

8. Conclusion

This chapter was intended to complement perspectives in other chapters of this volume, by drawing attention to applications of corpora and the corpus analysis process outside of variational studies of English. The perspective taken was that of negotiated identities in a South African political organisation. We were particularly interested not so much in the sayings of the political organisation but in other bases on which its being is fashioned. Such an approach required a broadly social semiotic approach to a corpus-based political discourse analysis. The notion of semiotic potential was elaborated upon. Following a theorisation of identity and of the heuristic value of onomastics (specifically, ethnicity of person names, gender-indexical names, place names) and of discursive mono/multilingualism, a small diachronic corpus of the Democratic Alliance and its predecessors was built and queried using the pre-identified heuristics.

Even if to differing extents, the corpus evidence pointed consistently to a political organisation that was semiotically responsive to diversity. The corpus showed some shifts in the racial and gender constitution of the party, from one time frame to the other. Similarly, greater diversity was observed in the languages used in the party's campaign materials, and in the sorts of information various languages were being used to convey. From at best fragmentary multilingualism in the subcorpus of time frame one, we see duplicating and complementary multilingualism in the subcorpus of time frame two.

While the very nature of our subject matter (politics) means that there can be no consensus on the findings, we would hope that, beyond drawing attention to other applications of corpus and the corpus analysis process, this effort has provided hypotheses for further research and has hopefully met van Dijk's (1997: 37) criterion for successful political discourse analysis (PDA): "PDA will only be accepted by political scientists if it has something to offer, preferably something political scientists would not otherwise (get to) know – at least not as well – through other methods, such as polls, participant observation or content analysis."

For now, though, we of course recognise the possible limitations of a rather small corpus as well as of descriptive, rather than inferential, statistics with respect to making claims of semiotic potentials that are socially significant.

References

- African National Congress. 1997. 50th National Conference: Resolutions – Economic Transformation. <<http://www.anc.org.za/content/50th-national-conference-resolutions-economic-transformation>> (25 June 2017).
- Ahmad, Khurshid (ed.). 2011. *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*. Dordrecht: Springer.
- Ahmad, Khurshid. 1996. A terminology dynamic and the growth of knowledge: A case study in nuclear physics and in the philosophy of science. In *TKE '96: Terminology and Knowledge Engineering*, Christian Galinski & Klaus-Dirk Schmitz (eds), 1–11. Frankfurt: Indeks.
- Aijmer, Karin. 2009. *Corpora and Language Teaching* [Studies in Corpus Linguistics 33]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.33>
- Akosu, Nicholas & Selamat, Ali. 2013. Enhancing the effectiveness of the spelling checker approach for language identification. In *Recent Developments in Computational Collective Intelligence*, Amelia Badica, Bogdan Trawinski & Ngoc Thanh Nguyen (eds), 155–166. Cham: Springer.
- Antia, Bassey E. 2017. Shh, hushed multilingualism! Accounting for the discreet genre of translanguage siding in lecture halls at a South African university. *International Journal of the Sociology of Language* 243, 183–198.
- Antia, Bassey E. & Oliver Razum. 2012. HIV/AIDS messaging in Germany and Nigeria: A corpus linguistics study. *Stellenbosch Papers in Linguistics PLUS* 41: 1–23.
- Antia, Bassey E. 2000. *Terminology and Language Planning: An Alternative Framework of Discourse and Practice* [Terminology and Lexicography Research and Practice 2]. Amsterdam: John Benjamins. <https://doi.org/10.1075/tlrp.2>
- Antia, Bassey E. 2009. Corpora and technology: Opportunities and strategic options for Nigerian Higher Education. In *Language Development Beyond Borders: A Festschrift in Honour of Professor Dr. Dafydd Gibbon*, Eno-Abasi Urua, Moses Ekpenyong & Firmin Ahoua (eds), 79–107. Uyo: Emsel Groups.
- Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.
- Benson, Morton, Benson, Evelyne & Ilson, Robert F. 1986. *Lexicographic Description of English*. Amsterdam: John Benjamins. <https://doi.org/10.1075/slcs.14>
- Berry, Bruce. 1996. *Divisions of South Africa (1910–1994)*. <http://www.crwflags.com/fotw/flags/za_old-.html> (20 August 2017).
- Bhatt, Rakesh M. 2008. In other words: Language mixing, identity representations, and third space. *Journal of Sociolinguistics* 12(2): 177–200. <https://doi.org/10.1111/j.1467-9841.2008.00363.x>
- Block, Jr., Ray & Onwunli, Chinonye. 2010. Managing monikers: The role of name presentation in the 2008 presidential election. *Presidential Studies Quarterly* 40(3): 464–481. <https://doi.org/10.1111/j.1741-5705.2010.03781.x>
- Bosch, Sonja E. & Eiselen, Roald. 2005. The effectiveness of morphological rules for an isiZulu spelling checker. *South African Journal of African Languages* 25(1): 25–36. <https://doi.org/10.1080/02572117.2005.10587246>
- Chege, Kamau, Wagacha, Peter, de Pauw, Guy, Muchemi, Lawrence, Ng'ang'a, Wanjiku, Ngure, Kenneth & Mutiga, Jayne. 2010. Developing an open source spellchecker for Gikũyũ. *Proceedings of the Second Workshop on African Language Technology – AfLaT 2010*. <http://aflat.org/files/chege_0.pdf> (24 August 2017).

- Chilton, Paul. 2004. *Analysing Political Discourse: Theory and Practice*. London: Routledge.
<https://doi.org/10.4324/9780203561218>
- City Press. 2014. ANC not concerned about DA's 'rent a black' – Mantashe. <<https://www.news24.com/Archives/City-Press/ANC-not-concerned-about-DAs-rent-a-black-Mantashe-20150429>> (20 June 2017).
- De Pauw, Guy, Maajabu, Naomi & Waiganjo Wagacha, Peter. 2010. A knowledge-light approach to Luo machine translation and part-of-speech tagging. *Proceedings of the Second Workshop on African Language Technology – AfLaT 2010*. <<https://www.clips.uantwerpen.be/sites/default/files/depauw.pdf>> (24 August 2017).
- De Pauw, Guy, de Schryver, Gilles-Maurice & van de Loo, Janneke. 2012. Resource-light Bantu part-of-speech tagging. *Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*. <<https://www.clips.uantwerpen.be/sites/default/files/depauwetel.pdf>> (24 August 2017).
- de Vos, Pierre. 2010. What do we talk about when we talk about “transformation”? <<https://constitutionallyspeaking.co.za/what-do-we-talk-about-when-we-talk-about-transformation-2/>> (24 June 2017).
- Dunmire, Patricia L. 2012. Political discourse analysis: Exploring the language of politics and the politics of language. *Language and Linguistics Compass* 6(11): 735–751.
<https://doi.org/10.1002/lnc3.365>
- Edelman, Murray. 1988. *Constructing the political spectacle*. Chicago: University of Chicago Press.
- Ekpenyong, Moses, Urua, Eno-Abasi, Wats, Oliver, King, Simon & Yamagish, Junichi. 2014. Statistical parametrical speech synthesis for Ibibio. *Speech Communication* 56: 243–251.
<https://doi.org/10.1016/j.specom.2013.02.003>
- Fairclough, Isabela & Fairclough, Norman. 2012. *Political Discourse Analysis: A Method for Advanced Students*. London: Routledge.
- Fishman, Joshua A. 1965. Who speaks what language to whom and when? *La Linguistique* 1(2): 67–88.
- Galasiński, Dariusz & Skowronek, Katarzyna. 2001. Naming the nation: A critical analysis of names in Polish political discourse. *Political Communication* 18(1): 51–66.
<https://doi.org/10.1080/10584600150217659>
- Halliday, Micheal. 1978. *Language as a Social Semiotic*. London: Edward Arnold.
- Hoey, Michael. 2001. *Textual Interaction: An Introduction to Written Discourse Analysis*. London: Routledge.
- Hough, Carole. 2014. Onomastics. <<http://onomastics.co.uk>> (15 March 2017).
- Kallen, Jeffrey L. & Ní Dhnna, Ester E. N. 2010. Language and inter-language in urban Irish and Japanese linguistic landscapes. In *Linguistic Landscape in the City*, Elana Shohamy, Eliezer Ben-Rafael & Monica Barni (eds), 19–26. Bristol: Multilingual Matters.
- Kress, Gunther & van Leeuwen, Theo. 2006. *Reading Images: The Grammar of Visual Design*. London: Routledge. <https://doi.org/10.4324/9780203619728>
- McNair, Brian. 2001. *An Introduction to Political Communication*. 5th edition. London: Routledge.
- Mottiar, Shauna. 2015. The Democratic Alliance and the role of opposition parties in South Africa. *Journal of African Elections (Special Issue on South Africa's 2014 Elections)* 14(1): 106–123.
- Myers-Scotton, Carole. 1993. *Social Motivations for Codeswitching. Evidence from Africa*. Oxford: Clarendon Press.

- Pavelenko, Aneta & Blackledge, Adrian. 2004. *Negotiations of Identities in Multilingual Contexts*. Clevedon: Multilingual Matters.
- Pitamber, Priya. 2016. South Africa's women in politics. <<https://www.brandsouthafrica.com/governance/developmentnews/south-africa-s-women-in-politics>> (15 March 2017).
- Selfe, James. 2015. *A Tribute to Helen Zille*. <<http://www.politicsweb.co.za/politics/a-tribute-to-helen-zille--james-selfe>> (20 December 2017).
- Sinclair, John McHardy (ed.). 2004. *How to Use Corpora in Language Teaching* [Studies in Corpus Linguistics 12]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.12>
- Statistics South Africa, 2012. *Census in Brief*. <<http://www.statssa.gov.za>> (12 June 2017).
- Swart, Sandra. 2003. An irritating pebble in Krugers shoe: Eugene Marais and Land en Volk in the ZAR 1891–1896. *Historia* 48(2), 66–87.
- Tang, Chihsia. 2008. Code switching in the political campaigning discourse: The case of 2008 presidential election debates in Taiwan. *UST Working Papers in Linguistics* 4: 99–116.
- van Dijk, Theo. 1997. What is political discourse analysis? In *Political Linguistics*, Jan Blommaert & Chris Bulcaen (eds). Special issue of *Belgian Journal of Linguistics* 11: 11–52.
- van Leeuwen, Theo. 2005. *Introducing Social Semiotics*. New York NY: Routledge.
- Zille, Helen. 2010. Sa today: The time is ripe for real transformation. <<http://www.da.org.za/anc/transformation>> (15 December 2017).

Index

A

academic writing 3, 39, 40, 44,
129, 215, 216, 267, 309, 331,
332, 336, 359, 369
see also student essays

acceptability 266, 311, 317, 318,
319, 320, 325
see also variational
acceptability

adjective phrase 248

adverb phrase 248

African languages (language
groups)

Acholi 308

Afrikaans 53, 99, 100, 107,
233, 234, 241, 250, 251, 390

Bantu 108, 232, 233, 296,
308, 391

isiXhosa 44, 390, 391

isiZulu 44, 390, 391

Khoisan 233

Kiswahili 308

Luganda 305, 308

Nilotic 296

Oshiwambo 233, 251

Setswana 9, 99, 234

Southern Sotho 390

annotation (*see also* tagging)
18, 19, 48, 64, 77, 80, 96, 148,
149, 383
see also Platform for
Annotated Corpora in
XML (Pacx)
pragmatic annotation
semantic annotation
syntactic annotation
(*see also* parsing)
time-aligned annotation

AntConc 10, 26, 46, 84, 106,
170, 210, 238, 289, 362, 383

B

balance 14, 15, 16, 42, 75

Bank of English 9, 14, 24, 91

Black South African English
108, 124, 186, 189, 201, 266

Borrowings 135, 240, 260, 261,
262, 265, 270, 276, 287, 296,
306ff
see also loanwords

British National Corpus (BNC)
19, 23f, 273

C

Cameroon(ian) English 39, 43,
45, 167, 206
see also Corpus of
Cameroonian English
(CCE)
Cameroon Pidgin English
Corpus of Cameroon MA
theses
Corpus of (Spoken)
Cameroon Pidgin English
(CPE)
Corpus of Nigeria and
Cameroon English
Learner Language
(CONACELL)
Cameroon Pidgin English 144
CLAWS 19, 48, 49, 56, 64,
74, 89
codeswitching 148, 158, 169,
170, 381, 382, 385, 390, 393
coherence 334, 347, 349
cohesion 336, 337
cohesive devices 333, 334, 336
collocate 18, 27, 52 (fnote), 53
(fnote), 273, 301, 303, 315, 347
comparable corpus 24f, 120,
167, 266, 357
concordance 87, 337
concordancer 63, 89

concordance lines 9, 17, 18, 27,
30, 34, 46, 47, 87, 222
see also keyword in context
(KWIC)

conjunction 307, 331, 332, 333,
334, 335, 341

corpus (corpora) 7, 8, 11,
14, 15, 16, 23, 37, 34, 38,
167, 373f
see also comparable corpus
general corpus (*see also*
reference corpus)
parallel corpus
saturation corpus
specialised corpus

corpus-driven versus corpus-
based 11f, 149, 150, 153

Corpus of Cameroon MA
theses 356, 360

Corpus of Cameroonian English
(CCE) 43, 143, 167

Corpus of English(es) Spoken
by Black Namibians post
Independence (ESBNaPI)
101, 105

Corpus of Namibian Online
Newspapers (CNamON)
232, 235, 236, 238, 239

Corpus of Nigeria and
Cameroon English Learner
Language (CONACELL)
22, 35, 55, 74

Corpus of (Spoken) Cameroon
Pidgin English (CPE) 143

Corpus of Xhosa-English
189, 193

D

Data Driven Learning 73

diachronic 25, 26, 29, 124,
382, 383

ditransitive 155

- dynamic model 124, 260f, 266
consonant cluster simplification 198, 202
- E**
East African English 186, 202
epistemic modality 172
- F**
frequency 17, 22, 106, 107, 110, 150ff, 167
see also normalised frequency
raw frequency
- G**
general corpus (*see also* reference corpus) 10, 18, 23f, 273, 310, 339
Ghanaianism 262, 263, 269, 271, 272, 276
Ghanaian English 58, 206, 262, 263, 273, 274, 285
see also
Historical Corpus of English in Ghana (HiCE)
International corpus of English (ICE) Ghana
Global Web-Based English (GloWbE) Corpus 43, 45, 53, 54, 123, 239, 294, 296, 312, 364
- H**
Historical Corpus of English in Ghana (HiCE) 120, 264
Historical native White and Black South African English 45
- I**
Identity 100, 120, 145, 374, 379, 390
Indian South African English 108
International corpus of English (ICE) 27, 38, 40, 45, 123, 146
International corpus of English (ICE) Ghana 264
- International corpus of English (ICE) Nigeria 41, 189, 236, 337
International corpus of English (ICE) Uganda 294, 305
International corpus of learner English 25, 27, 71
- K**
Kenyan English 58, 293, 361
keyword in context (KWIC) 46, 55, 87, 89
- L**
L1 interference 98, 368
L1 transfer 134, 239
learner corpus 25, 71, 72, 74
lexical verb 21, 90, 153
lexical expansion in Ghanaian English 260, 268, 287
lexis 12, 13, 240, 244, 254, 260
lexis-grammar interface 136, 262, 265
loanwords 51, 240, 241, 243
log-likelihood 337, 340
- M**
metadata 60, 77, 78, 81, 82, 156
modal verb 170, 208
multi-method approach 293, 311, 321
- N**
Namibian English (NamE) 97, 232, 234, 239, 251,
see also Corpus of English(es) Spoken by Black Namibians post Independence (ESBNaPI)
Corpus of Namibian Online Newspapers (CNamON)
Namibianisms 232, 239 (footnote)
Namlish 239 (footnote), 244, 255
Nativisation 106, 120, 122, 123, 125, 127, 180, 297 (footnote), 312, 321
News on the Web corpus (NOW) 43, 238, 239, 252
- Nigerian English 51, 52, 59, 80, 196, 197, 201, 206
see also Corpus of Nigeria and Cameroon English Learner Language (CONACELL)
International corpus of English (ICE) Nigeria /h/-deletion
Nigerian Learner English Corpus (NLEC) 45, 333
non-native standards 358
normalised frequency 16, 17, 53
- P**
parallel corpus 25, 44, 165 (footnote)
parsing 20, 49
person names 52, 147, 240, 265, 383, 386
phonological reduction 189, 192
place name 147, 240, 241, 261, 265, 384
Platform for Annotated Corpora in XML (Pacx) 41, 77, 133, 134
political discourse analysis 377, 380
pragmatic annotation 223
preposition 150, 248, 356, 361, 362, 364, 366
prepositional verb 246, 254, 265, 274, 279
- R**
raw frequency 17, 178
register 7, 9, 28, 38
register variation 9, 28
representativeness 16, 23, 38, 74, 76, 129
root modality 173
- S**
saturation corpus 16
semantic annotation 41, 49, 50, 266
semantic change 302ff, 310, 315, 316, 321
semiotic potential 377, 378
serial verb constructions 151, 152, 153, 156

- SketchEngine 9, 10, 17
 Skylight 10, 27
 South African English 44, 51, 54, 232, 254
 see also Black South African English
 corpus of Xhosa-English
 Historical native White and Black
 South African English
 Indian South African English
 White South African English
 specialised corpus 10, 24, 27, 44, 356
 stance markers 207, 208, 211, 212, 215, 218
 student essays 14, 17, 44, 210, 215
- syntactic annotation (*see also* parsing) 41, 49
- T**
 tagging 19, 21, 22, 23, 41, 49, 48, 64, 148
 Tanzanian English 41, 43, 48, 296, 299
 tense 55, 57, 58, 89, 90, 194, 196, 197, 199, 201
 time-aligned annotation 42
- U**
 Ugandan English (UgE) 293, 296, 305
 see International corpus of English (ICE) Uganda
 Web-UG
- V**
 variational acceptability 358
 variational pragmatics 209
- W**
 Web-UG 312, 313
 West African English(es) 206, 211, 214, 215, 218
 White South African English 107
 Word-formation 135, 240, 244, 260, 262, 265, 271, 303
 Wordsmith Tools 10, 26, 46
 World Englishes ix, 136, 188, 240, 273, 293
- Z**
 Zimbabwean English 42, 45

Corpus linguistics has become one of the most widely used methodologies across the different linguistic subdisciplines; especially the study of world-wide varieties of English uses corpus-based investigations as one of the chief methodologies. This volume comprises descriptions of the many new corpus initiatives both within and outside Africa that aim to compile various corpora of African Englishes. Moreover, it contains cutting-edge corpus-based research on African Englishes and the use of corpora in pedagogic contexts within African institutions. This volume thus serves both as a practical introduction to corpus compilation (Part I of the book), corpus-based research (Part II) and the application of corpora in language teaching (Part III), and is intended both for those researchers not yet familiar with corpus linguistics and as a reference work for all international researchers investigating the linguistic properties of African Englishes.

ISBN 978 90 272 0219 2



9 789027 202192

JOHN BENJAMINS PUBLISHING COMPANY