TECHNISCHE UNIVERSITÄT CHEMNITZ
PHILOSOPHISCHE FAKULTÄT
Anglistik/ Amerikanistik
Professur Englische Sprachwissenschaft

MAGISTERARBEIT

"Typicality of Lexical Items of Kenyan an Tanzanian English"

Betreuer: Prof. Dr. Josef Schmied

Eingereicht am 19.09.2008

Anne-Katrin Sacher
Berthelsdorfer Straße 30
09661 Hainichen
asac@hrz.tu-chemnitz.de
13. Fachsemester
Fachrichtung Anglistik/Amerikanistik

Acknowledgements

I would like to thank a number of people, who guided me through the process of accomplishing this thesis successfully.

First, I want to thank my supervisor Prof. Dr. Josef Schmied from the Chemnitz University of Technology, for his kind support, as well as giving me this opportunity to write this thesis. Furthermore, I want to thank Dr. Christoph Haase, as well from the Chemnitz University of Technology, for giving me always good advice for the improvement of this thesis.

Thanks go also to my friend Christin Kretzschmar, who offered to read through and correct my thesis again.

Last but not least, I would like to thank my family for always supporting and encouraging me. As well as my closest friends, who were always willing to listen to me whenever I had problems with this study.

Contents

List of Tables

List of Figures

List of Abbreviations

ACE	Australian Corpus of English
COLT	Corpus of London Teenage Language
EA	East Africa
Freq.	Frequency
ICAME	International Computer Archive of Modern and Medieval English
ICE	International Corpus of English
ICE-EA	East African Part of the International Corpus of English
LLC	London-Lund Corpus
LOB	Lancaster-Oslo/Bergen Corpus
SEC	Spoken English Corpus
SEU	Survey of English Usage Corpus
WC	Wellington Corpus of Written New Zealand English
WSC	Wellington Corpus of Spoken New Zealand English

1 Introduction

The sociolinguistic situation in Kenya and Tanzania is quite problematic and complex. More than 40 languages in Kenya and even 120 languages in Tanzania co-occur with English. The two official languages are English and Kiswahili. In Kenya, English is the primary language of governmental institutions, in education and in the media. Furthermore, it is the language of the élite. Kiswahili has its origin in the Bantu languages, and it is referred to as the "National Language", which is spoken in the rural regions and used by governmental officials in public. It is also used in business relations and as the communication tool with Central Africa. Both languages are exerted on official governmental documents, e.g. passports and registration forms, and on labels of all goods and products.

Nevertheless, English as the lingua franca should be taught and learned, so that communication outside of Africa is feasible and that foreign knowledge can be implemented. Due to the fact that Kenya has a heterogeneous status of languages and a number of Kenyans grow up bilingual, English is seen as the lingua franca, which brings together the different parts of the country. All in all, there is an imbalance between English, as the former colonial language, and the indigenous languages, e.g. teaching material as well as teachers are not available for all indigenous languages. Since Kiswahili language classes were established in schools and universities in 1985, the status of the language increased until today, and more and more books as well as magazines are printed.

But English is still the status language, followed by Kiswahili, and the mother tongues play only a minor role. It is very difficult for a country, in which more than 40 languages are spoken, to obtain equality and to implement all linguistic rights. Furthermore, the Kenyan élite does not respect the indigenous languages, because they are seen as non-scientific languages. There is a need for clear governmental instruction and financial aid to develop and secure the language varieties.

The linguistic situation in Tanzania is similar to the one in Kenya. In Tanzania more than 120 languages are spoken, but the major languages are English and Kiswahili as well. The Tanzanian government is a role model for implementing African languages. In primary schools it is compulsory to use an indigenous language, and therefore English has become the language of instruction only in higher education. Educational policies

support the studying of African languages and benefit the public. In colonial times, language was a symbol of status, wealth and separation. Nowadays, many people in African countries look at the "western life", and the entrance to this life is the English language. Moreover, there is a lack of confidence of using indigenous languages, but still it is spoken in public life and institutions, although English is usually the official language.

As can be seen, the linguistic situation in both countries is diverse. The English language differs immensely from the English varieties, spoken in other countries across the world. As a result, the English language itself is often mingled with the indigenous languages, which makes Kenyan and Tanzanian English so unique.

The aim of this thesis is to detect this uniqueness, if possible at all. Uniqueness, in this sense, refers to specific lexical items, which are typical for the English, used in Kenya and Tanzania. In other words, this study tries to examine the lexicon, in order to retrieve appropriate linguistic data, which reflects in some way the culture or politics of Kenya and Tanzania. Retrieving linguistic data can only be accomplished with the help of corpora, which enable us to study the language under consideration effectively. The East African Part of the International Corpus of English serves as the basic corpus for this study. The exploitation of corpora can only be successfully completed with the help of analysis tools, which enable us to retrieve exactly the data we are about to examine. The received linguistic data can then be analysed according to the desires of the researcher. The desires for this study are: firstly, to obtain appropriate results which can be said to be typical for Kenyan and Tanzanian English, and secondly, finding out similarities and differences of the received Kenyan and Tanzanian keywords, if there are any.

The study can only be conducted if one is familiar with the theoretical background, surrounding this complex field of corpus research. Therefore, the paper will start with a section about the scope of corpus linguistics, which tries to give an insight of what corpus linguistics is about, and will further hint to the limitations of this linguistic area. But knowledge about corpus linguistics is not sufficient to understand the problems I was faced with during this study. Thus, I will also present the different corpora, which were relevant for this research, as well as the ICE-EA, the corpus under consideration. The last theoretical chapter will deal with the notion of typicality and keywords, which have been hardly described in linguistics until now, due to the reason that it is relatively young research field, in comparison to others. Furthermore, I decided to include a

chapter about the procedures used when working with a corpus analysis tool, as a demonstration to other people, who might be interested in information retrieval as well.

Whether the results I obtained are appropriate or not, and whether they can be linked to the culture, way of life, politics or any other field of Kenya and Tanzania, and whether similarities or differences between these two varieties could be detected, will be resolved in the paper now.

2 An Outline of Corpus Linguistics

The concern of this chapter is to give an overview of what corpus linguistics is about and how a corpus is defined. It, furthermore, deals with the crucial role of computers in corpus linguistics, i.e. how they enhanced corpus studies. It presents the various types of corpora as well as the most important English corpora, which are of great interest either in corpus linguistics or for this underlying study. The chapter will be concluded by a section about the limitations and problems corpus linguists have to cope with.

2.1 What is Corpus Linguistics about?

The interest in corpus linguistics has increased tremendously since the middle of the 20th century, when the advent of the computer began to facilitate the handling of the vast amount of data. Many linguists, such as Leech, claim that corpus linguistics is not a subject, which underlies studies, as syntax or semantics, it is rather a methodology for analysing the language according to the various subfields of linguistics (cf. Ooi 1998: 35). But how can *corpus linguistics* be defined? The most appropriate definition, in my opinion, is the one given by Ooi, who states that "corpus linguistics is a field of study that involves the study of language on the basis of textual or acoustic corpora, almost always involving the computer in some phase of storage, processing, and analysis of this data" (Ooi 1998: 35)[1]. This definition already implies the notion of *corpora*, which therefore needs to be further explained. Kennedy describes the *corpus* as "a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description" (Kennedy 1998: 1). He points out the differentiation between written and spoken corpora, which will be closer analysed in the section of corpus typology, chapter 2.3. I want to concentrate on the second part of the definition that corpus linguistics is the basis for analysis and description.

Corpora are intended to give insights how languages work and how we use them. In analysing them, linguists can identify patterns of specific aspects of language. Corpus research can help the analyst to describe language functions concerning e.g. pragmatics, grammar or discourse patterns. These new insights might be very helpful for various

[1] The original definition given by Ooi dates from his 1994 article: Ooi, V. (1994). "Corpus Linguistics". *SAAL Quarterly (Journal of the Singapore Association for Applied Linguistics)*, 28: 2-4.

applications, of which language learning and teaching is probably the most significant one. Syntax and grammar had been the original focuses of corpus studies, but were replaced by lexis in the late 20th century (cf. Scott and Tribble 2006: 4). Research in lexis is concerned with analysing the meaning of a text or corpus, of what they are about. But corpus studies are not only interested in retrieving linguistic aspects of one language, they also deal with comparing corpora of different languages or varieties of a language. Thus, the underlying study falls clearly into these domains, trying to investigate lexical items, keywords, of Kenyan and Tanzanian English. Nevertheless, the definition given by Kennedy seems to be more general. What we need, is a more specific definition of *corpus*, as the one given by Francis, who was one of the compilers of the important Brown Corpus, which I will present in chapter 2.4. He defines corpus as "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis" (Partington 1998: 2). The second definition reveals a further issue about the aims of corpus linguistics, i.e. the "representativeness" of a corpus.

This representativeness is difficult to achieve. How can a corpus be representative to a language, if it only contains about one million words? Is representativeness therefore dependent on the size of a corpus? According to Biber, it is more a question of "well-balancing" the corpus or text. He defines representativeness as "the extent to which a sample [text] includes the full range of variability in a population" (Ooi 1998: 53). Corpora should therefore be compiled in such a way that they cover as many as possible varying language categories. Nonetheless, many linguists claim that "a corpus, no matter how large and varied, is only ever representative of itself" (Partington 1998: 146) and that they "will only ever provide a glimpse of what has been said" (Teubert and Čermáková: 2004: 157)[2]. The compiling of corpora is only one main aspect of corpus linguistics. Kennedy claims that there are four areas in which corpus researchers are involved.

The first one, as I already mentioned, is the compilation or design of corpora which includes assembling relevant texts, in order to receive the desired representativeness, as well as the annotation of texts to make corpora accessible to later research. The group of researchers who develop analysing tools falls into the second activity area of corpus

[2] Čermáková, A., M. Halliday, W. Teubert, and C. Yallop, (2004). *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.

linguistics. These researchers are primarily concerned with processing tools for the exploration of corpora. In this case, corpus linguistics enters partly the domain of computational linguistics. Laurence Anthony of the Waseda University in Japan is particularly involved in this research area, developing his analysis tool called *AntConc* in 2002, which I used for the investigation of the different corpora in my study. The third group of researchers is concerned with descriptive linguistics. They exploit the compiled corpora in order to obtain linguistic patterns of the language in use. As Kennedy puts it: "descriptive linguistics is concerned not only with what is said or written, where, when and by whom, but how often particular forms are used" (Kennedy 1998: 9). The last research area deals with the application of the new gained descriptive information. This linguistic knowledge could be applied in language learning and teaching, in dictionary making or in several other fields involving language processing. The division given by Kennedy implies another issue already explained by Leech, that the focus of corpus linguistics is on "performance rather than competence, and on observation of language in use leading to theory rather than vice versa" (Kennedy 1998: 7). Corpus linguists perform by compiling corpora and developing analysis tools, and then observe corpora to draw conclusions about language functions.

The terms *computer* and *analysis tools* had been mentioned several times so far, which gives the impression that they play a crucial role in corpus linguistics. This crucial role will be further examined in the next section.

2.2 Computers and Corpus Linguistics

Through the advent of the computer, corpus linguistics experienced an enormous boost. The new developed hardware enabled linguists to store vast amounts of data in electronic form. Until then, assembling texts and counting as well as transcribing their words was not only time consuming, it was also susceptible to errors. Since the middle of the 20th century corpus linguistics is inevitably linked to the computer. Therefore Leech claims that the right expression should be "computer corpus linguistics" (cf. Ooi 1998: 34). The computer was a great progress in storing and classifying the data, but it could not analyse the corpora by itself. New software tools had to be developed and programmed to achieve this purpose. At the beginning, corpus research was a slow process since computer and software technology was still in its development. Thus, corpus linguists examined only, what we would call nowadays, small corpora, containing about one million words. But with the improvement of technology, research process became faster

and faster, and subsequently, larger corpora were increasingly subject to studies. New technologies affected not only the handling of huge material, but also the reliability of the achieved data. Accurate measuring is an important issue for linguists to trust the results they received. But how do the software tools work? Scott offers a rather trivial explanation: "modern corpus software […] is capable of ploughing through vast quantities of text in a relatively short time, possibly accessing it remotely, and reducing it to a set of potential patterns" (Scott and Tribble 2006: 5). Especially the last part of the definition points out the task of software tools. A corpus consisting of millions of words, scattered over thousands of texts about hundreds of topics is reduced to a simple wordlist either in alphabetical or frequency order. The investigation of these lists is the real corpus analysis, at least in terms of lexical analysis. The word lists enable researchers to identify frequencies, keywords and concordances, just to name a few. This specific procedure will be explained in more detail when analysing the East African corpus.

Today, nobody can imagine corpus linguistics without computers anymore. The computer has become a great tool of coping with corpora. And corpora have been designed with the help of computers to fit the researchers' purposes, which are manifold. Therefore, the next section will cover the broad range of different corpus types.

2.3 Corpus Typology

Considering corpus typology, I will follow the outline given by Kennedy in his book *An Introduction to Corpus Linguistics*. Corpora can differ for various reasons or purposes for which they were compiled. Firstly, I want to mention the distinction between *corpus* and *archive* (Leech and Fligelstone refer to a *collection* rather than an *archive*)[3]. A *corpus* is systematically planned and structured whereas an *archive* is often opportunistically collected, not structured and new material is added ongoing.

This will be only a brief and general overview of the different corpora types, in order to determine, to which types the corpora of study belong.

General vs. specialized corpora: A *general corpus* is compiled for the usage of any linguistic research and contains various texts about different language categories. In contrast to a *specialized corpus* which is compiled for specific research projects such as the English used in child or business language.

[3] cf. Leech and Fligelstone 1992: 120.

Written vs. spoken corpora: A *written corpus* is by far the most common type. This is due to the fact that written texts are easier to scan and process with the computer. Spoken corpora, on the other hand, are difficult to compile. Although spoken language is much more used than written language in our daily use, the transcription of spoken utterances takes a lot of time and is expensive.

Sample-text vs. full-text corpora: A *sample-text corpus* is compiled to be representtative to a language or any other subfield of language as genres, dialects and so on. The complete works of e.g. Charles Dickens would be a *full-text corpus*, which means that it contains the complete texts of the author.

Synchronic vs. diachronic corpora: An example of a *synchronic corpus* is the Brown Corpus, which was compiled of written texts of American English only published 1961. If texts are sampled over a period of time in order to observe language change, we speak of a *diachronic corpus*.

Another corpus type which Kennedy mentions is the *monitor* or *dynamic corpus*. This corpus is basically open-ended, because new texts can be added all the time, and it is normally not structured, which tends to go into the direction of an *archive*.

Teubert and Čermáková mention a further type, which is not in the outline given by Kennedy. This is the *parallel* or *translation corpus*, which consists of original texts of one language and their equivalent translations into other languages. (cf. Teubert and Čermáková: 2004: 122)[4]

Kennedy throws up the term "balanced" in his outline although it is not explicitly classified by him. This is another distinction of corpora, which I consider as vital for corpus typology.

Balanced vs. imbalanced corpora: A *balanced corpus* should consist of written and spoken material. It should as well contain various different language categories to be representative. *General corpora* are intended to be balanced. A corpus which consists only of spoken utterances or only of written texts or has a severe unequal distribution of these two is considered an *imbalanced corpus*. A *monitor corpus* would clearly fall into this category.

Corpus typology is crucial in order to classify the different corpora which are in existence up to now, since many of them were created for varying purposes. The corpus

[4] Čermáková, A., M. Halliday, W. Teubert, and C. Yallop, (2004). *Lexicology and Corpus Linguistics: An Introduction*. London: Continuum.

of this study, namely the East African Corpus, would be classified as a general, sample-text, synchronic and balanced corpus, subdivided into written and spoken parts. The reference corpora of this study would be categorized the same way. Detailed information about the underlying corpora will follow in the next chapter.

2.4 Important English Corpora for Research Studies

The reason why I chose to present only the most important English corpora is that the underlying study is about the East African varieties of English compared with other varieties of English, and therefore other language corpora need no inspection since they had no influence on this study. The following corpora are again drawn from Kennedy's book *An Introduction to Corpus Linguistics*.

The first electronic corpus of English was the *Brown Corpus*. It was conducted between 1961 and 1964 at the Brown University, USA, by Nelson Francis and Henry Kučera. The corpus consists of written texts printed in the USA in the year 1961. It is thus representative to the American variety of English. The corpus was designed to include 500 texts of about 2,000 words each, accounting to a total of about one million words. The corpus has been and still is very influential in terms of corpus design. Many major corpora are based on this innovative Brown Corpus as e.g. all the corpora compiled for the ICE project, which I will refer to in the following chapter. The Brown Corpus launched the creation of several further corpus projects.

One of these is the *Lancaster – Oslo/Bergen* (*LOB*) *Corpus* compiled between 1970 and 1978. Since the corpus was intended to be an exact British counterpart to the American Brown Corpus, it contains only printed texts published in Great Britain in the year 1961 and has the same structure as its American fellow.

Both corpora were followed by the *Kolhapur Corpus*, which consists of writings printed in India in the year 1978. Although it has a different year as reference point, it is compiled according to the same design as the Brown Corpus and LOB. The next two corpora developed, were the *Wellington Corpus of Written New Zealand English* (*WC*) and the *Australian Corpus of English* (*ACE*). Both corpora take 1986 as reference year and were designed in the same way as the already existing corpora. They were compiled to match the Brown Corpus and LOB. As a consequence, especially the WC had problems to be representative to New Zealand English. For some categories texts had to be sampled over a period of four years, and one popular fiction type was not included because it did not fit the pattern of the Brown Corpus. In 1991, 30 years after the

compilation of the Brown Corpus and the LOB, the University of Freiburg started a project, commonly known as *Freiburg-Brown* (*Frown*) and *Freiburg-LOB* (*FLOB*), to design equivalent corpora of American and British texts published in 1991. The project was intended to enable research in language change of written American and British English. The corpora considered so far, are called first-generation corpora by Kennedy (cf. Kennedy 1998: 30). They are all based on the design of the Brown Corpus, and thus consisting of only one million words. Especially the Brown Corpus and the LOB are considered to be the most important corpora for corpus research and still serve as linguistic source nowadays. Nevertheless, they further lack of spoken material, which is vital to language research since we use language through the medium of speech more frequently than through the medium of writing.

The first electronic spoken corpus was the *London – Lund Corpus* (*LLC*). It was created in 1975 from the spoken material of the *Survey of English Usage Corpus* (*SEU*) from 1959. It contains only about half a million words, divided into 100 texts of about 5,000 words each. Until the mid 1990s it was by far the largest spoken corpus, but it suffered from the shortage of broadly varying categories. A rather small spoken corpus was sampled between 1984 and 1987 before its compilation in 1992. The *Lancaster/IBM Spoken English Corpus* (*SEC*) contains spoken texts of British adults and accounts to 52,600 words. The counterpart is the American English variant, the *Corpus of Spoken American English* (*CSAE*). The *Wellington Corpus of Spoken New Zealand English* (*WSC*) was accomplished to complement the WC, thus containing as well about one million words, sampled between 1988 and 1993 and designed after the same categories. The last spoken corpus I want to mention is the *Corpus of London Teenage Language* (*COLT*), which consists of half a million words, sampled in texts spoken by London teenagers between the age of 13 and 17 years. The COLT was included into the British National Corpus in 1994.

Although the corpora above were subject to many linguistic research studies, their size was too small for most studies of lexical and semantic language aspects (Kennedy 1998: 45). Therefore Kennedy claims that “a corpus must be big enough to provide a substantial number of instances of a particular linguistic feature from a number of different texts in order to give us a reliable picture of how that feature is used” (Kennedy 1998: 30). During the Nineties so-called second-generation corpora appeared, consisting of a hundred million words. The first electronic mega project already began in 1980. The *Cobuild* (*Collins Birmingham University International Language Database*) project was

a cooperation of the commercial publisher Collins and a research team of the University of Birmingham. It is thus commonly known either as the *Cobuild Corpus* or as the *Birmingham Corpus*. This project resulted in the production of a new English dictionary, published in 1987 under the name *Collins Cobuild English Language Dictionary*, widely known as *Cobuild Dictionary*. In 1990 the *Cobuild Corpus* was expanded to become *The Bank of English*, which, by 1997, contained over 300 million words.

Another mega database is the *Longman Corpus Network*, consisting of three large corpora, the *Longman/Lancaster English Language Corpus* (*LLELC*), the *Longman Spoken Corpus* (*LSC*) and the *Longman Corpus of Learners' English* (*LCLE*). The database has been used for the creation of dictionaries, especially for non-native English speakers. The LSC is also part of the third major corpus, the *British National Corpus* (*BNC*). The BNC contains about 100 million words of written and spoken British English and was compiled to be representative to the whole British variety of English, rather than particular genres, as the one million-word corpora. This results in an enormously imbalanced structure. There is a huge cleft between ten million words of spoken texts and 90 million words of written material. Many of these mega corpora were designed to be monitor corpora, which often tend to be imbalanced.

The choice of the here presented corpora seems to be somewhat arbitrary. The second-generation corpora are, of course, the most important concerning the English language. However, the smaller corpora of the first-generation were chosen because they are accessible on the ICAME collection (except the CSAE) and were thus used as reference corpora for this study.

2.5 Limitations of Corpus Linguistics and Corpora

The previous sections were intended to give an overview about corpus linguistics in general, whereas this section is aimed to point at specific limitations, which not only linguists criticize, but which I experienced myself during this study. The first limitation is the availability of corpora. Most corpora of commercial or research facilities are not available. This is caused by copyright restrictions, since most mega corpora are used for dictionary making, as the Cobuild Corpus. There are some corpora which indeed can be accessed, but only through expensive acquisition, as the BNC. Only word frequency lists of the BNC can be downloaded from the internet without any purchase. Free accessible corpora are rare and often either small or compiled for special purposes. Especially for investigating the lexis of a corpus, the researcher needs a reference corpus which should

be five times the size of the examined corpus; at least this is the suggested size by Berber Sardinha (cf. Scott and Tribble 2006: 64/65). But this is hard to accomplish. The first choice of reference corpus for my study was the BNC, which was not accessible without purchase. The frequency list of the BNC turned out to be inappropriate for my purpose. Therefore, I turned to the corpora which are available on the ICAME CD. Kennedy commented on ICAME:

> This ICAME collection is one of the most valuable sources of text for corpus-based linguistic analysis, not only because it provides a convenient way of studying some of the classic one million-word corpora but also because of the accompanying software for corpus analysis. (Kennedy 1998: 86)

Since I study English linguistics I had the ability to receive the ICAME collection from the Department of Linguistics. But for other students as well as non-students this is not an option either.

However, the quotation of Kennedy reveals a further limitation of corpus linguistics, namely the access to corpus analysis software. Free software tools can be downloaded from the internet. But they are rather constructed for simple usage, such as offering frequency or alphabetical word lists. If a researcher wants to get more statistical data about the text or corpus, he has to rely on more complex software tools, which are again either not available or too expensive. The version of *WordSmith*, I originally intended to use for my study, is only a demo version, since the full version is accessible only through expensive purchase. Fortunately, the researcher Laurence Anthony provides his multi-purpose analysis tool on the internet for free. But even if the researcher has the access to every existing software tool, he might be obstructed in his study by the lack of the appropriate software which he actually needs. Thus a researcher with a humanities background is always dependent on already existing software tools. If he wants to analyse a specific feature where no existing software tool can help, he has to rely on a computational linguist or computer specialist who has the talent to program such a tool. Summing up, researchers will often face obstacles, caused by various reasons, while conducting their studies.

3 The Corpora of Study

In order to accomplish lexical studies, the researcher needs two different corpora, the corpus of study, of which he wants to retrieve lexical aspects, and the reference corpus, with which the corpus of study is compared. These two corpora should be compatible to receive best results. The corpus of this study is the East African Corpus, a component of the International Corpus of English. Both corpora will be referred to in the next two sections, constituting the basis for this current study. The last section of this chapter will be concerned with the reference corpora used for this study and the influence of reference corpora on research results.

3.1 The ICE Project

The idea behind this whole project was to enable comparative studies of the different varieties of English throughout the world. Until then, the only existing corpora designed for comparison were the Brown Corpus and the LOB, as well as the later compiled Kolhapur Corpus, the ACE and the Wellington Corpus. However, these were all written corpora and spoken corpora were only available for British, American and New Zealand English. Beware of this lack, Sidney Greenbaum, at that time director of the SEU, suggested the idea of parallel English corpora in 1988. The objective was to create corpora with a homogeneous design. (cf. Kennedy 1998: 54)

Participating countries include not only countries where English is used as a first language, but also countries where it is used as a second language or lingua franca. Today, 20 participants are engaged in this project, of which seven corpora are already completed: East Africa (Kenya and Tanzania), Great Britain, Hong Kong, India, New Zealand, Philippines and Singapore. The other corpora still under construction include the following countries: Australia, Canada, Fiji, Ireland, Jamaica, Malaysia, Malta, Nigeria, Pakistan, South Africa, Sri Lanka, USA and the recently added Trinidad & Tobago.

A specific guideline was defined to ensure compatibility. The sample texts should date from 1990 onwards. The speakers and authors of the texts were ought to be 18 or older. Furthermore, they must have been born in the country in question or moved there at an early age. A third prerequisite is that they must have received their education in English at least to the end of secondary school. Based on the design of the Brown

Corpus, the corpora should consist of 500 texts, each containing about 2,000 words, which make up a total of about one million words. These 500 texts are further subdivided into written material (200 samples) and spoken material (300 samples). The spoken material is subdivided into monolog and dialog, and the written texts into printed and non-printed. This subdivision continues to the level of 15 text categories in speech and 17 text categories in writing.[5]

Table 1: ICE Text Categories (Numbers in brackets indicate the number of 2,000-word texts)

Spoken (300)	**Dialogues** (180)	**Private** (100)	Conversations (90) Phonecalls (10)
		Public (80)	Class Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Cross-examinations (10) Business Transactions (10)
	Monologues (120)	**Unscripted** (70)	Commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		Scripted (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)
Written (200)	**Non-printed** (50)	**Student Writing** (20)	Student Essays (10) Exam Scripts (10)
		Letters (30)	Social Letters (15) Business Letters (15)
	Printed (150)	**Academic** (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Popular (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Reportage (20)	Press reports (20)
		Instructional (20)	Administrative Writing (10) Skills/hobbies (10)
		Persuasive (10)	Editorials (10)
		Creative (20)	Novels (20)

http://www.ucl.ac.uk/english-usage/ice/textcats.htm (retrieved April 24, 2008)

[5] cf. http://www.ucl.ac.uk/english-usage/ice/design.htm (retrieved April 24, 2008).

Table 1 shows the classification of the text categories according to the ICE guidelines. The division of written and spoken material as well as the different text categories enables the corpora to be representative and balanced as far as possible. The different ICE corpora should all be designed after these regulations in order to keep compatibility. This is of great advantage, since the researcher is able to use each variety corpus separately for his studies or for comparisons with other varieties of English.

3.2 The Design of the East African Corpus

The corpus project was launched in 1989 to investigate the linguistic variability of the East African English. It was part of the project "Identity in Africa" at the University of Bayreuth, and was concluded at the Department of English language and linguistics at the University of Chemnitz in 1999 with the creation of a website for online research on that corpus. The project was conducted by the project director Josef Schmied, the co-ordinator Diana Hudson-Ettle and assistant co-ordinator Barbara Krohne.[6]

The corpus includes two East African varieties of English, i.e. Kenyan English and Tanzanian English. The material was assembled between 1991 and 1996, which could not be proceeded without facing many problems. The linguistic situation in Kenya and Tanzania is different from the countries where English is used as first language. In both countries several languages coexist with English, and even the role of English is partly different. However, the linguistic situation has already been explained in the introductory chapter. The difficult situation of English in East Africa affected the design of the corpus, which should be based on the ICE regularities. This could be accomplished only through slight deviations from the classification given by the ICE.[7]

Originally it was intended to compile one corpus of both varieties, but while conducting this project the researchers experienced that this was hard to realize. Hence they decided to include two written parts, each for one variety, based on the written guidelines of the ICE. This is the reason why the ICE-EA contains more words, about 1.4 million, than the other ICE corpora. Because of the difficulties in acquiring data from spoken sources, as well as the ability to use Hansards and cross-examinations, a new category, named written and spoken, was included to compensate the lack of the spoken

[6] cf. http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/index.html (retrieved April 24, 2008)

[7] cf. ICE-EA manual: http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/index.html (retrieved April 24, 2008)

material. Many ICE categories, especially areas covering natural science and technology, were designed for developed countries and, consequently, could not be fulfilled in Third World countries such as Kenya and Tanzania. It was not intended to create an exact counterpart to other ICE corpora, though this would have been the ideal case, the objective was to compile a corpus representative to East African English. In order to obtain the best results the ICE categories had to be adjusted.[7]

Table 2: ICE-EA Text Categories

Spoken (250)	**Dialogues** (130)	**Private** (30)	Conversations (30) Phonecalls (0)
		Public (100)	Class Lessons (20) Broadcast Discussions (35) Broadcast Interviews (45) Parliamentary Debates (0) Cross-examinations (0) Business Transactions (0)
	Monologues (120)	**Unscripted** (0)	Commentaries (0) Unscripted Speeches (0) Demonstrations (0) Legal Presentations (0)
		Scripted (120)	Broadcast News (40) Broadcast Talks (40) Non-broadcast Talks (30) School-broadcast (10)
Written as Spoken (50)			Hansards (25) Legal cross-examinations (25)
Written (200-200)	**Non-printed** (50-50)	**Student Writing** (20-30)	Student Essays (10-20) Exam Scripts (10-10)
		Letters (20-10)	Social Letters (10-0) Business Letters (10-10)
	Printed (150-150)	**Academic** (40-40)	Humanities (10-10) Social Sciences (10-10) Natural Sciences (10-10) Technology (10-10)
		Popular (40-40)	Humanities (10-10) Social Sciences (10-10) Natural Sciences (10-10) Technology (10-10)
		Reportage (20-20)	Press reports (0-0) Feature/reportage (10-10) Splash (10-10)
		Instructional (10-10)	Administrative Writing (10-10) Skills/hobbies (0)
		Persuasive (20-20)	Editorials (0-0) Institutional (10-10) Personal columns (10-10)

		Creative (20-20)	Novels (20-20)

http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/studentprojects/ice /ice design2.htm (retrieved April 24, 2008)

The table above displays the categorization of the ICE texts. Table and numbers were retrieved from the above website, but the ICE-EA categories were arranged according to the table of the ICE categories. The table has been designed to point out the differences to the ICE guidelines. Red marked categories indicate a reduced number of texts in these categories and blue marked ones indicate an increased number of texts in these categories. As the written part consists of the Kenyan and Tanzanian components, and both are not identical to each other, a distinction was set up as well. The first number in brackets refers to the Kenyan variety and the second number to Tanzanian English. It can be seen that especially the spoken part differs enormously from the defined ICE regulations. There is not such a great divergence in the written component.

For more detailed information about the deviation of the different categories and its justifications, as well as further information about the coding system, please read the manual of the East African Component of the ICE.

3.3 Reference Corpora

The choice of reference corpora heavily depends on the availability of corpora. When starting this study, I planned to use the BNC as reference corpus, but had to realize soon that I could not access it without expensive purchase. The ideal corpora would have been the British, Indian and New Zealand varieties of the ICE project, since they were compiled to be compatible with the ICE-EA. But they are not accessible without purchase either. Thus, I turned to the ICAME collection which I received from the Department of English linguistics. The collection contains, amongst others, several corpora of different English language varieties. The reference corpus for this study was created of seven written corpora: the Brown and Frown Corpora (representing the American variety), the LOB and FLOB Corpora (representing the British variety), the Kolhapur Corpus (Indian variety), and the ACE and WC (representing the Australian and New Zealand variety). Four spoken corpora were also included, which, unfortunately, represent only two varieties: the LLC, SEC and COLT for the British variety and the WSC for the New Zealand one.

The choice of reference corpus plays a crucial role in lexical studies. Selecting an inappropriate reference corpus can result in receiving inappropriate outcomes. Therefore, to gain the best results possible, the choice has to be considered carefully. Mike Scott demonstrated this effect by using different reference corpora for analysing the keywords of the Shakespeare play "Romeo & Juliet" (cf. Scott and Tribble 2006: 59-65). In order to examine keywords of Kenyan and Tanzanian English, it would be ineffective to take a corpus dealing with the same varieties, since no outstanding words would be detected during the analysis process. To receive reasonable outcomes, the reference corpora are supposed to deal with different varieties of English. The reference corpus for the underlying study serves this criterion representing five different varieties. Nevertheless, it lacks of compatibility with the EA Corpus. If this lack has a negative effect on the research outcomes, will be resolved in chapter 6.

4 Keywords vs. Typicality of Lexical Items

The title of this thesis is called "Typicality of Lexical Items of Kenyan and Tanzanian English". This needs to be further discussed as it is fundamental to this study. The sociolinguistic situation in Kenya and Tanzania has already been explained in the introduction. The present, rather short, chapter, completing the theoretical background, is aimed at considering "typicality of lexical items" which is the study's main concern.

Unfortunately, there is no clear definition of typicality. The common assumption is that it is just the state of being typical. But what do we regard as being typical? Considering birds, the question seems to be obvious; at least in most people's minds or concepts, which is in fact another important issue in linguistics, but will not be paid attention to in this paper. A sparrow would be considered a typical bird whereas we would classify an ostrich as atypical. But what can be typical for a language? Since the fundamental bases of languages are words or signs, we would assume them to be typical. Nonetheless, Scott claims that words can only be typical for texts, but not for languages; or they can be typical for a culture, according to Williams and Stubbs.[8] But is it not culture that influences the language the most? Therefore, we can assume that certain words, describing people, food, politics or other cultural aspects, are typical for a language, in order to reflect the culture. Nevertheless, most linguists do not speak of typical words; they consider words to be "key" in a text or language. We can claim that "typical" equates to "key", because words that are key in a language are typical for that language as well. Thus we speak of keywords and their keyness when analysing lexical items of languages.

4.1 Keywords and Keyness

Williams defined keywords in 1983 as:

> … strong, difficult and persuasive words in everyday usage … common in description of wider areas of thought and experience … they are significant, binding words in certain activities and their interpretation; they are significant, indicative words in certain forms of thought. (Scott and Tribble 2006: 55)

[8] cf. Powerpoint by M. Scott (2005). *The Behaviour of Key Words (KWs)*: http://www.lexically.net/downloads/writing/bham/talk.ppt (retrieved July 9, 2008)

The definition given by Williams points out that keywords are closely tied to people's concepts, as I already mentioned. Twenty years ago, software tools to analyse keywords with statistical measures were not available, because software technology was still in its development. Therefore, keyword analysis was highly dependent on the researcher's point of view, i.e. what he perceived to be keywords in a text or language. Nowadays, keywords can be drawn from a corpus by using software tools, which means that results rely only on statistical probabilities. This is a further important issue which may cause problems as well, e.g. we might receive keywords which we would not consider to be key. This problem will be discussed in more detail in the next chapter, which faces the problem of obtaining the best results. In this section, however, I just want to concentrate on the theoretical scope of keywords. A keyword is thus a word reflecting some kind of importance and is associated with its "keyness".

Scott describes keyness as "a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about, avoiding trivia and insignificant detail" (Scott and Tribble 2006: 55/56). When speaking of keyness we speak of the text's aboutness, about words dealing with the important topics of a text. These words are naturally lexical items, which are content words carrying a specific meaning. Investigating keywords of a language indicates that we receive only lexical items, or content words, which reflect the culture of the country where the language is spoken. Investigating keywords of a text, we might also obtain function words as results, which give an impression of the textual style. But how can keyness be identified? It can be identified by using probabilistic measures which compare the frequencies of lexical items of the corpus being studied (either a single text or a set of texts representing a language) with their frequencies in a reference corpus. If lexical items appear more (positive keywords) or less (negative keywords) often in the target corpus than in the reference corpus, they seem to be outstanding, having a higher keyness than other lexical items, and are therefore considered to be keywords. There are several statistical measures which can be used to detect keywords, of which the two most important ones, log-likelihood and chi-square, will be presented in the following chapter. Scott explains the procedure on the trivial example of the definite article *the*, which is usually the most frequent word in an English corpus. If the frequency of *the* as a percentage of all the words of the target corpus is nearly the same as in the reference corpus, *the* will not be outstanding and therefore gets sorted out (cf. Scott and Tribble 2006: 59). The result is that most of the corpus' words are filtered out, and only a few

words remain, which are those words with a high level of keyness. Those keywords with the highest keyness will be ranked at the top of a keyword list, which enables us to distinguish between dominant and weak keywords.

Another distinction can be made between *local* and *global* keywords (cf. Scott and Tribble 2006: 66). Global keywords are spread across the corpus, occurring in most of the texts, whereas local keywords appear only in a few texts or even just in a single text, because they are mostly restricted to specific topics. My assumption is now that dominant keywords could then be classified as global keywords and weak keywords as local keywords. Whether this assumption proves to be true will be found out by analysing the keyword results in chapter 6. Global and local keywords can be easily identified by a plot, which is called a dispersion plot in Scott's *WordSmith* tool, and a concordance plot, in Anthony's *AntConc* tool.

4.2 Concordances

What we have to bear in mind is that keywords have always to be studied in context, because it is the context which reveals the usage of keywords. According to Cruse (1986): "It is assumed that the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with the actual and potential contexts" (Scott and Tribble 2006: 8). Scott distinguishes between nine contextual scopes: from scope 1, which means that the keyword is only examined in a range of a few words to both sides, to scope 9, which includes the whole culture (cf. Scott and Tribble 2006: 9). In order to take these considerations into account, researchers developed an extremely useful tool, called "concordancing", which enables us to study keywords in context. "A concordance is a list of all the examples of the target item [keyword], normally accompanied by enough context to enable a human being to study the item's occurrence in detail" (Leech and Fligelstone 1992: 127).

A concordance can be displayed in two different outputs: either as a KWIC or as a KWOC concordance. The KWIC concordance is the most common type and means "Key Word in Context". This concordance shows the keyword in the centre of the lines with as much words as preferred on both sides of the word, reflecting the context. The "Key Word out of Context (KWOC)" concordance, on the other hand, displays the keyword on the margin of the lines, which means on the margin of a sentence or paragraph which builds up the context. (cf. Mason 2000: 43)

The concordance tool has proved to be important for analysing corpora. The researcher can determine the scope of the context by setting a span on either side of the keyword. A span of four words can be sufficient for most studies; nevertheless it can also restrict the contextual horizon and reveal not enough information about the keyword. On the other hand, if a span is too large, it can become "noisy", which means that the researcher can lose track, because he is confronted with a huge bulk of words. The researcher has to find out what suits best for his purposes. (cf. Scott 1997: 235/236)

The concordance tool proved to be a very useful tool for my study. It helped me to investigate inappropriate keywords during the search for adequate results as well as analysing the appropriate keywords. The last three chapters were aimed at giving insights into the theoretical background of this underlying study, and hopefully constituted the basis for a sufficient understanding of what the study is about and how it has been achieved. The next chapter will cover the topic of finding relevant keywords before I discuss the received results in the chapter after.

5 The Search for Appropriate Results

Keyword analysis can not be accomplished without the help of software tools relying on statistical procedures, at least when ploughing through a vast majority of words in a short period of time. There are numerous analysis tools available, of which *AntConc* appealed most to me. *AntConc* is not only a free software toolkit, but also relatively easy to handle and has a clearly structured interface design. It was developed by Laurence Anthony in 2002 and is used in its third edition, *AntConc 3.2.1* (2007), for the present study. The toolkit offers a number of important analysis tools: *Concordance*, *Concordance Plot*, *File View*, *Clusters*, *Collocates*, *Word List* and *Keyword List*, of which the first and the last two are of greatest interest for this study.

Although *AntConc* is a very powerful toolkit, it does not produce appropriate results without the help and adjustments of the researcher. Extracting keywords from a corpus is not an easy task to perform, as it might seem. The researcher should be aware of what he actually wants to achieve, in order to adjust the settings correctly. This chapter tries to illuminate the steps taken to receive suitable results, as well as the difficulties and problems I was confronted with while conducting this research.

5.1 Wordlists Sorted by Frequency

The fundamental function of analysis tools is to reduce the vast amount of texts to a simple list of words. These simple wordlists can be a good starting point for further investigations. The most useful way to list words is to rank them by their frequency in descending order. But what do we regard as a word? What about numbers, capital letters, hyphens or symbols? *AntConc* offers the researcher to determine what to count as a word in the *Global Settings* menu under the *Token (Word) Definition* category. The notion of *token* has to be further explained. A *token* is each occurrence of a word in a corpus. On the other hand, *types* are the total number of distinct words in a corpus, which make up the vocabulary. Scott illustrates this with the definite article *the*: a wordlist contains the type *the* only once, together with its frequency, which is the number of all the tokens found (cf. Scott and Tribble 2006: 13).

A few more words have to be said about the division of the corpus. The ICE-EA consists of two written corpora, the Kenyan and the Tanzanian part, but only of one spoken corpus containing both varieties. Since the texts are marked according to their

variety, it is also possible to divide the spoken corpus into two subparts. The following frequency lists are thus based on the *spoken* and written part of the Kenyan variety as well as the spoken and written part of the Tanzanian one.

Table 3: Frequency Results of the Kenyan and Tanzanian Part of ICE-EA

Kenya types: 26,393 tokens: 779,885				**Tanzania** types: 23,373 tokens: 619,449			
Rank	**Word**	**Freq.**	**%**	**Rank**	**Word**	**Freq.**	**%**
1	***the***	47,460	6.09	1	***the***	41,892	6.76
2	***to***	23,267	2.98	2	***of***	22,553	3.64
3	***of***	22,274	2.86	3	***to***	18,687	3.02
4	***and***	20,974	2.69	4	***and***	17,487	2.82
5	***in***	15,563	2.00	5	***in***	14,555	2.35
6	***a***	14,620	1.87	6	***a***	10,769	1.74
7	***that***	12,925	1.66	7	***is***	8,728	1.41
8	***is***	12,171	1.56	8	***that***	8,229	1.33
9	***I***	9,687	1.24	9	***for***	6,071	0.98
10	***you***	8,491	1.09	10	***it***	4,726	0.76
□				□			
68	***said***	1,498	0.19	43	***who***	1,624	0.26
69	***like***	1,493	0.19	44	***people***	1,608	0.26
70	***very***	1,465	0.19	45	***all***	1,568	0.25
71	***then***	1,452	0.19	46	***can***	1,568	0.25
72	***your***	1,412	0.18	47	***Tanzania***	1,518	0.25
73	***should***	1,404	0.18	48	***his***	1,506	0.24
74	***time***	1,349	0.17	49	***also***	1,488	0.24
75	***Kenya***	1,329	0.17	50	***what***	1,467	0.24
76	***our***	1,319	0.17	51	***said***	1,406	0.23
77	***her***	1,285	0.16	52	***other***	1,402	0.23
□				□			
16,533	***aaah***	1	0.0001	14,420	***aa***	1	0.0001
16,534	***aaword***	1	0.0001	14,421	***aah***	1	0.0001
16,535	***aawrd***	1	0.0001	14,422	***aat***	1	0.0001
16,536	***aba***	1	0.0001	14,423	***abaru***	1	0.0001
16,537	***abadios***	1	0.0001	14,424	***abas***	1	0.0001
16,538	***abagusil***	1	0.0001	14,425	***abated***	1	0.0001
16,539	***abaluhya***	1	0.0001	14,426	***abatwele***	1	0.0001

The complete frequency lists can be found as text files on the added CD-ROM. The frequency results of the Kenyan and Tanzanian part are almost the same. The most frequent item in both varieties is the definite article *the* with over 40,000 occurrences in each corpus, making up 6.09% of the Kenyan and 6.76% of the Tanzanian part. The top ten ranked words are almost identical and belong to the category of function words. Function words are high-frequency items and have the purpose of combining lexical items to a flowing text, and therefore carry no information of the text's aboutness, but they can be style markers. There are only a few content words ranked among high-frequency items. These words are mostly concerned with what humans do, say, know and see.

Medium-frequency items are characterized by content words as shown in the table above. Among them is the first typical word for Kenyan English, *Kenya*, ranking at 75th position. It is thus the most frequent typical lexical item, occurring 1,329 times in the sub-corpus, amounting to only 0.17% of all the running words. The first characteristic word of the Tanzanian part is *Tanzania* with a frequency of 1,518 tokens and a percentage of 0.25. Although it ranks already at 47th position, in comparison to *Kenya*, the percentage is relatively small if compared to 6.76% of the determiner *the*.

Most types belong to the category of low-frequency items, which are only characterized by content words. Items with a frequency of three or lower make up incredible 60% of both sub-corpora. The so-called *hapax legomena* are words which have a frequency of one. As the table above shows, *hapax legomena* start at rank 16,533 in the Kenyan corpus, which, calculated as a percentage, amounts to more than 37%. This means that 37% of the words occur only once in the entire Kenyan part of the EA corpus. In the Tanzanian part even 38% of the types have a frequency of one. Many of these words may be spelling mistakes as well as typos, proper names or African words. But among them are, surprisingly, a few English words, such as *pencil* at rank 23,187 of the Kenyan corpus, or *beef* ranking at position 15,160 in the Tanzanian part.

However, these frequency wordlists do not tell us anything about the characteristics of Kenyan and Tanzanian English. As Scott puts it: "we have done [not] more than scratch at the surface of the discoveries that are there to be made about them" (Scott 2001: 57)[9].

[9] Scott, M. (2001). "Comparing Corpora and Identifying Key Words, Collocations, and Frequency Distributions through the WordSmith Tools Suite of Computer Programs." In M. Ghadessy, A. Henry, and R.L. Roseberry, eds. (2001). *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam and Philadelphia: John Benjamins Publishing Company, 47-67.

5.2 Statistical Procedures

Retrieving keywords from a corpus is based on two concepts: frequency and probability. We already gained frequency lists of the two varieties in the last section, which are the basis for further calculations. In order to extract keywords, we have to compare these lists with frequency lists of other varieties. Thus we need more sophisticated procedures such as probabilistic measures.

The two most important measures are the chi-squared statistic and the log-likelihood statistic. One important earlier study, based on the chi-squared test, was the comparison of the Brown and LOB corpora, which was conducted by Hofland and Johansson in 1982. Many other researchers have used this statistical measure as well to investigate differences between corpora. However, it has become apparent that the chi-squared test tends to be unreliable for two reasons: when the expected frequency, and sometimes also the observed frequency, is too small, occurring less than 5 times; and when a small corpus is compared to a significantly larger one. It, furthermore, seems to overestimate high-frequency words. Therefore, it is suggested to use the log-likelihood ratio to rectify these limitations. (cf. Rayson and Garside 2000: 2)

Both statistical measures can be chosen to investigate keywords with the *Keyword List* tool of *AntConc*. The researcher can choose in the *Tool Preferences* menu under the category *Keyword List* which keyword generation method to use. To illustrate the difference between the keyword outcomes for both statistics, I compared the whole Tanzanian part of the ICE-EA with the frequency list created of all reference corpora already presented in chapter 3.3. These corpora are used for all further keyword analyses and will not be mentioned again. If the assumption, that the chi-squared statistic is unreliable, proves to be true, or if it has to be rejected due to the fact that there are hardly any deviations between the two statistics, will be resolved by regarding the table below.

Table 4: Tanzanian Keyword Results for the Chi-Squared and Log-Likelihood Statistics

Chi-Squared				Log-Likelihood			
Rank	**Word**	**Freq.**	**Keyness**	**Rank**	**Word**	**Freq.**	**Keyness**
1	***uh***	3,839	49,731.175	1	***uh***	3,839	19,451.537
2	***Tanzania***	1,518	20,212.462	2	***Tanzania***	1,518	7,989.907
3	***Dar***	473	6,310.352	3	***Dar***	473	2,498.007
4	***Salaam***	457	6,111.166	4	***Salaam***	457	2,424.119
5	***Zanzibar***	303	3,971.383	5	***the***	41,892	1,941.631
6	***Kiswahili***	296	3,968.173	6	***African***	506	1,641.114
7	***ndugu***	288	3,860.926	7	***Kiswahili***	296	1,579.245
8	***African***	506	3,766.929	8	***Zanzibar***	303	1,558.274
9	***Africa***	503	3,091.481	9	***ndugu***	288	1,536.563
10	***development***	895	2,606.948	10	***development***	895	1,495.373
11	***nineteen***	486	2,379.379	11	***is***	8,728	1,448.142
12	***appellant***	213	2,370.226	12	***Africa***	503	1,427.500
13	***Tanzanian***	177	2,312.605	13	***are***	4,606	1,399.208
14	***countries***	646	2,296.998	14	***of***	22,553	1,299.273
15	***the***	41,892	2,078.864	15	***countries***	646	1,247.864
16	***CCM***	153	2,051.117	16	***to***	18,687	1,181.499
17	***are***	4,606	1,697.671	17	***nineteen***	486	1,177.460
18	***arusha***	124	1,662.343	18	***in***	14,555	1,103.565
19	***is***	8,728	1,660.553	19	***government***	1,154	1,040.722
20	***Tanzanians***	119	1,595.313	20	***which***	3,228	960.851

The table displays the first twenty keywords with their frequency and keyness, and verifies the assumption that the chi-squared method tends to produce higher keyness results. The first four items are identical for both methods, except that they have a different calculated keyness. The four are followed by *Zanzibar* for the chi-squared test as well as *the* for the log-likelihood method, which gives the impression that the first method produces better keyword results. The definite article *the* appears also in the chi-squared statistic, but only on 15th position. Though it has a higher keyness of 2,078.864 in contrast to 1,941.631 for the log-likelihood method, it is ranked much lower. Regarding the other words the impression gets even stronger, since the retrieved chi-squared results contain only three function words whereas the log-likelihood results contain twice as much. Although the chi-squared method tends to overestimate and, therefore, produces much higher keyness results, it is more suitable to my research interests. The study does not depend on receiving correct

keyness results; it deals with detecting keywords and analysing them. On the whole, both measures seem to produce approximately the same results, but the outcomes received with the chi-squared method are slightly better than the outcomes of the log-likelihood measure. Therefore, I will use the first method for all further investigations. This choice does not rely only on the performed searches so far, it is based on the experiences I received while experimenting with the tool and corpora to gain adequate results.

However, the so far received outcomes can not be satisfactory since they still contain function words, proper names and lexical items which I would not consider to be key for that language. Therefore, the settings of the analysis tool have to be adjusted correctly to detect the best keywords possible.

5.3 Adjustments of the Tool Settings

The *AntConc* toolkit is extremely versatile due to its various adjustments. The *Tag Settings* category of the *Global Settings* menu provides two options, either to display or to hide tags. The preferred option for this study is to hide tags, so I had to change the default according to the tags which were used for marking the ICE-EA corpus. The two most important categories are *Word List* and *Keyword List* in the *Tool Preferences* menu. Both categories offer the option to treat all data as lowercase, which should be implemented in any case. If it is not applied, the program considers, e.g. capitalized *The*, at the beginning of a sentence, and usual *the* as two types instead of one type, which has an enormous effect on the outcomes. These are the only modifications I did so far, as well as the choice of the chi-squared method, and the determination of the threshold value to the option *average value*, which means that the average number of keywords will be displayed.

Nevertheless, using only these adjustments is not sufficient, since the keyword results still contain words which are not considered to be typical, as shown in table 4. The numerous occurrences of function words appear to be one problem, which can be solved with a so-called stoplist. This list enables the program to ignore all the words the list contains. Professor Jian-Yun Nie from the Department of Computer Science and Operations Research of the Université de Montréal is especially involved in Information Retrieval and created such a stoplist, which I used for this study.[10]

[10] The stoplist can be retrieved from: http://www.iro.umontreal.ca/~nie/IFT6255/common_words. (retrieved July 22, 2008)

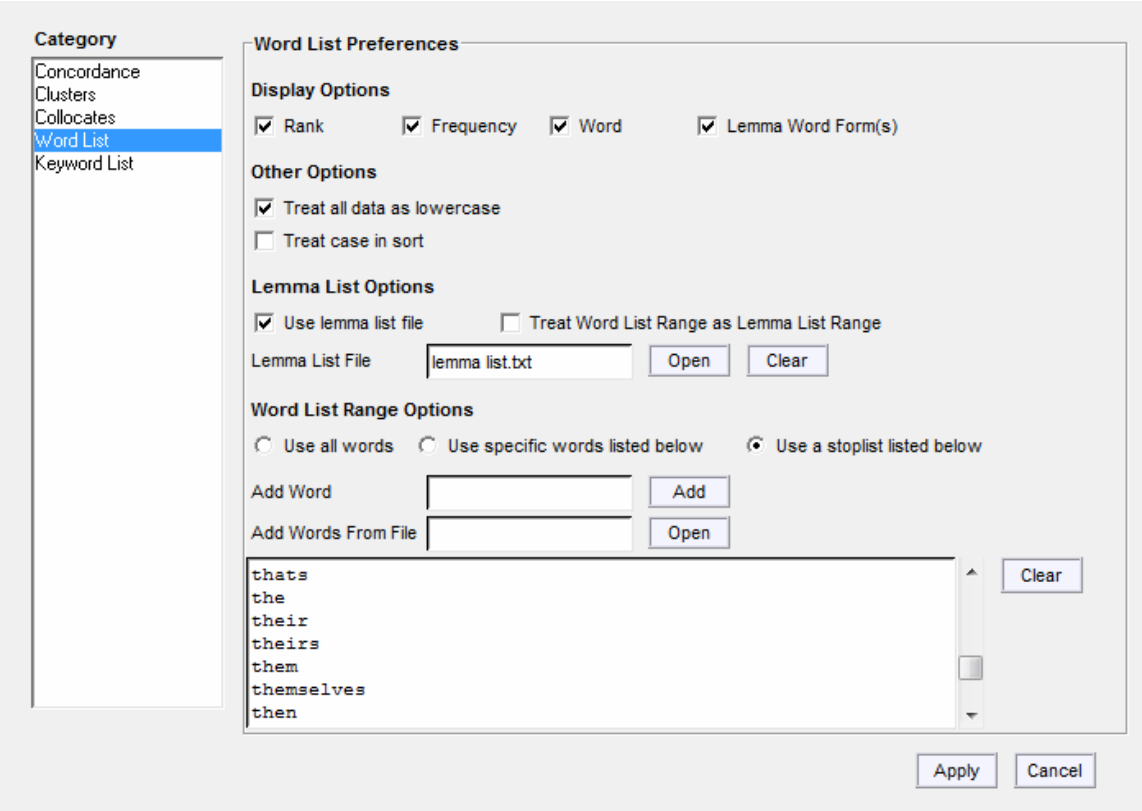


Figure 1: Word List Tool Preferences

Figure 1 shows a screenshot of the *Word List* category. A part of the stoplist can be seen at the lower end of the shot, whereas the middle part of the shot reveals a further option. Besides using a stoplist, the researcher can also load a so-called lemma list, which assigns lemmas such as *wants* or *wanted* to the type *want*. Lemma lists can also be retrieved from the internet, including various verbs and nouns. Many verb forms are already excluded anyways, due to the stoplist, and therefore I created a lemma list by myself, including only the lemma types influencing the keyword results. Examples of these lemma types are *Tanzanian* and *Tanzanians*, which are assigned to the type *Tanzania*; *Kenyan* and *Kenyans* to *Kenya*; and *African* and *Africans* to *Africa*. A further assignment had to be made concerning shillings. The Kenyan wordlist contains three different types: the right expression *shillings* as well as its abbreviations *shs* and *Kshs* (for Kenya shillings). But both abbreviations refer to only one meaning, shillings, and are therefore assigned to this type. When regarding the first keyword results, I had to experience that plural words were counted as distinct types, as well as some adjectives, e.g. *regional*. I also discovered that *Rwanda* is spelled in a further variation, *Ruanda*, which I had to take into consideration as well. The next screenshot displays the Kenyan wordlist together with the lemma word forms of *Kenya* and *Africa*, as well as some plural lemma forms.

Hits | Total No. of Word Types: 20614 | Total No. of Word Tokens: 158979

Rank	Freq	Lemma	Lemma Word Form(s)
1	1079	kenya	kenya 748 kenyan 172 kenyans 159
2	716	people	
3	680	children	child 183 children 497
4	599	women	
5	582	africa	africa 285 african 264 africans 33
6	559	language	language 436 languages 123
7	546	country	countries 216 country 330
8	506	development	
9	409	government	
10	408	areas	area 169 areas 239
11	368	school	school 261 schools 107
12	351	problem	problem 183 problems 168
13	332	world	
14	323	services	service 121 services 202
15	315	education	

Figure 2: Wordlist Using a Lemma List

Using a lemma list results in a change of the word's keyness and consequently in boosting up its rank. Without lemma list, the type *shillings* ranks at 22nd position, *kshs* at rank 25 and *shs* at rank 47. Yet, the type *shillings*, including the lemma forms *kshs* and *shs* ranks at 9th position when using a lemma list. *Kenya* still ranks at 2nd position, including the lemma forms *Kenyan* and *Kenyans*, which rank at place 7 and 9 without lemma list. The same consequences affect the types *Africa* and *Tanzania*. Although lemma lists drive up the types' ranks, they have no influence on the keyword outcomes, only on the keyness results. As I already mentioned, the aim of this study is to retrieve keywords, it does not matter if the keyword *shillings* ranks at 9th or at 22nd position, what matters is that it is detected as a keyword.

If we now apply the discussed considerations, i.e. using a stop and a lemma list, we receive the following keyword outcomes for the Tanzanian part of the ICE-EA.

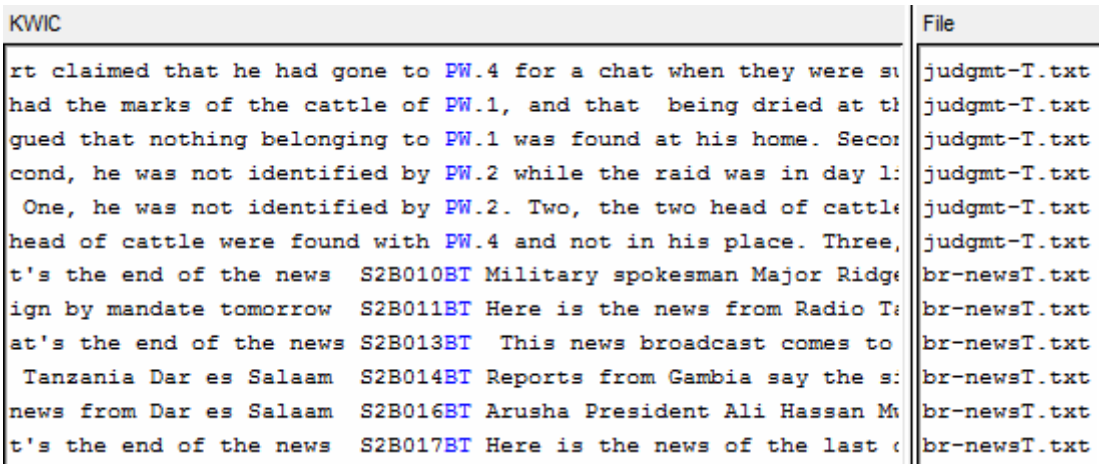
Hits | Keyword Types Before Cut: 22815 | Keyword Types After Cut: 3423

Rank	Freq	Keyness	Keyword
1	3839	114505.144	uh
2	1814	55556.649	tanzania
3	1058	22194.510	africa
4	473	14497.646	dar
5	457	14037.614	salaam
6	1514	13635.572	country
7	303	9136.000	zanzibar
8	296	9113.387	kiswahili
9	288	8867.080	ndugu
10	306	8339.095	appellant
11	954	7576.604	problem
12	604	7467.864	programme
13	895	7396.029	development
14	774	6999.515	areas
15	513	6561.039	region
16	486	6118.048	nineteen
17	1154	5355.275	government

Figure 3: Tanzanian Keyword Results Using a Lemma List and Nie's Stoplist

The figure shows the first 17 keywords as produced and displayed by the *Keyword List* tool, and it seems that the adjustments helped to produce far better results. Nevertheless, the keyword list still contains items, which are not key to Tanzanian English. Therefore I created a stoplist by myself. Such an item is the first keyword *uh*, which is used by people for filling a pause while searching for the next word to say. The type *uh* occurs only in the spoken part of the Tanzanian corpus, and is to be seen as a result of a different transcribing than the one used for the reference corpora. I do not consider this to be a keyword and, therefore, added *uh* to the stoplist. The next word, which comes to my notice, but can not be seen in the screenshot, is the word *CCM*, ranking at the 20th position. As all data is treated as lower case, it first seems that *ccm* is the abbreviation of cubic centimetres, but when using the *Concordance* tool, it becomes immediately apparent that it is the abbreviation of an important Tanzanian party, the "Chama Cha Mapinduzi", and can thus clearly be classified as a keyword.

Two more keyword results need some closer examination. These are *PW*, ranking at the 29th position, and *BT* placed at the 44th rank. At first sight, it is not clear what these types stand for. That is why a further contemplation with the *Concordance* tool is indispensable.



KWIC	File
rt claimed that he had gone to PW.4 for a chat when they were su	judgmt-T.txt
had the marks of the cattle of PW.1, and that being dried at th	judgmt-T.txt
gued that nothing belonging to PW.1 was found at his home. Secor	judgmt-T.txt
cond, he was not identified by PW.2 while the raid was in day li	judgmt-T.txt
One, he was not identified by PW.2. Two, the two head of cattle	judgmt-T.txt
head of cattle were found with PW.4 and not in his place. Three,	judgmt-T.txt
t's the end of the news S2B010BT Military spokesman Major Ridge	br-newsT.txt
ign by mandate tomorrow S2B011BT Here is the news from Radio Ta	br-newsT.txt
at's the end of the news S2B013BT This news broadcast comes to	br-newsT.txt
Tanzania Dar es Salaam S2B014BT Reports from Gambia say the si	br-newsT.txt
news from Dar es Salaam S2B016BT Arusha President Ali Hassan Mw	br-newsT.txt
t's the end of the news S2B017BT Here is the news of the last c	br-newsT.txt

Figure 4: KWIC Concordances of *PW* and *BT*

This "Key Word In Context" concordance shows the search terms in the centre of the lines. The right-hand column displays the files where the search terms appear in. As can be seen, *PW* appears in texts dealing with judgements. It is the abbreviation of *Prosecution Witness* and clearly not a keyword of Tanzanian English. The KWIC display reveals that the second term *BT* is used for marking the texts, and can therefore not be used for text analysis. As the texts of the reference corpora are marked with other letters, *BT* appears to be outstanding, although it should not be included in text analysis.

There are, of course, further types, which the analysis tool detected as keywords, such as year numbers, e.g. *nineteen* (rank 16) or *seventy* (rank 97), but it would be of no use to present them all and justify why I do not consider them to be typical. I just want to mention a few words of the Kenyan results, such as the term *coz*, the short form of *because*, as well as the expressions *yah, yeah* and *aha* used in conversations, which, apart from *uh*, are important keywords for analysing the style of the language, but not for analysing culture specific keywords. The written corpora parts constituted another problem because they contain the category of creative writing. The result is that a whole bunch of fictitious proper names are detected as keywords. This problem applies not to the whole corpora, but since I want to compare the spoken and written parts, I had to take this into consideration as well. Much of these proper names were added to the stoplist, whereas the proper names of real people were left unchanged.

Although the keyword analysis tool is an easy and quick way to analyse typical words, it has a severe problem, i.e. it can not distinguish between homonyms. An example is the item *case*, which is ranked at 96th position in the Tanzanian corpus and at 107th in the Kenyan corpus. *Case* can take the meaning of court case or of instance in such as expressions as *in case of*, as well as the minor meaning of suitcase. Though *case*,

referring to court case, appears 86 times in judgement texts in the Tanzanian corpus and 92 times in cross-examination texts in the Kenyan corpus, it is outnumbered by the meaning of instance. The court case could be a keyword, whereas the expression *in case of* is rather a style marker. Nevertheless, I decided to add *case* to the stoplist, because the meaning of court case is already presented by the keyword *court*.

However, the stoplist proved to be a solution, for another problem I was confronted with. It can be argued if this is a good solution or not, but I found no other way to solve this problem. Figure 4 shows the keyword results of Tanzania, containing the type *Dar* at rank 4 and S*alaam* at rank 5. It is obvious that these words belong to the largest city of Tanzania, *Dar es Salaam*. Although *Dar* appears 16 times more often than *Salaam*, these two words are inextricably linked together, since D*ar* is often used as short form of *Dar es Salaam*. This conclusion can be verified with the *Clusters* tool of *AntConc* as shown in the Figure below.

Rank	Freq	Cluster
1	416	Dar es Salaam
2	16	Dar-es-Salaam
3	10	DAR ES SALAAM
4	7	Dar es salaam
5	2	Dar-es-salaam

Figure 5: Three-Word Cluster of *Dar*

The cluster *Dar es Salaam* in five different spelling variations already accounts to 451 tokens. The remaining 22 occurrences of *Dar* also refer to the city. The problem arises from the city's spelling since the computer program recognizes it as three different words, due to the spaces. The spelling variation with dashes could be used when modifying the *Token (Word) Definition* category in the *Global Settings* menu. But this not an option since the majority of the tokens is written with spaces. Therefore, I was forced to solve this problem by adding *es* and *Salaam* to the stoplist. The consequence is that the tool removes *Salaam* and *es* from the wordlist and displays only the word *Dar*. Thus, the type *Dar* stands for the city *Dar es Salaam* in all further keyword results. The same problem applies to the expression *lingua franca* as well as the plural form *lingua francas*. The type *franca* occurs 47 times and *francas* 28 times in the Kenyan corpus. Adding up the tokens we receive 75, which is exactly the number of occurrences of the type *lingua*. This means that *lingua* already contains the items *franca* and *francas*. As a consequence, I chose to add both items to the stoplist. The result is that only the type

lingua represents the expression *lingua franca* as well as its plural form in the final keyword outcomes.

The stoplist as well as the lemma list I created can be found in appendix 7 and on CD-ROM, which also contains the stoplist by Nie, and all interim results I used for the discussion of finding appropriate keywords. The stop and lemma words were not chosen arbitrary. The aim was to create a list with the first 120 adequate keywords for further analysis. After these 120 words, the list probably contains a whole bunch of words, which are not considered to be key as well as lots of lemma forms. But since each list consists of more than 3,000 keywords, it is just too time consuming to go through the whole list to exclude all inappropriate items.

5.4 Keywords of Kenyan English

After I have discussed the problems of finding the best outcomes, it is now time to apply these considerations to the Kenyan part of the ICE-EA, in order to get the final results. This section serves only to create the keywords. An analysis of the results will follow in the next chapter. The first keyword list is produced of the written and spoken part of the Kenyan corpus.

Table 5: Final Keyword Results of the Kenyan Part of the ICE-EA

Rank	Word	Freq.	Keyness
1	***Kenya***	1,888	53,568.082
2	***Africa***	999	19,207.262
3	***Nairobi***	602	17,018.699
4	***speaker***	791	13,146.960
5	***accused***	586	12,228.080
6	***children***	1,392	10,487.340
7	***shillings***	358	9,310.626
8	***Moi***	288	8,090.491
9	***country***	1,142	7,464.289
10	***hon***	340	6,902.483
11	***people***	2,065	6,878.270
12	***areas***	791	6,751.531
13	***programme***	575	6,403.195
14	***KANU***	219	6,281.633
15	***women***	1,085	6,115.879

16	***language***	812	5,474.526
17	***development***	794	5,471.126
18	***MFI***	182	5,245.852
19	***Kiswahili***	179	5,159.382
20	***problem***	812	5,159.348
21	***environment***	422	4,667.682
22	***members***	677	4,555.966
23	***minister***	655	4,221.474
24	***teacher***	411	4,103.780
25	***officer***	399	4,054.300
26	***crops***	226	3,996.877
27	***learner***	149	3,703.908
28	***services***	520	3,579.135
29	***students***	523	3,539.509
30	***Tanzania***	129	3,458.916

This table presents only the first 30 keywords. The first 120 keywords can be found in appendix 1. The entire keyword results are added to CD-ROM. In case, somebody is interested in comparing the final results of the chi-squared statistic with the final results of the log-likelihood statistic, I added the latter results to the CD-ROM. The keywords calculated with the log-likelihood measure seem to be adequate results, at least until the 70th rank, but further down the list, more and more inappropriate items appear.

The next keyword lists are based on either the written part or on the spoken part of the Kenyan corpus. I have decided to bring them up here, to show the differences in using English as a written or spoken medium. The first 30 keywords already reveal the effect of the Hansards and cross-examination categories, included into spoken material, which will be explained in detail when analysing the Kenyan and Tanzanian keyword results in detail. The list of the first 120 keywords can be looked at in appendix 2 for the written part, and appendix 3 for the spoken part. The entire lists can be found on CD-ROM if somebody is interested in further research.

Table 6: Final Keyword Results of the Written and Spoken Part of the Kenyan Corpus

	Written				Spoken		
Rank	**Word**	**Freq.**	**Keyness**	**Rank**	**Word**	**Freq.**	**Keyness**
1	***Kenya***	1079	54,618.378	1	***Kenya***	809	50,205.245
2	***Africa***	582	16,524.503	2	***accused***	553	25,668.487
3	***Nairobi***	257	12,790.668	3	***speaker***	704	25,453.308
4	***learner***	142	6,353.566	4	***Nairobi***	345	21,468.542
5	***language***	559	5,325.117	5	***Hon***	332	15,148.579
6	***pest***	118	5,250.808	6	***shillings***	241	13,414.630
7	***KANU***	102	5,248.999	7	***Africa***	417	12,431.542
8	***children***	680	5,234.535	8	***Moi***	196	12,150.067
9	***environment***	301	4,982.342	9	***MFI***	182	11,698.138
10	***discourse***	155	4,817.162	10	***people***	1349	7,882.661
11	***crops***	165	4,687.249	11	***minister***	547	7,541.161
12	***shillings***	117	4,596.224	12	***KANU***	117	7,454.518
13	***development***	506	4,582.742	13	***children***	712	7,351.551
14	***Moi***	92	4,431.446	14	***Kiswahili***	107	6,877.477
15	***areas***	408	3,837.143	15	***officer***	320	6,624.699
16	***women***	599	3,772.700	16	***members***	484	6,062.269
17	***Kiswahili***	72	3,742.913	17	***assistant***	198	5,748.330
18	***fuelwood***	73	3,639.338	18	***country***	596	5,368.511
19	***country***	546	3,477.806	19	***deputy***	183	5,269.006
20	***Malaria***	92	3,251.794	20	***programme***	309	5,217.582
21	***Imanyara***	62	3,223.064	21	***Wananchi***	73	4,692.110
22	***lingua***	74	3,211.376	22	***district***	278	4,683.714
23	***programme***	266	3,141.412	23	***problem***	461	4,381.689
24	***farmers***	212	3,098.743	24	***areas***	383	4,327.135
25	***maize***	74	2,931.486	25	***caliban***	67	4,306.457
26	***services***	323	2,850.636	26	***president***	377	4,243.557
27	***Kipsigis***	53	2,755.200	27	***Tanzania***	74	4,224.600
28	***Kenyatta***	55	2,754.990	28	***receipt***	99	3,836.943
29	***pesticides***	67	2,589.006	29	***Kiliku***	55	3,535.152
30	***parastatal***	50	2,495.429	30	***teacher***	229	3,507.952

5.5 Keywords of Tanzanian English

The so far received results of the Tanzanian part of the ICE-EA were not appropriate, but after applying all adjustments, we finally obtain the following list.

Table 7: Final Keyword Results of the Tanzanian Part of the ICE-EA

Rank	Word	Freq.	Keyness
1	***Tanzania***	1,814	58,740.958
2	***Africa***	1058	23,502.013
3	***Dar***	473	15,328.560
4	***country***	1514	14,525.084
5	***Zanzibar***	303	9,660.148
6	***Kiswahili***	296	9,635.566
7	***Ndugu***	288	9,375.145
8	***appellant***	306	8,820.677
9	***problem***	954	8,081.493
10	***programme***	604	7,933.117
11	***development***	895	7,885.488
12	***areas***	774	7,455.823
13	***region***	513	6,968.037
14	***government***	1154	5,754.565
15	***organisation***	332	5,272.879
16	***environment***	417	5,226.934
17	***CCM***	153	4,980.546
18	***language***	701	4,768.595
19	***village***	436	4,648.463
20	***people***	1608	4,619.012
21	***education***	625	4,496.347
22	***Arusha***	124	4,036.521
23	***parastatal***	125	4,001.058
24	***economic***	583	3,918.874
25	***services***	494	3,748.880
26	***accused***	223	3,581.035
27	***nation***	370	3,544.278
28	***Kenya***	132	3,498.927
29	***issue***	444	3,497.375
30	***women***	752	3,353.690

The list of the first 120 keywords is given in appendix 4, as well as the complete results on CD-ROM, together with the results received with the log-likelihood measure.

As I have done with the Kenyan corpus, the Tanzanian one will also be displayed as written and spoken part separately. It can be seen that the list of the first 30 results contains quite different keywords. Appendix 5 shows the first 120 keywords of the written Tanzanian part and appendix 6 contains the spoken part. Both complete lists can be viewed at the CD-ROM for any research.

Table 8: Final Keyword Results of the Written and Spoken Part of the Tanzanian Corpus

	Written				Spoken		
Rank	**Word**	**Freq.**	**Keyness**	**Rank**	**Word**	**Freq.**	**Keyness**
1	***Tanzania***	1123	53,122.725	1	***Tanzania***	691	69,907.581
2	***Africa***	627	16,857.743	2	***Dar***	210	21,320.560
3	***Kiswahili***	288	13,737.436	3	***Africa***	431	21,127.047
4	***appellant***	306	12,955.594	4	***Zanzibar***	188	18,664.661
5	***Dar***	263	12,446.302	5	***country***	684	11,878.455
6	***Ndugu***	233	11,113.968	6	***leprosy***	103	9,352.398
7	***country***	830	7,296.396	7	***problem***	488	8,270.056
8	***language***	653	6,504.015	8	***parastatal***	80	7,997.069
9	***CCM***	130	6,200.926	9	***Rwanda***	76	7,587.281
10	***region***	347	5,464.761	10	***organisation***	194	7,418.285
11	***development***	575	5,318.429	11	***environment***	243	7,053.184
12	***Zanzibar***	115	5,202.031	12	***Arusha***	59	6,048.234
13	***programme***	352	4,761.558	13	***programme***	252	5,951.634
14	***education***	505	4,663.973	14	***pesticides***	73	5,716.648
15	***accused***	203	4,628.242	15	***Ndugu***	55	5,638.184
16	***UDSM***	93	4,436.047	16	***people***	836	5,036.835
17	***areas***	462	4,418.012	17	***areas***	312	4,902.604
18	***Mrema***	78	3,720.556	18	***nation***	214	4,605.156
19	***Swahili***	79	3,534.556	19	***women***	443	4,563.088
20	***HIV***	97	3,270.091	20	***government***	512	4,339.683
21	***problem***	466	3,140.834	21	***hundred***	280	4,204.568
22	***Arusha***	65	3,100.463	22	***Bawata***	41	4,203.010
23	***Amref***	64	3,052.764	23	***Uganda***	52	4,177.872
24	***offence***	138	2,978.181	24	***thousand***	202	4,076.616

25	***shillings***	88	2,928.029	25	***development***	320	4,033.734
26	***services***	342	2,884.219	26	***Zambia***	40	3,901.427
27	***Mwinyi***	58	2,766.567	27	***village***	194	3,753.827
28	***respondent***	100	2,758.877	28	***Nigeria***	47	3,753.827
29	***economic***	385	2,718.515	29	***workshop***	84	3,614.359
30	***Kenya***	78	2,696.715	30	***Kenya***	54	3,608.658

6 Analysis of the Keywords Extracted from the ICE-EA

This chapter discusses the results obtained through the complex retrieving process. It will start with analysing the keywords of the Kenyan corpus, followed by a discussion of the Tanzanian keywords. The results of both varieties will then be compared by investigating similarities and differences.

6.1 Discussion of the Kenyan Keyword Results

The section will be divided into two main parts. The first part is aimed at analysing in detail the first 30 keywords displayed in table 5. The discussion will then be expanded to the first 120 keywords shown in appendix 1.

6.1.1 A Detailed Analysis of the First Thirty Kenyan Keywords

It is not surprising that the keyword, ranking at the top of the list, is *Kenya*. It is not surprising either that it is followed by the items *Africa* and *Nairobi*, the capital of Kenya. The next typical items, at least a layman would assume, are the currency of Kenya, *shillings*, placed at the 7th rank, as well as Kenya's lingua franca, *Kiswahili*, ranked at position 19, and its neighbouring country *Tanzania*, placed at the end of the list. Regarding only these six lexical items, the observer immediately associates them with Africa, Kenya and Tanzania, without exactly knowing which corpus is under consideration. This seems to be a first hint that the extracted keywords tend to be strong indicators of what the corpus is about. Though these six keywords are self-explanatory, I want to examine how they are distributed within the corpus.

The *Concordance* tool of *AntConc* displays 1,329 hits for *Kenya*, the first keyword to be considered. It should be mentioned that the number of these hits does not equal to the frequency number of *Kenya*, which accounts to 1,888 tokens. In the cases of *Nairobi* and *Kiswahili* the number of concordance hits and the frequency number match exactly, but usually the first number is lower than the frequency number. Concordance hits can be further analysed according to the categories where *Kenya* appears in. This can be accomplished by the so-called *Concordance Plot*, which enables us to identify the hits in each category.

Figure 6: Concordance Plots of *Kenya*

The figure shows the concordance plots in six different ICE-EA text categories. They were chosen to demonstrate the divergent distribution of *Kenya*. Hit file 2 (representing one conversation category), hit file 7 (Hansards) and hit file 8 (broadcast news) belong to the spoken part of the Kenyan corpus. It can be seen that the keyword hardly appears in conversation with only 13 entries whereas it occurs 162 times in the broadcast news category, which is as well the highest number of hits in all categories. The other three files belong to the written part of the Kenyan corpus. Hit file 12 (business letters) contains the keyword only 43 times, hit file 27 (popular social sciences) shows a slightly higher number with 56 hits, and hit file 30 (splash) outnumbers both with 90 entries. This plot figure makes obvious that the keyword *Kenya* appears more in the public, or rather news section, than in the, what I would call, “private life” section. A detailed distribution including percentages is given in the table below.

Table 9: Distribution of Kenya across the Kenyan Corpus

		Categories	Distribution of *Kenya* in %
Spoken Part 43.8%	Private	Conversation	1.2%
	Public 15.8%	Broadcast Discussions	6.2%
		Broadcast Interviews	5.4%
		Class Lessons	4.2%
	Scripted 19.1%	Broadcast News	12.2%
		Broadcast Talks	5%
		School-Broadcast	1%
		Non-Broadcast	0.9%
	Written as Spoken 7.7%	Hansards	5.3%
		Cross Examinations	2.4%
Written Part 56.2%	Letters 3.7%	Business Letters	3.2%
		Social Letters	0.5%
	Student Writing 1.3%	Essays	0.9%
		Exams	0.4%
	Academic 15.7%	Humanities	2.9%
		Natural Sciences	2.9%
		Social Sciences	4.8%
		Technology	5.1%
	Popular 14.6%	Humanities	4.1%
		Natural Sciences	2.2%
		Social Sciences	4.2%
		Technology	4.1%
	Reportage 11.1%	Reportage-Feature	4.3%
		Splash	6.8%
	Persuasive 7.5%	Institutional	4.5%
		Personal Columns	3%
	Instructional	Administrative Writing	1.4%
	Creative	Novels/Stories	0.9%

The table shows that *Kenya* is spread over the entire corpus, occurring in each ICE-EA (Kenya) category. It is not an equal distribution, since it appears more often in the "public sector", including such categories as broadcast news and splash, than in the "private sector", e.g. conversation and social letters. Coming back to the assumption I made in chapter 4.1, that dominant keywords can be classified as global keywords, I can state that this assumption proves to be true for the case of *Kenya*. Kenya ranks at the top of the list as the most dominant keyword. As the table has shown, it is distributed over the entire corpus and therefore clearly a global keyword. Whether the other five keywords show the same distribution will be determined now.

I will not consider these keywords as detailed as I did with *Kenya*, since applying this procedure to every keyword detected, would far exceed the scope of this paper. The keyword *Africa* is spread over almost all categories, only missing the exam category of the written part. In the spoken part, it occurs mostly in the broadcast discussions (41 hits) as well as broadcast news (34) categories. Regarding the written part, *Africa* appears only 17 times in the splash category, but 49 times in the essay and 39 times in the academic social sciences categories. The conclusions drawn for *Kenya*, that it is more a keyword of the public sector, can be applied to *Africa* as well; and since *Africa* is distributed over the most part of the corpus, it is also classified as a global keyword, which verifies my assumption.

The next keyword under consideration is *Nairobi*. One third of the hits in the spoken part, occur in the broadcast news category (106). Taking a look at the written part, we recognize that it appears only 27 times in the whole part of academic writing, whereas it occurs in the single category of business letters twice as much (58). As Nairobi is the capital and also largest city of Kenya, most businesses are established there, and therefore the high number is not surprising. Although *Nairobi* has no entries in 5 categories, it is still spread across the corpus, and can be regarded as a global keyword.

The investigation of the item *shillings* proved to be interesting. The keyword appeared 140 times in the spoken part of the corpus, of which considerable 101 hits occurred in the single category of broadcast news. In the early nineties, Kenya underwent a structural and economic reform. After it had suffered three devaluations of the shilling in 1993, another exchange rate system was implemented. The texts for the corpus were sampled between 1991 and 1996, a period when the topic about shillings was prominent in the news. One could assume now, that the result reflects the situation of Kenya at that time. But taking a closer look of the concordances with the *File View*

tool, we recognize that the item is mostly mentioned together with donations and expenses, and therefore the assumption has to be rejected. However, in the written part, *Shillings* has only 23 entries, but these are spread evenly across ten categories, sometimes occurring only once or twice. Though the item is not presented in 9 categories (8 categories belong to the written part), I would still consider it to be a global keyword.

The distribution of the keyword *Kiswahili* is somewhat arbitrary. It is the first item which does not appear in the broadcast news category at all, and occurs only once in the splash category. Of 107 hits in the spoken part, 94 appear in the class lessons category. Since Kiswahili is the lingua franca of Kenya, it will surely be taught at school, which explains the high number of entries in contrast to the other categories. In the written part, *Kiswahili* appears mostly in humanities texts, either academic or popular. The term *Kiswahili* is far more used than *Swahili*, as the language is called by other people. It is ranked at the 52nd place with only 75 tokens, in contrast to 179 for *Kiswahili*.

The item *Tanzania* has the fewest entries of all regarded keywords so far. It is almost equally distributed across the corpus, occurring 50 times in the spoken and 44 times in the written part. The class lessons category contains the most entries (39); the remaining 56 hits are spread evenly across the 22 categories. This is already the answer to the question if *Tanzania* is a global keyword.

These six discussed keywords seem to be the most eye-catching words for laymen. However, the list contains as well some items, which are not interpretable at first glance if one is not familiar with the subject. These items are *Moi*, placed at the 8th rank, *hon* (rank 10), *KANU* (14) and *MFI* (18). To investigate the meanings of these words, one can either use the *Concordance* or the *File View* tool. These reveal that *Moi* is the proper name of Kenya's president at that time. *Moi* was governing the country for a long period, from 1978 to 2002. It occurs, as one might suspect, 202 times, of all 288 entries, in the broadcast news and splash categories. But not all entries belong to the collocate *president*. Since several institutions have been named after him, *Moi* also refers to the leading university of Kenya, Moi University, as well as e.g. the Moi Airbase.

The next three items are all abbreviations. Since the settings were adjusted to treat all data as lower case, distinctions can not be drawn anymore, which has especially an effect on the keyword *hon*. *Hon* is the abbreviation of "honorary" in lower case, and of "Honourable" in upper case. Nevertheless, both imply the same meaning and refer to members of the parliament. The majority of the tokens is written in lower case, and of

340 entries, 332 hits appear in the Hansards category. The remaining 8 occurrences are distributed over three further categories. This is the first item to be considered a local keyword, and it contradicts the assumption that dominant keywords are spread globally across the corpus.

The next item to be considered, ranking at the 14th position, is the abbreviation *KANU*, which stands for the "Kenya African National Union", a leading party in Kenya. It is distributed similar to *Moi*, occurring more in the broadcast news and splash categories, but also in the personal columns category as well as administrative writing.

The next keyword, *MFI*, seems to be, at first sight, an abbreviation of some Kenyan party or institution. But through analysis it becomes apparent that it is the abbreviation of "Marked for Identification" used in cross-examinations. This becomes apparent in the distribution of the keyword, because it appears only in the category of cross-examinations. This is an additional, even stronger contradiction to my assumption. The keyword is placed at rank 18, although it occurs only in one single category.

Nevertheless, *MFI* is not the only English item. The keyword list contains a whole bunch of common English words, which tend to be no indicators about the culture of Kenya. Yet, they can be indicators of the texts' aboutness. Among the list of the top 30 keywords are 20 English items, which will not be all presented. I will just pick out some outstanding ones. Most of these keywords appear in more than 20 categories, but there are also items such as *speaker* (rank 4) or *crops* (26) which occur in only 12. The keyword *learner* (27) even appears in only 3 categories, and of its 46 hits in the whole corpus, 43 occur just in the single category of exams. Thus, it should be questioned why *learner* is detected as keyword. The case of *members* (22) and *minister* (23) is different. Both have most hits in the Hansards category, which the reference corpus does not contain.

Outstandingness can also be caused by spelling variations, which is the case for *programme* (13). The reference corpus contains the spelling variation as well, but the spelling *program* is far more preferred. The last item, which will be mentioned in detail, is *women* (15). It occurs in 27 out of 28 categories, and in most of these categories (academic social sciences and humanities, broadcast talks and conversations have the most entries) *women* is referred to the role of women and their emancipation. This is an indicator of the difficult situation of women in Kenya. The keywords *women* and *crops* are the only English items, which represent in some way Kenya's culture.

Therefore, we should examine why the analysis tool detected these 20 English items as outstanding words. This can be done by taking a closer look at the frequency percentages of those items in the reference corpus. The corpus contains over 8.3 million words, which is ten times the size of the Kenyan part of the ICE-EA. The table below displays some of these English items with their number of tokens in the target and in the reference corpus as well as their percentages. The lemma word forms have been included as well, which means that the number of tokens of e.g. *problem* includes the number of *problems*.

Table 10: Frequency Percentages of 12 English Keywords from the Kenyan Part of the ICE- EA Compared with their Frequency Percentages of the Reference Corpus

Keyword	Kenyan Part of ICE-EA 779,885 Tokens		Reference Corpus 8,304,328 Tokens		Difference
	Tokens	Percentage	Tokens	Percentage	Percentage
accused	586	0.075%	204	0.003%	0.072%
children	1,392	0.178%	4,836	0.058%	0.12%
crops	226	0.029%	325	0.004%	0.025%
development	794	0.102%	1,978	0.024%	0.078%
learner	149	0.019%	44	0.0005%	0.019%
members	677	0.087%	3,006	0.036%	0.051%
minister	655	0.084%	2,072	0.025%	0.059%
people	2,065	0.265%	10,203	0.123%	0.142%
problem	812	0.104%	3,904	0.047%	0.057%
programme	575	0.074%	1,105	0.013%	0.061%
speaker	791	0.101%	519	0.006%	0.095%
women	1,085	0.139%	3,341	0.04%	0.099%

Table 10 shows that all keywords have a higher frequency percentage in the target corpus than in the reference corpus, even if the difference is very small. The token numbers of the items *accused*, *learner* and *speaker* are even higher in the target corpus than in the reference corpus, although the target corpus contains much fewer words. Therefore, it is not surprising that these words were extracted as outstanding. The table demonstrates a further issue, that keyness calculation is based on a complex mathematical procedure. If it was based only on the percentage differences, then *people* would be highest in ranking among these 12 items, followed by *children* and *women*. Yet, if we

look at the keyword list, we recognize that *people* is preceded by *accused* (ranking highest), *speaker* and *children*. This reveals that statistical calculations do not rely only on frequencies, but also on probabilities as well as the distribution of words. This is why the distinction between global and local keywords is important. A good method should determine keywords which are distributed across the corpus. The case of *learner* reveals that the chi-squared method seems to partially lack of this feature. The log-likelihood measure proceeds slightly better, ranking the keyword only at 43rd place.

Table 11: Classification into Global and Local Keywords

Keyword	Number of Categories	Classification
Kenya	28	global
Africa	27	global
Nairobi	24	global
speaker	12	local
accused	16	local
children	27	global
shillings	18	global
Moi	19	global
country	27	global
hon	4	local
people	28	global
areas	27	global
programme	23	global
KANU	11	local
women	27	global
language	22	global
development	26	global
MFI	1	local
Kiswahili	15	local
problem	26	global
environment	24	global
members	27	global
minister	22	global
teacher	17	global
officer	20	global

crops	12	local
learner	3	local
services	26	global
students	22	global
Tanzania	22	global

The table above displays all 30 keywords and the number of categories they appear in as well as their classification. For most of the keywords the classification can be easily drawn. But in some cases the decision is not only based on the number of text categories, but also on the concentration of hits in these categories. *Kiswahili* has been especially a borderline case. It occurs in 15 categories, which is more than the half. But its hits are concentrated only in three of them: class lessons, academic humanities and popular humanities. Therefore, I chose to classify it as a local keyword. The item *accused* appears in 16 categories, but of its 586 concordance hits, 92% (540 hits) occur in the single category cross-examinations. The same is true for *speaker*, because 85% of all entries occur in the Hansards category. Local keywords seem to be concentrated in these two categories, such as *speaker*, *accused*, *hon* and *MFI*. The last item is even the only keyword which appears in just one category. The reason why these two categories produce the most local keywords is that they have no equivalents in the reference corpus. If I had taken other varieties of the ICE corpora, I guess it would not have been different, since these categories were only included in the ICE-EA to compensate the lack of spoken material.

I think this classification is essential, to investigate if the statistical measures work well. Among the first 30 keywords of the Kenyan part, which I consider to be dominant since they are ranked at the top of the list, are 8 local keywords, of which 4 are caused by the reason I explained above. If we ignore these 4 items, then only 4 keywords are considered to be local. This means that the assumption I made in chapter 4.1 can be applied for the most part. This implies further that the chi-squared statistic works fairly well in extracting proper keywords.

Nevertheless, of the keywords I analysed so far, I would only consider 9 keywords to be indicators of Kenya and its culture. These are the six items I analysed at the beginning of this section: *Kenya*, *Africa*, *Nairobi*, *shillings*, *Kiswahili* and *Tanzania*, as well as the proper name *Moi*, referring to Kenya's president, the abbreviation *KANU*, which stands for one leading party of Kenya, and the English word *crops*, which is the basic food of the people. The item *women* might be also grouped among them as it

mostly refers to the role of women in Kenya. But analysing only the first 30 keywords is a too restricted horizon if we want to investigate culture specific words. Thus, we have to expand the scope to at least 120 detected keywords.

6.1.2 A Global Analysis of the First 120 Kenyan Keywords

It would be of no sense to analyse the first 120 keywords the way I did in the previous section. Therefore, I tried to group them together, in order to detect to which areas the keywords refer.

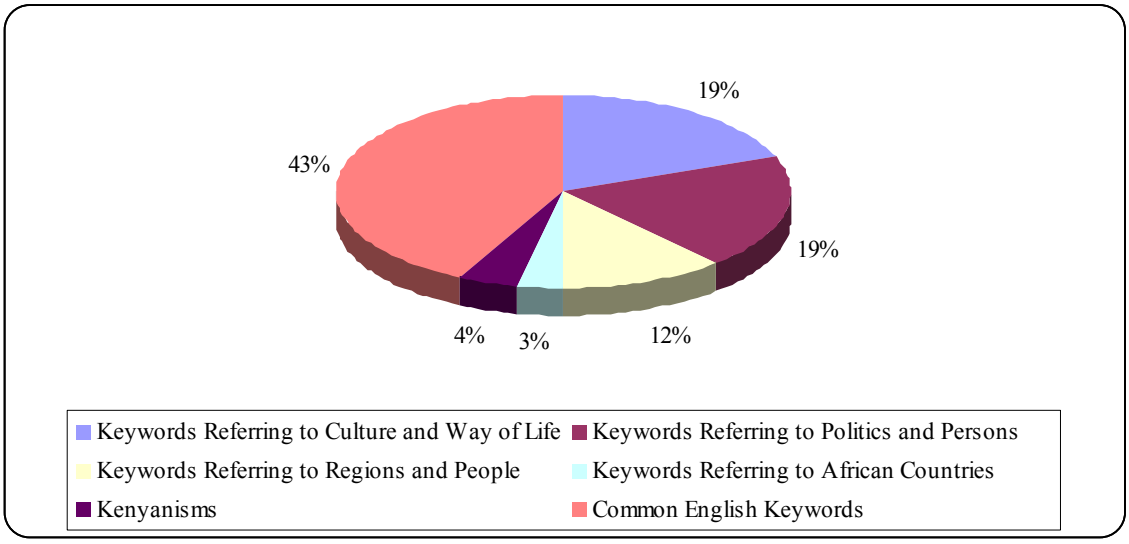


Figure 7: Distribution of Kenyan Keywords across Six Different Areas

The different areas or categories, I have chosen, can be seen in the legend of the diagram above. The figure reveals that the largest area is common English words with 43%, followed by keywords referring to culture and way of life (19%), keywords referring to politics and persons (19%), keywords referring to regions and people (12%) as well as African countries (3%). Kenyanisms make up only 4% of all 120 keywords. Which keywords have been categorized in the specific areas is given in the table below.

Table 12: Classification of Kenyan Keywords into First Five Areas

Culture and Way of Life	Politics and Persons	Regions and People	African Countries	Kenyanisms
shillings	*speaker*	*Kenya*	*Africa*	*wananchi*
Kiswahili	*Arap Moi*	*Nairobi*	*Tanzania*	*matatu*
crops	*hon*	*Mombasa*	*Uganda*	*harambee*
farmers	*KANU*	*Turkana*	*Somalia*	*orkoiik*
maize	*members*	*Kisumu*		*kali*
pest	*minister*	*Kikuyu*		
Swahili	*president*	*Kitui*		
fuelwood	*Kenyatta*	*Kipsigis*		
livestock	*government*	*Nakuru*		
disease	*Imanyara*	*Eldoret*		
traditional	*Kiliku*	*Luo*		
malaria	*Odinga*	*Machakos*		
lingua	*Mungai*	*Kakamega*		
KBC	*ministry*	*Nyanza*		
health	*parastatal*			
oral	*colonial*			
pesticides	*national*			
arid	*leaders*			
agricultural	*Shikuku*			
aphids	*international*			
narratives	*parliament*			
tribalism	*Kariuki*			
pidgin				

The keywords are listed according to the order they appear in the keyword list. The categorization may not be perfect, since some of these keywords could be classified to two or more categories, and some people may have different views about the meanings. Therefore, I want to pick out and explain a few keywords, why they were grouped into this category and which important role they play for Kenya and its people. Some items have already been explained, and are thus not mentioned in detail again here.

Among the first 120 keywords, geographical terms seem to be prominent. Besides the continent *Africa* (rank 2), three neighbouring countries are extracted as keywords:

Tanzania (30), *Uganda* (78), and *Somalia* (105). Keywords referring to regions and people turn up several times among the top 120 words: the country *Kenya* itself, as well as cities and towns such as *Nairobi* (3), *Mombasa* (40), *Kisumu* (57) – the capital of *Nyanza* (120) province, *Kitui* (70), *Nakuru* (85), *Eldoret* (95), *Machakos* (106) and *Kakamega* (110) appear. The keywords *Turkana* (54), *Kikuyu* (66) and *Luo* (101) refer not only to regions, but also to the people living there and their language.

The area politics and persons contains English words such as *government* (38), *ministry* (76), *international* (11), *parliament* (112) or *colonial* (82), referring to the history of Kenya. Two keywords have been grouped together, since they belong to one and the same person. *Arap* (68) is part of the name of President *Moi* (8). As the *Collocates* tool reveals, it occurs 44 times out of 63 along with *Moi*. The keyword *president* (33) appears even 208 times together with *Moi*, of which 33 times belong to the full name *President Daniel Arap Moi*. Thus, among the 21 keywords of that category, three are linked to each other. But president could also be linked to *Kenyatta* (36), which governed the country from 1964 – 1978. But as the analysis tool reveals *Kenyatta* comes up only a few times with *president*. This is due to the reason that many things were named after him, such as the Kenyatta University, Kenyatta Hospital, Kenyatta National Airport or Kenyatta Darts Festival. This reflects the importance of Kenyatta to the Kenyan people. The category contains several more proper names. When conducting this study I decided to exclude only fictitious proper names. This proved to be a right decision. The proper names, which come up among the first top 120 keywords, mostly refer to politicians, which were topical at the beginning of the nineties when the material was sampled. Examples are *Kiliku* (69), *Odinga* (72), *Mungai* (75) and *Shikuku* (104), but also the renowned lawyer *Imanyara* (67). The proper name *Kariuki* (116) refers to several people.

The next category contains keywords referring to culture and way of life. Lexical items such as *Kiswahili* (19), *Swahili* (52), *lingua* (franca) (65) as well as *pidgin* (115) refer to the language, and terms such as *shillings* (7), *traditional* (61), *KBC* (74) – *Kenya Broadcasting Cooperation*, and *tribalism* (113) reflect the culture of Kenya. The keywords *oral* (81) and *narratives* (102) have been classified as a cultural term, since they appear 34 times as *oral narratives*, and *oral* occurs as well 62 times as *oral literature* out of its 136 overall appearances. The way of life is reflected by words such as *crops* (26), *farmers* (32), *maize* (39), *fuelwood* (53), *livestock* (59) and *agricultural* (98). And people have to cope with agricultural problems such as *pest* (41), *pesticides*

(88), *arid* (94) and *aphids* (99). Keywords referring directly to people are *disease* (60), *malaria* (63) and *health* (79). Through the grouping of these keywords, becomes apparent that many are linked with each other.

Originally, I assumed that a good proportion of Kenyanisms would be detected by the tool, but unfortunately, only five appeared to be outstanding, at least among the first 120 keywords. These Kenyanisms are *wananchi* (47), meaning "people", *matatu* (49), *harambee* (83), *orkoiik* (96), the plural form of orkoiyot, and *kali* (118). Using clusters and collocates, we determine that *matatu* comes along with driver(s) 15 times, which confirms that it means "share taxi". The keyword *kali* means "hot", but occurring together with *jua*, "sun", it is assigned to people working under the tropical sun, the so-called *Jua Kali*. *Orkoiik* are the "ritual leaders" or "witchdoctors" of the Kipsigis and Nandi people. The keyword is mentioned 20 times together with *Kipsigis* and 6 times with *Nandi*. The probably most important Kenyanism is *harambee*, literally meaning "let us pull together". It has become a way of life and traditional custom in Kenya, but was adopted by *Kenyatta* to be the official slogan for Kenya, and is therefore also displayed on the Kenyan coat of arms. The category of Kenyanisms shows that no clear cut lines can be drawn between these areas, since all could be classified as cultural terms as well, and *harambee* is even a political term. The keywords *orkoiik* and *Kipsigis* are linked to each other, as well as *Kenyatta* and *harambee*.

The keywords discussed so far, reflect Kenya's people and culture. But nevertheless, they only make up 57% of the top 120 keywords. The remaining 43% belong to the category of common English keywords, which are shown in the table below.

Table 13: Common English Keywords of the Kenyan Corpus

Common English Words				
accused	*environment*	*school*	*receipt*	*Mfl*
children	*teacher*	*nation*	*stamp*	*population*
country	*officer*	*society*	*security*	*house*
people	*learner*	*deputy*	*deceased*	*arrested*
areas	*services*	*drugs*	*person*	*land*
programme	*students*	*money*	*statement*	*activities*
women	*district*	*organisation*	*province*	*information*
language	*issue*	*police*	*baby*	*identify*

development	*education*	*office*	*question*	*faithfully*
MFI	*assistant*	*story*		
problem	*discourse*	*food*		

Most of these words can be linked to the other five categories as well. Keywords such as *women* (15), *language* (16), *development* (17), *environment* (21), *district* (31), *education* (35), *drugs* (48), *food* (62) and *province* (89) might play a larger role in Kenya than in the countries, of which the reference corpora texts were sampled, but I do not consider them to be as typical as the keywords discussed above.

Reasonable 69 words, more than half of the first 120 keywords, can be said to be good indicators of Kenya, its culture and its people, and are therefore regarded as typical lexical items. Several more tend to reflect the situation in Kenya in some way or the other. It can be argued that looking only at the first 120 words, is a too narrow horizon, and probably much more Kenyanisms and keywords could be detected if regarding the entire keyword list. But analysis tools are constructed to extract words which seem to be outstanding, and some words show indeed that they are outstanding in contrast to the reference corpora, although one would not consider them to be typical. My aim was as well to examine if the analysis tool works properly in retrieving typical lexical items. If my aim was only to detect Kenyanisms, I could have used the frequency lists instead.

However, if the Tanzanian keyword results show the same distribution and characteristics will be investigated in the next section.

6.2 Discussion of the Tanzanian Keyword Results

6.2.1 A Detailed Analysis of the First Thirty Tanzanian Keywords

The analysis of the Tanzanian keywords will be carried out according to the procedure I used for Kenya, but not as detailed since most of the aspects have already been explained. The list of the first 30 keywords can be seen in table 7 on page 35.

Regarding the first 30 keywords, we recognize that the list shows striking similarities to the Kenyan keyword list. *Tanzania* is the most typical keyword, ranking at the top of the list. It is followed by *Africa* in second position and *Dar* in third place. Remember that *Dar* represents the city Dar es Salaam. The Tanzanian list seems to contain more typical words at the beginning, since *Zanzibar* is ranked at 5th place, followed by *Kiswahili* (6) and *Ndugu* (7). The next words, which come to my notice, are the

abbreviation *CCM* (17), *Arusha* (22), and *Kenya* (29). The words will be analysed according to this order.

The most prominent keyword is by no means *Tanzania*. The partial concordance plot in figure 7 shows its distribution across the Tanzanian corpus.

Figure 8: Concordance Plots of *Tanzania*

The files have not been chosen arbitrary, but to allow a comparison to the distribution of *Kenya*. The parliamentary file (Hansards) of figure 6 has been replaced by the broadcast-interview file. This was done due to the reason that the spoken part of the Tanzanian corpus contains only 7 categories (in comparison to 10 in the Kenyan corpus) missing the Hansards and cross-examinations categories, as well as the school broadcast category. Because Tanzania is missing the political and judicial files in the spoken as written part, it contains a further category, legal presentations, in the written part, but it as well lacks of social letters. To compensate this shortage, a general category had been included in the popular writing section. If the different categorization has effects on the keyword outcomes will be analysed later.

The distribution of *Tanzania* is quite similar to that of Kenya. The most entries, 150, appear in hit file 3 representing broadcast interviews, followed by 117 hits in the broadcast news category. The conversation part contains only 26 occurrences. The hit files 8 (business letters) and 26 (splash) contain an almost identical number of entries, whereas that of the popular sciences category is half as much. A detailed distribution is given in the table below.

Table 14: Distribution of Tanzania across the Tanzanian Corpus

		Categories	Distribution of *Kenya* in %
Spoken Part 39.6%	Private	Conversation	1.7%
	Public 13.7%	Broadcast Discussions	3.5%
		Broadcast Interviews	9.9%
		Class Lessons	0.3%
	Scripted 24.2%	Broadcast News	7.7%
		Broadcast Talks	9%
		Non-Broadcast	7.5%
Written Part 60.4%	Letters	Business Letters	4.4%
	Legal Presentations	Judgement Texts	1.3%
	Student Writing 12.1%	Essays	7.2%
		Exams	4.9%
	Academic 11.9%	Humanities	3.6%
		Natural Sciences	3.5%
		Social Sciences	1.6%
		Technology	3.2%
	Popular 15.6%	Humanities	2.2%
		Natural Sciences	1.9%
		Social Sciences	2.2%
		Technology	6.1%
		General	3.2%
	Reportage 8.1%	Reportage-Feature	3.4%
		Splash	4.7%
	Persuasive 4%	Institutional	3.1%
		Personal Columns	0.9%
	Instructional	Administrative Writing	2.7%
	Creative	Novels/Stories	0.3%

The table shows that *Tanzania* occurs more in the "public sector" as the percentages of broadcast news and broadcast talks reveal. Yet, it has a slightly lower number, 4.7%,

in the splash category as *Kenya*, 6.8%. The greatest distinction between *Kenya* and *Tanzania* can be seen in the student writing section. Whereas Kenya appears only 0.9% in the essay and 0.4% in the exam category, *Tanzania* outnumbers them both with incredible 7.2% and 4.9%. However, the table reveals that the keyword occurs in each ICE-EA (Tanzania) category, and is therefore clearly classified as a global keyword.

The next keyword under consideration is *Africa*, ranking at second place. Its distribution across the corpus is almost similar to that in the Kenyan corpus, although it does not appear in each category. The only significant difference is that it occurs more in the broadcast news (87 hits) as well as broadcast talks (57) and in the class lessons (33) category, compared with the hits of *Africa* in the other corpus.

In third place ranks the keyword *Dar*, representing the largest and most important city of Tanzania, Dar es Salaam, which is also the seat of the government. The importance of the city is reflected in the distribution of the concordance hits. It mostly appears in broadcast news with 115 entries, in the splash category with 50 hits, and is mentioned as well in business letters 35 times. It is spread evenly across the corpus, missing only the class lessons category.

It is interesting that *Zanzibar* is ranked at 5th position. It is mostly mentioned in conversations, accounting to 80 hits, and in the non-broadcast category with 74 hits. Although it occurs only in 18 out of 26 categories, the keyword is distributed evenly across these. I would have suspected *Zanzibar* to occur more in the public sections, such as news, since Zanzibar is part of Tanzania. But maybe, it is not an interesting issue for the Tanzanian people to talk about.

The keyword is followed by *Kiswahili* at position 6, ranking 13 places higher than in the Kenyan corpus. This is due to the fact, that Kiswahili plays a larger role as lingua franca in Tanzania than in Kenya. The distribution of *Kiswahili* equals that in the Kenyan corpus only in academic humanities texts, although it has twice as much hits. The greatest divergence can be seen in the class lessons category, with 94 hits in the Kenyan corpus, in contrast to one hit in the Tanzanian one, as well as in the student writings section with a reversed distribution of 70 entries in the essay and 114 in the exam categories in the Tanzanian corpus, in comparison to only 5 hits in the Kenyan one. The term *Swahili* is used slightly more, with 90 tokens, than in the Kenyan corpus, but it is still outnumbered by *Kiswahili* (296 tokens), and therefore ranking only at 43.

The Tanzanian term *Ndugu* (rank 7) is the highest ranking Tanzanianism, as Schmied calls it (cf. Schmied 2004: 254). It is originally a Kiswahili term, meaning

"brother, sister, friend and family", but is associated nowadays with the political term "comrade". Although it appears only in 14 categories, I would consider it to be a global keyword, since its distribution is evenly spread across these categories, with most entries in popular general texts (32), as well as in the reportage section, with 52 hits in the feature and 97 hits in the splash category.

The abbreviation *CCM* comes in 17th position and stands for the Tanzanian political party "Chama Cha Mapinduzi". It appears mostly in the administrative writing (48) and splash (55) category. Compared to the distribution of the Kenyan party KANU, it enormously differs from the entries of *KANU* in the broadcast news category. Whereas *KANU* appears 99 times in this category, *CCM* occurs only 5 times. But the Tanzanian party is spread over more categories, and can thus be seen as a global keyword.

The keyword *Arusha*, which is placed at the 22nd rank, refers to several meanings, the region Arusha, as well as its capital, which is named the same way, and the Arusha National Park. It is distributed evenly over 21 categories, without sticking out in one single category.

The next keyword to be considered is Tanzania's neighbouring country *Kenya*. It is placed relatively low (rank 28) in the list of the first 30 keywords, as it is the case for *Tanzania* within the Kenyan keyword list. The entries are spread evenly across 20 categories. The only categories which seem to be more prominent are broadcast news with 21 hits and the essay category with 17. The distribution of *Kenya* in the Tanzanian corpus is similar to that of *Tanzania* in the Kenyan corpus.

The so far discussed keywords are the most typical items of the list, at least at first glance. Their categorization into global or local keywords can be seen in table 13. But before I will discuss this issue, I want to take a closer look at some English items occurring in the list. The keyword list contains 21 English words, of which only a few outstanding ones will be investigated.

The first keyword, I want to mention, is ranked at 8th position. *Appellant* appears in only 5 categories, and of its 213 entries 82% (175 hits) occur in the single category judgement texts. This is the first item, which can be clearly categorized as local keyword. The second word to be considered is *village* (rank 19). The item does not only refer to the meaning we usually imply this word with, but also to the organisation "International Village of Science and Technology". Most entries of the keyword appear in the public spoken part, accounting to 138 hits. The word *village* mostly refers to the traditional settlements of Tanzanian people, but they reason why it appears more often in

the Tanzanian corpus than in the Kenyan corpus (ranking very low at place 1268), is that the organisation is not mentioned in the Kenyan corpus, and therefore, seems to be of more significance to Tanzania.

The political term *parastatal* ranks at the 23rd place. It occurs for the most part (43 hits) in the non-broadcast category. If using the *Clusters* tool, it becomes apparent that *parastatal* is mostly associated with the items *organisation* and *sector*. Although it appears in 15 categories, I consider the item to be a local keyword, since 60% of the hits appear in a single category, and the remaining entries are scarcely scattered over 14 categories. The last keyword to be considered is *women* ranking at the end of the list. It appears in 23 categories, with most entries in broadcast discussions (83) and talks (192), as well as in academic (36) and popular (80) social sciences. The item refers mostly to the role of women in Tanzania, as it is the case for the Kenyan corpus, although the issue seems to be more prominent in Kenya, since it ranks at position 15 in the Kenyan keyword list with a much higher frequency.

As I did with the Kenyan keywords, the first 30 keywords of the Tanzanian corpus will also be classified into global and local keywords.

Table 15: Classification into Global and Local Keywords

Keyword	Number of Categories	Classification
Tanzania	26	global
Africa	24	global
Dar	25	global
country	26	global
Zanzibar	18	global
Kiswahili	12	local
Ndugu	14	global
appellant	5	local
problem	26	global
programme	23	global
development	25	global
areas	25	global
region	25	global
government	26	global
organisation	20	global

environment	23	global
CCM	16	global
language	18	global
village	24	global
people	26	global
education	23	global
Arusha	21	global
parastatal	15	local
economic	24	global
services	24	global
accused	13	local
nation	24	global
Kenya	20	global
issue	23	global
women	23	global

For most of the keywords the classification seems to be obvious. Nevertheless, there are four borderline cases: *CCM*, *parastatal*, *accused* and *Kiswahili*. The first two have already been explained. Although *accused* occurs in 13 categories, 76% appear in only three of them: judgement texts, essays and exams. Therefore, I classified it as a local keyword. The situation of *Kiswahili* is similar to that in the Kenyan corpus. It is mostly concentrated in the student writings section (184 hits), and in academic humanities texts with 63 hits, which is the reason why I chose to consider it to be a local keyword. The Tanzanian corpus contains only 5 local keywords (in contrast to 8 in the Kenyan corpus), of which 4 of them are borderline cases, and one of them, the only clear local keyword *appellant*, is caused by the additional text category of judgements. Thus, we can claim that my assumption proves to be true for most of the Tanzanian results as well.

The different text categorization of the Tanzanian corpus had the effect that the political terms, which were prominent in the Hansards category of the Kenyan corpus, *speaker* and *hon* do not appear at all in the Tanzanian corpus, and the items *members* and *minister* are ranked much lower in the keyword list at positions 36 and 71. On the other hand, the Tanzanian keyword *appellant* does not occur in the Kenyan corpus, which does not contain a legal presentations category.

We can say that the chi-squared method produced fairly well results. More than one third of the keyword list can be said to be indicators of Tanzania and its culture: *Tanzania*, *Africa*, *Dar*, *Zanzibar*, *Kiswahili*, *Ndugu*, *CCM*, *village*, *Arusha*, *Kenya* and

women, the first six even ranking among the top ten. If the first 120 keywords are good indicators of Tanzania as well, will be examined in the next section.

6.2.2 A Global Analysis of the First 120 Tanzanian Keywords

The list will be analysed the same way as the list of the Kenyan results. The first figure shows the distribution of the top 120 keywords of the Tanzanian corpus across the categories, which were introduced in section 6.1.2.

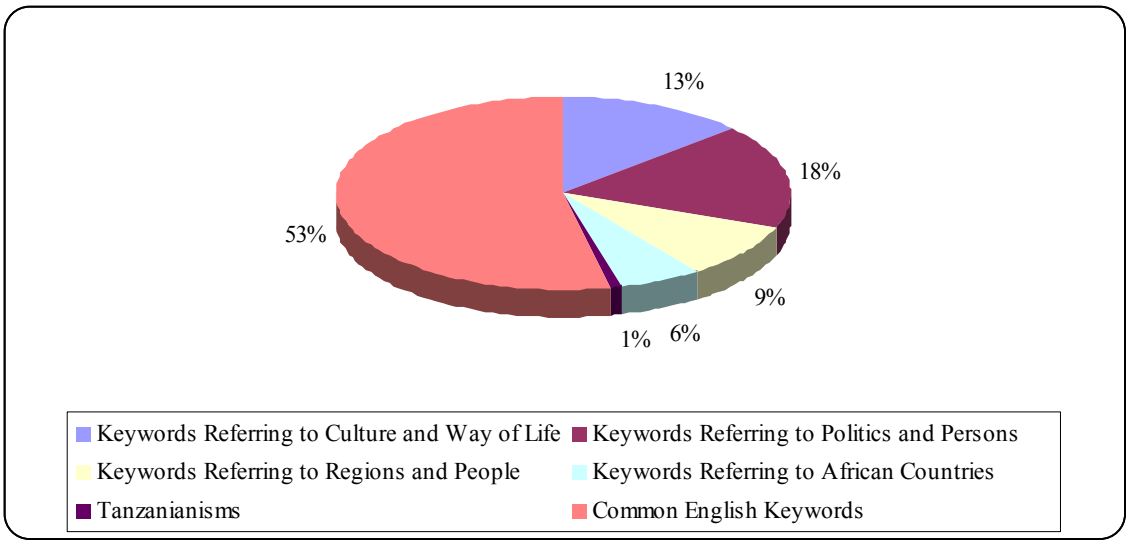


Figure 9: Distribution of Tanzanian Keywords across Six Different Areas

The diagram reveals that more than half of the keywords (53%) are common English words, 18% refer to politics and people, 13 % to culture and way of life, 9% to regions and people, and 6% to African countries. Tanzanianisms make up only 1% of all 120 keywords. In comparison to the Kenyan results, the Tanzanian ones contain 10% more common English words, 3% more African countries, but in all other categories they contain less keywords. The classification of the different keywords can be seen in the two tables below.

Table 16: Classification of Tanzanian Keywords into First Five Areas

Culture and Way of Life	Politics and Persons	Regions and People	African Countries	Kenyanisms
Kiswahili	*government*	*Tanzania*	*Africa*	*Ndugu*
village	*CCM*	*Dar*	*Kenya*	
health	*parastatal*	*Zanzibar*	*Uganda*	
crops	*nation*	*Arusha*	*Rwanda*	
UDSM	*national*	*Tanganyika*	*Zambia*	
leprosy	*international*	*Mwanza*	*Mozambique*	
shillings	*members*	*Tanga*	*Nigeria*	
Swahili	*Mwinyi*	*Dodoma*		
HIV	*Mrema*	*Shinyanga*		
disease	*democracy*	*Moshi*		
AIDS	*president*	*Morogoro*		
witchcraft	*political*			
AMREF	*rights*			
pesticides	*minister*			
malaria	*united*			
rape	*ministry*			
	party			
	trade			
	republic			
	independence			
	constitution			

Table 17: Common English Keywords of the Tanzanian Corpus

Common English Words				
country	*services*	*students*	*primary*	*gender*
appellant	*accused*	*rural*	*process*	*social*
problem	*issue*	*science*	*code*	*situation*
programme	*women*	*material*	*participants*	*conference*
development	*institutions*	*food*	*marketing*	*human*
areas	*children*	*land*	*respondent*	*secondary*

region	*cooperative*	*offence*	*school*	*person*
organisation	*activities*	*sector*	*seminar*	*local*
environment	*technology*	*district*	*court*	*management*
language	*workshop*	*resources*	*developing*	*research*
people	*project*	*production*	*rodent*	*thousand*
education	*teacher*	*university*	*foreign*	*law*
economic	*population*	*teaching*	*registration*	

I decided to place this last table right behind the table of the other categories, since the same characteristics of the common English words of the Kenyan results can be applied to the Tanzanian results, and will, therefore, not be mentioned again, as well as the keywords shown in table 16, which have already been explained in the previous section.

Unfortunately, this table is not as interesting as table 12, as the percentage figures already implied. In contrast to the Kenyan results, the Tanzanian list contains 5 neighbouring countries: *Kenya* (rank 28), *Uganda* (41), *Rwanda* (42), *Zambia* (78), *Mozambique* (99), as well as *Nigeria* (100) and the continent *Africa* (2). The area of regions and people only applies to regions, since none of the regions' names refer to their people as well. The category contains the country *Tanzania* (1) itself, which was formed of the independent states *Zanzibar* (5) and *Tanganyika* (55) in 1964. It is interesting that *Zanzibar* is ranked relatively high at 5th place, in contrast to *Tanganyika*. Other keywords referring to cities are *Moshi* (102), as well as cities including their regions, *Dar es Salaam* (3), *Arusha* (22), *Mwanza* (57), *Tanga* (74), *Dodoma* (80), *Shinyanga* (91) and *Morogoro* (114). *Dodoma* is the real capital of Tanzania, but is preceded by four cities, which gives an impression of the relatively unimportance of the city, in contrast to *Dar es Salaam*.

The political area contains far less politicians in contrast to the political area of the Kenyan results. Only two persons appear among the top 120 keywords: *Mwinyi* (47), who was the president from 1985 – 1995, and the politician *Mrema* (49), who was Home Affairs Minister in the early – mid nineties. President *Mwinyi* seems to play not such an important role in Tanzania as President *Moi* in Kenya, as the rankings of these two reveal. *Speaker*, *members*, *minister* and *president* have been prominent among the Kenyan keywords, which is probably the result of the Hansards category. The detected Tanzanian political terms are rather different, such as *democracy* (53), *rights* (69), *united* (72), *republic* (111), *independence* (117) and *constitution* (120). These items do not play a crucial role in Kenyan politics, as the ranking shows: *democracy* (2620), *rights* (459),

united (498), *republic* (2345), *independence* (190) and *constitution* (2620). Further analysis of the Tanzanian results reveals that *republic* is collocated with *united* 64 times, with *Tanzania* 36 times and with *constitution* 10 times. The collocation *United Republic of Tanzania* even appears 31 times.

The keywords referring to Tanzanian culture and way of life are rather similar to those of Kenya. *Kiswahili* (6) and *Swahili* (43) referring to the language; *village* (19), *UDSM* (38) – *University of Dar es Salaam*, *shillings* (40), *witchcraft* (66) and *AMREF* (67) – *African National Research Foundation* referring to the culture, as well as *crops* (33), which is as an important basic food. The Tanzanian results contain also problems people have to cope with: *leprosy* (39), *HIV* (48), *disease* (58), *AIDS* (63), *pesticides* (68), *malaria* (92), *rape* (93). The only Tanzanianism among the first 120 keywords is *Ndugu* (7), meaning "comrade", as already mentioned in the previous section, which can be classified as a political term as well.

I was in some way surprised to receive only one Tanzanianism. Not even half of the first 120 keywords can be said to be really typical for Tanzania. Only 56 keywords, making up 47%, can be regarded as indicators of the Tanzanian culture, politics, people and life.

6.3 Comparison of the Kenyan and Tanzanian Keywords

At the beginning, when analysing the first 30 keywords, it seemed that the results extracted from the Tanzanian part of the ICE-EA were slightly better than the results retrieved from the Kenyan part of the ICE-EA. But it turned out that the Kenyan corpus produced more typical lexical items than the Tanzanian one, regarding the list of 120 keywords. A comparison has already been made in the last two sections when analysing the Tanzanian results. This section is aimed at given a short summarizing overview of the similarities and differences between the Kenyan and Tanzanian keyword results, which have been detected in this chapter.

6.3.1 Similarities

In the following statements, the first keyword refers always to the Kenyan corpus and the second to the Tanzanian corpus.

- The first three keywords of both corpora refer to: the country itself, *Kenya* and *Tanzania*, to the continent *Africa*, and to the most important and largest city of the country, *Nairobi* and *Dar es Salaam*.
- Among the first 30 keywords both corpora contain the following words: *Africa*, *shillings*, *Kiswahili*, *country*, *people*, *areas*, *programme*, *development*, *language*, *women*, *problem*, *environment* and *accused*. These are more than one third, thus both corpora tend to deal with the same subjects.
- Among the first 30 keyword results, the Kenyan list contains 10 typical keywords and the Tanzanian list 11.
- Both corpora contain the name of the other country among the first 30 results.
- *Kenya* appears in the Kenyan corpus more in the public sector, as well as *Tanzania* in the Tanzanian corpus.
- The distribution of the keyword *Africa* is almost the same in both corpora.
- Both keyword lists contain the president's name and one important political party: *Moi* and *KANU* for Kenya, as well as *Mwinyi* and *CCM* for Tanzania.
- Both lists contain global as well as local keywords.
- Regarding the first 120 keywords, a whole bunch of words is contained in both corpora.
- The keyword lists of both corpora contain African countries, names of cities and regions, political terms, cultural terms and lots of common English words.
- The keywords of both lists refer to problems people have to cope with: *disease*, *malaria* and *pesticides*.

6.3.2 Differences

- Kiswahili is placed 13 ranks higher in the Tanzanian keyword list than in Kenyan.
- The differences among the first 30 keywords are the result of the different text sampling strategies: the Hansard and cross – examination categories (which are not included in the Tanzanian corpus) cause the existence of *speaker* and *hon* in the Kenyan corpus, whereas they do not appear at all in the Tanzanian corpus. They, furthermore, boost *members* and *minister* in the top 30 Kenyan list, whereas they are ranked much lower in the Tanzanian list.

- The category judgment texts (which is not included in the Kenyan corpus) causes the ranking of *appellant* at 8th place in the Tanzanian list, whereas the item does not appear at all in the Kenyan corpus.
- Among the first 30 keywords of the Tanzanian results are only 5 local keywords, whereas the Kenyan results contain 8 local keywords.
- In the Kenyan list of the first 120 keywords contains 5 Kenyanisms: *wananchi*, *matatu*, *harambee*, *orkoiik* and *kali*, whereas the Tanzanian list contains only the Tanzanianism *Ndugu*.
- The keywords referring to politics are quite different: *members*, *minister*, *hon*, *speaker* and *parliament* are prominent among the Kenyan keyword results. *Democracy*, *rights*, *united*, *party*, *republic*, *independence* and *constitution* among the Tanzanian keyword results.
- The Kenyan keyword list contains far more politicians (7), in contrast to the Tanzanian keyword list (2).
- The Tanzanian keywords refer to more diseases: *leprosy*, *HIV*, *AIDS* and *malaria*, whereas *malaria* is the only disease detected as Kenyan keyword.

7 Conclusion

The aim of this study was to detect and analyse typical lexical items of Kenyan and Tanzanian English. Typicality of lexical items is inextricably linked to the notion of keywords. Although some linguists claim that keywords can not be typical for a language, we can nevertheless state that they can be key to a culture. But, we should ever bear in mind that the keywords we receive are extracted from a corpus, not from a language or culture. Therefore, corpora should be always compiled in such a way that they represent the language, as well as the culture of the country where the language is spoken, as best as possible. Detection of keywords relies on two types of corpora, the target corpus, of which we want to extract the words, as well as one or more reference corpora, with which the target corpus is compared. These two corpora should always be compatible, since incompatibility can distort the results we are to obtain.

The target corpus of this study, the ICE-EA, and the reference corpus, compiled of six written corpora (Brown, LOB, Frown, FLOB, Kolhapur, ACE and WC) and four spoken (LLC, SEC, WSC and COLT) corpora, did not serve this criterion exactly. The written corpora of the reference corpus were assembled due to the same guidelines. But the East African corpus is part of the International Corpus of English and was thus compiled according to the guidelines of the ICE, which were established to ensure compatibility of all ICE corpora. The design of the ICE corpora is nevertheless based on the design of the Brown Corpus. But due to difficulties to match these guidelines exactly, the text categories of the ICE-EA had to be adjusted according to the linguistic situation in Kenya and Tanzania. Nevertheless, at least some compatibility between the target and reference corpus was given, and the proper results, obtained in the study, confirm the decision to use these reference corpora. Yet, the results might have been different if another reference corpus had been chosen, but I assume that the majority of the keywords would have been detected as well.

The choice of the reference corpus had not been the only problem of this study. The availability of multi-purpose corpus analysis tools, is very restricted, at least if one is not inclined to pay an expensive price. Fortunately, the *AntConc* toolkit, developed by Laurence Anthony, is available for download on the internet without any purchase. Yet, the tool will not do alone in retrieving appropriate results reflecting the culture under consideration. Several adjustments had to be made, which are explained in chapter 5.

First, frequency lists had been created to demonstrate the necessity of statistical procedures in order to detect proper keywords. There are several procedures, which can be applied, but only two were chosen and weighed up against each other. These were the chi-squared and the log-likelihood methods, which were offered in the *AntConc* toolkit. The chi-squared measure proved to produce slightly better results, although most linguists prefer the log-likelihood measure for various reasons. But the aim of this study was not to calculate exact keyness values (which tend to be higher for the chi-squared method), but to extract as many appropriate keywords as possible from the target corpus to investigate typicality of lexical items of Kenyan and Tanzanian English.

The choice of the statistical generation method was not the only adjustment, which had to be made, as first keyword results showed. A lemma list was created to assign the plural as well as some adjective forms to their related words. To exclude all fictitious proper names and the most English common words, stoplists have been used as well. After adjusting the settings correctly, I obtained the final results of this study, retrieved from the Kenyan as well as the Tanzanian part of the ICE-EA. Keyword lists are included in chapter 5.4 and appendices 1 - 3 for the Kenyan corpus, as well as in chapter 5.5 and appendices 4 - 6 for the Tanzanian corpus.

As the detailed analysis in chapter 6 has shown, the received results are good indicators of the culture in Kenya and Tanzania. The most prominent items, ranking at the top of the list, are the names of the countries itself: *Kenya* and *Tanzania*. The list of the first 30 Kenyan keywords contains 10 words, which I considered to be typical for Kenya and its culture. The list of the first 30 Tanzanian keywords contains even 11 typical words for Tanzania and its culture. A further analysis was done to classify the first 120 keywords according to different categories. The analysis showed that almost half of the Kenyan keywords could be grouped into the category common English words, and even more than the half of the Tanzanian keywords as well. But it also revealed that geographical items, referring to African countries as well as regions and cities within the two countries, are prominent among the keywords. The Kenyan list even contained several lexical items, such as *Luo* and *Turkana*, which refer to Kenya's people, as well as some proper names referring to politicians, who were topical at the time when the material was sampled. But the most important lexical items are those referring to the culture or way of life. In the Tanzanian keyword list, 16 typical items have been detected among the first 120 keywords, which are dominated by diseases people have to cope with in these countries, such as *malaria*, *leprosy*, *AIDS* and *HIV*.

Other cultural terms are *crops*, *Kiswahili* and *witchcraft*. The Kenyan keyword list contains even 23 typical lexical items for the culture and way of life in Kenya. Some of these items refer to agricultural problems, such as *pest*, *aphids*, *arid* and *pesticides*. But also traditional words as *tribalism* and *maize* come up.

Although a whole bunch of usual English words were detected as keywords, though several of them may refer to Kenya and Tanzania in some way or the other, the extracted results show that the statistical chi-squared method is indeed a trustworthy method, producing good indicators of culture. It retrieves surely not the best results ever, since the prerequisites of this study were not optimal, regarding the choice of the reference corpora. Further studies could be conducted, using different reference corpora or different statistical measures. A method, worth mentioning, is the TF-IDF measure, which seems to produce extremely good results. The results obtained could then be compared to the results gained in this study. But I am convinced that the majority of the keywords will be the same.

References

BOOKS

Aijmer, K., and B. Altenberg, eds. (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* Essex and New York: Longman Group UK Limited.

Butler, C. (1992). *Computers and Written Texts*. Oxford and Cambridge, USA: Basil Blackwell Ltd.

Charniak, E. (1993). *Statistical Language Learning*. London and Cambridge, USA: The MIT Press.

Facchinetti, R. (2007). *Corpus Linguistics 25 Years on*. Amsterdam and New York: Rodopi B.V.

Ghadessy, M., A. Henry, and R. Roseberry, eds. (2001). *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam and Philadelphia: John Benjamins B.V.

Garside, R., G. Leech, and G. Sampson, eds. (1987). *The Computational Analysis of English: A Corpus-Based Approach*. Essex and New York: Longman Group UK Limited.

Halliday, M., W. Teubert, C. Yallop, and A. Čermáková, eds. (2004). *Lexicology and Corpus Linguistics: An Introduction.* London and New York: Continuum.

Jurafsky, D., and J. Martin, eds. (2000). *Speech and Language Processing: An* Introduction to Natural Language Processing, Computational Linguistics and *Speech Recognition.* Upper Saddle River: Prentice-Hall Inc.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics.* London and New York: Addison Wesley Longman Limited.

Klavans, J., and P. Resnik, eds. (1996). *The Balancing: Combining Symbolic and Statistical Approaches to Language.* London and Cambridge, USA: The MIT Press.

Leitner, G. (1992). *New Directions in English Language Corpora: Methodology, Results, Software Development.* Berlin and New York: Mouton de Gruyter.

Mason, O. (2000). *Programming for Corpus Linguistics: How to Do Text Analysis with Java.* Edinburgh: Edinburgh University Press.

Nelson, G. (2006). "World Englishes and Corpora Studies." In: B. Kachru, Y. Kachru, and C. Nelson, eds. *The Handbook of World Englishes.* Malden, USA, Oxford, UK, and Carlton, Australia: Blackwell Publishing, 733-750.

Oakes, M. (1998). *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Ooi, V. (1998). *Computer Corpus Lexicography.* Edinburgh: Edinburgh University Press.

Pennington, M., and V. Stevens, eds. (1992). *Computers in Applied Linguistics: An International Perspective.* Clevedon, UK, Bristol, USA, and Adelaide, Australia: Multilingual Matters Ltd.

Rubagumya, C. (1990). *Language in Education in Africa*. Clevedon, UK, and Bristol,

USA: Multilingual Matters Ltd.

Schmied, J. (1991). *English in Africa: An Introduction.* Essex and New York: Longman Group UK Limited.

Schmied, J.(2006). "East African Englishes." In: B. Kachru, Y. Kachru, and C. Nelson, eds. *The Handbook of World Englishes.* Malden, USA, Oxford, UK, and Carlton, Australia: Blackwell Publishing, 188-202.

Scott, M., and G. Thompson, eds. (2001). *Patterns of Text: In Honour of Michael Hoey*. Amsterdam and Philadelphia: John Benjamins B.V.

Scott, M., and C. Tribble, eds. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education.* Amsterdam and Philadelphia: John Benjamins B.V.

Skandera, P. (2003). *Drawing a Map of Africa: Idiom in Kenyan English*. Tübingen: Gunter Narr Verlag Tübingen.

Souter, C. and E. Atwell (1993). *Corpus-Based Computational Linguistics*. Amsterdam and Atlanta: Rodopi B.V.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford and Malden, USA: Blackwell Publishers Ltd.

JOURNALS

Oakes, M., and M. Farrow (2007). "Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries." *Journal of The Association for Literary and Linguistic Computing and The Association for Computers and the Humanities*, 22.1, 85-99.

Schmied, J. (2004). "Cultural Discourse in the Corpus of East African English and beyond: Possibilities and Problems of Lexical and Collocational Research in a One Million-Word Corpus." *World Englishes*, 23.2, 251-260.

Scott, M. (1997). " The Right Word in the Right Place: Key Word Associates in Two Languages." AAA – Arbeiten aus Anglistik und Amerikanistik, 22.2, 239-252.

ARTICLES

Berber Sardinha, T. (1999). „Using Key Words in Text Analysis: practical aspects." *DIRECT Papers 42,* LAEL, Catholic University of São Paulo, 1-9.

Connell, L., and M. Ramscar (2001). " Using Distributional Measures to Model Typicality in Categorization." Proceedings of the *Twenty-Third Annual Conference of the Cognitive Science Society,* 1-4 August 2001, Edinburgh, 226-231.

Kilgarriff, A. (1996). " Which Words are Particularly Characteristics of a Text? A Survey of Statistical Approaches." Proceedings of the *Workshop on Language Engineering for Document Analysis and Recognition*, April 1996, Brighton, 33-40.

Rayson, P., and R. Garside (2000). "Comparing corpora using frequency profiling." Proceedings of the *Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics*

(ACL 2000). 1-8 October 2000, Hong Kong, 1-6.

Scott, M. (1997). "PC Analysis of Key Words - and Key Key Words." *System*, 25.1, 1-13.

INTERNET SOURCES

Hudson, D., and J. Schmied (1999). "Manual to Accompany the East African Component of the International Corpus of English." Retrieved 2008-02-20 <http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/eafrica/index.htm>

Laurence, A. (2007). "Users-Manual for AntConc." Retrieved 2008-02-20 <http://www.antlab.sci.waseda.ac.jp/software/README_antconc3.2.1.txt>

Laurence, A. (2008). "The AntConc Homepage." Retrieved 2008-08-26 <http://www.antlab.sci.waseda.ac.jp/antconc_index.html>

University of Montreal <http://www.iro.umontreal.ca/~nie/IFT6255/common_words.> Retrieved July 22, 2008

ICE Project

ICE Text Categories Table <http://www.ucl.ac.uk/english-usage/ice/textcats.htm>, Retrieved April 24, 2008

ICE-EA Text Categories Table <http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/studentprojects/ice /icedesign2.htm> Retrieved April 24, 2008

Powerpoint Presentations

Powerpoint on Investigating Keyness, Conference on Keyness in Text, Siena, June 2007. Retrieved 2008-26-06 <http://www.lexically.net/downloads/corpus_linguistics/Problems%20in%20Investigating%20Keyness%20Siena%202007.ppt>.

Powerpoint about KeyWords, Birmingham July 2005. Retrieved 2008-26-06 < http://www.lexically.net/downloads/writing/bham/talk.ppt>.

Appendices

First 120 Keyword Results of the Kenyan Part of ICE-EA

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
1	1888	53568.082	kenya	37	223	2979.383	assistant
2	999	19207.262	africa	38	915	2954.042	government
3	602	17018.699	nairobi	39	122	2942.054	maize
4	791	13146.960	speaker	40	101	2911.160	mombasa
5	586	12228.080	accused	41	118	2896.825	pest
6	1392	10487.340	children	42	163	2813.001	discourse
7	358	9310.626	shillings	43	782	2732.191	school
8	288	8090.491	moi	44	339	2628.544	nation
9	1142	7464.289	country	45	506	2580.993	society
10	340	6902.483	hon	46	196	2521.157	deputy
11	2065	6878.270	people	47	84	2421.163	wananchi
12	791	6751.531	areas	48	208	2246.432	drugs
13	575	6403.195	programme	49	77	2219.399	matatu
14	219	6281.633	kanu	50	569	2177.168	money
15	1085	6115.879	women	51	207	2147.092	organisation
16	812	5474.526	language	52	75	2017.279	swahili
17	794	5471.126	development	53	73	2015.290	fuelwood
18	182	5245.852	mfi	54	68	1959.989	turkana
19	179	5159.382	kiswahili	55	431	1949.101	police
20	812	5159.348	problem	56	439	1944.754	office
21	422	4667.682	environment	57	67	1931.165	kisumu
22	677	4555.966	members	58	403	1915.131	story
23	655	4221.474	minister	59	102	1878.988	livestock
24	411	4103.780	teacher	60	244	1868.551	disease
25	399	4054.300	officer	61	283	1820.051	traditional
26	226	3996.877	crops	62	365	1817.315	food
27	149	3703.908	learner	63	93	1803.922	malaria
28	520	3579.153	services	64	104	1801.150	receipt
29	523	3539.509	students	65	75	1798.182	lingua
30	129	3458.916	tanzania	66	70	1796.368	kikuyu
31	385	3452.690	district	67	62	1787.049	imanyara
32	324	3396.439	farmers	68	63	1785.531	arap
33	539	3320.474	president	69	60	1729.402	kiliku
34	462	3267.409	issue	70	59	1700.578	kitui
35	572	3263.418	education	71	58	1671.755	kipsigis
36	111	3138.839	kenyatta	72	58	1641.455	odinga

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
73	105	1638.066	stamp	113	41	1151.669	tribalism
74	56	1614.108	kbc	114	339	1140.985	information
75	56	1614.108	mungai	115	53	1131.041	pidgin
76	207	1588.963	ministry	116	40	1122.864	kariuki
77	56	1554.590	parastatal	117	134	1114.976	identify
78	66	1546.417	uganda	118	43	1101.360	kali
79	350	1533.128	health	119	61	1098.548	faithfully
80	251	1530.718	security	120	38	1095.288	nyanza
81	136	1513.261	oral				
82	132	1500.278	colonial				
83	52	1498.815	harambee				
84	78	1461.064	deceased				
85	51	1439.762	nakuru				
86	406	1427.336	person				
87	237	1425.459	statement				
88	67	1420.864	pesticides				
89	106	1415.833	province				
90	402	1370.806	question				
91	217	1370.788	baby				
92	47	1354.698	mfl				
93	273	1348.990	population				
94	66	1327.674	arid				
95	46	1325.875	eldoret				
96	46	1325.875	orkoiik				
97	454	1310.574	national				
98	166	1303.566	agricultural				
99	50	1276.065	aphids				
100	565	1267.848	house				
101	45	1266.899	luo				
102	68	1261.719	narratives				
103	205	1240.639	leaders				
104	43	1239.405	shikuku				
105	56	1230.652	somalia				
106	42	1210.581	machakos				
107	105	1208.583	arrested				
108	371	1203.671	land				
109	225	1190.009	activities				
110	41	1181.758	kakamega				
111	277	1159.562	international				
112	189	1152.366	parliament				

First 120 Keyword Results of the Kenyan Written Part

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
1	1079	54618.378	kenya	37	60	2139.984	arid
2	582	16524.503	africa	38	282	2121.954	students
3	257	12790.668	nairobi	39	71	2070.926	livestock
4	142	6353.566	learner	40	51	2058.285	uganda
5	559	5325.117	language	41	61	2015.221	faithfully
6	118	5250.808	pest	42	52	2011.663	pidgin
7	102	5248.999	kanu	43	315	2004.984	education
8	680	5234.535	children	44	351	1932.711	problem
9	301	4982.342	environment	45	36	1818.931	sandflies
10	155	4817.162	discourse	46	182	1797.610	teacher
11	165	4687.249	crops	47	220	1796.777	population
12	117	4596.224	shillings	48	35	1766.987	odinga
13	506	4582.742	development	49	235	1751.257	issue
14	92	4431.446	moi	50	135	1738.784	agricultural
15	408	3837.143	areas	51	33	1715.502	kws
16	599	3772.700	women	52	33	1715.502	turkana
17	72	3742.913	kiswahili	53	188	1695.835	nation
18	73	3639.338	fuelwood	54	238	1573.700	food
19	546	3477.806	country	55	37	1568.368	kikuyu
20	92	3251.794	malaria	56	30	1559.547	hydatid
21	62	3223.064	imanyara	57	30	1559.547	kisumu
22	74	3211.376	lingua	58	31	1559.235	nakuru
23	266	3141.412	programme	59	29	1507.562	eom
24	212	3098.743	farmers	60	29	1507.562	ouko
25	74	2931.486	maize	61	28	1455.577	njonjo
26	323	2850.636	services	62	27	1403.592	imbuga
27	53	2755.200	kipsigis	63	27	1403.592	jica
28	55	2754.990	kenyatta	64	27	1403.592	matatu
29	67	2589.006	pesticides	65	27	1403.592	taarab
30	50	2495.429	parastatal	66	171	1389.271	management
31	48	2495.275	mombasa	67	31	1379.110	circumcision
32	55	2441.654	tanzania	68	38	1367.687	somalia
33	51	2405.414	swahili	69	70	1355.742	civilization
34	46	2391.305	orkoiik	70	27	1351.536	moreno
35	50	2310.053	aphids	71	27	1351.536	tribalism
36	188	2256.230	disease	72	26	1299.623	aphid

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
73	25	1299.623	gambiae	113	23	973.946	silage
74	716	1286.034	people	114	22	960.962	creole
75	24	1247.638	eldoret	115	80	950.701	conservation
76	136	1206.905	institutions	116	32	941.286	amnesty
77	23	1195.653	kakamega	117	18	935.728	kiambu
78	23	1195.653	kbc	118	18	935.728	naivasha
79	23	1195.653	leishmaniasis	119	18	935.728	nyanza
80	23	1195.653	marakwet	120	18	935.728	ruganda
81	79	1181.268	colonial				
82	42	1169.311	sustainable				
83	243	1167.719	society				
84	102	1163.737	organisation				
85	22	1143.668	harambee				
86	22	1143.668	issp				
87	22	1143.668	kisii				
88	22	1143.668	muite				
89	252	1141.795	land				
90	368	1118.742	school				
91	32	1112.274	communicative				
92	33	1109.902	sudan				
93	21	1091.683	shamba				
94	31	1090.961	ethiopia				
95	106	1078.575	soil				
96	28	1057.844	uterus				
97	409	1054.305	government				
98	21	1040.153	arap				
99	20	1039.698	matiba				
100	20	1039.698	shikuku				
101	20	1039.698	thome				
102	27	1039.238	giraffe				
103	291	1034.303	social				
104	24	1024.268	bantu				
105	35	1014.336	dialects				
106	23	1011.716	insecticides				
107	32	999.235	firewood				
108	105	991.114	communication				
109	20	988.285	luo				
110	20	988.285	piglets				
111	20	988.285	silo				
112	19	987.713	headteachers				

First 120 Keyword Results of the Kenyan Spoken Part

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
1	809	50205.245	kenya	37	334	3100.485	police
2	553	25668.487	accused	38	74	3092.598	deceased
3	704	25453.308	speaker	39	47	3020.948	mfl
4	345	21468.542	nairobi	40	45	2892.397	kitui
5	332	15148.579	hon	41	42	2634.836	arap
6	241	13414.630	shillings	42	381	2603.096	money
7	417	12431.542	africa	43	286	2544.808	story
8	196	12150.067	moi	44	61	2474.095	narratives
9	182	11698.138	mfi	45	37	2378.193	kisumu
10	1349	7882.661	people	46	506	2353.716	government
11	547	7541.161	minister	47	36	2313.917	quorum
12	117	7454.518	kanu	48	35	2249.642	turkana
13	712	7351.551	children	49	34	2185.366	kggcu
14	107	6877.477	kiswahili	50	306	2166.213	question
15	320	6624.699	officer	51	33	2121.091	kbc
16	484	6062.269	members	52	227	2099.573	issue
17	198	5748.330	assistant	53	48	2085.518	maize
18	596	5368.511	country	54	120	2068.482	drugs
19	183	5269.006	deputy	55	32	2056.815	epz
20	309	5217.582	programme	56	176	2046.378	statement
21	73	4692.110	wananchi	57	36	2022.961	excellency
22	278	4683.714	district	58	36	2022.961	kali
23	461	4381.689	problem	59	293	2010.944	person
24	383	4327.135	areas	60	241	1987.316	students
25	67	4306.457	caliban	61	414	1973.816	school
26	377	4243.557	president	62	30	1928.264	harambee
27	74	4224.600	tanzania	63	93	1885.276	oral
28	99	3836.943	receipt	64	45	1883.766	angola
29	55	3535.152	kiliku	65	288	1876.791	development
30	229	3507.952	teacher	66	263	1831.039	office
31	56	3471.448	kenyatta	67	263	1810.116	society
32	99	3451.793	stamp	68	28	1799.713	isiolo
33	53	3406.601	mombasa	69	28	1799.713	machakos
34	53	3406.601	mungai	70	97	1709.708	blind
35	50	3213.774	matatu	71	257	1701.590	education
36	486	3163.790	women	72	33	1694.366	kikuyu

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
73	28	1676.001	rwanda	113	141	1270.145	security
74	27	1671.530	quantifier	114	24	1268.382	swahili
75	26	1671.163	kangwana	115	32	1241.237	brewing
76	26	1671.163	sunkuli	116	127	1236.217	baby
77	42	1661.070	czechoslovakia	117	80	1233.281	documents
78	382	1634.477	house	118	66	1224.959	liquor
79	129	1632.813	ministry	119	24	1224.508	bosnian
80	231	1614.174	court	120	20	1222.391	nakuru
81	63	1605.386	counsel				
82	105	1554.295	organisation				
83	73	1553.354	arrested				
84	25	1543.161	luo				
85	24	1542.612	nyayo				
86	66	1537.793	province				
87	32	1478.462	cults				
88	23	1478.336	oyondi				
89	23	1478.336	shikuku				
90	156	1458.597	traditional				
91	69	1448.937	bags				
92	103	1440.635	guys				
93	30	1431.237	bwana				
94	86	1421.962	examined				
95	23	1414.823	odinga				
96	22	1414.061	eldoret				
97	22	1414.061	lwali				
98	22	1414.061	nssf				
99	151	1407.360	nation				
100	96	1393.760	poem				
101	94	1364.553	motion				
102	23	1356.392	initialled				
103	22	1350.667	kariuki				
104	21	1349.785	busia				
105	21	1349.785	maasai				
106	253	1347.342	language				
107	197	1345.128	services				
108	36	1291.561	complainant				
109	20	1285.510	muranga				
110	20	1285.510	mwangi				
111	20	1285.510	nyanza				
112	46	1283.370	cheques				

First 120 Keyword Results of the Tanzanian Part of the ICE-EA

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
1	1814	58740.958	tanzania	37	693	3028.835	children
2	1058	23502.013	africa	38	93	3027.391	udsm
3	473	15328.560	dar	39	105	3018.368	leprosy
4	1514	14525.084	country	40	118	2888.351	shillings
5	303	9660.148	zanzibar	41	101	2863.017	uganda
6	296	9635.566	kiswahili	42	88	2797.060	rwanda
7	288	9375.145	ndugu	43	90	2766.071	swahili
8	306	8820.677	appellant	44	145	2763.504	cooperative
9	954	8081.493	problem	45	322	2751.549	activities
10	604	7933.117	programme	46	278	2707.605	technology
11	895	7885.488	development	47	83	2701.865	mwinyi
12	774	7455.823	areas	48	110	2605.207	hiv
13	513	6968.037	region	49	80	2604.207	mrema
14	1154	5754.565	government	50	147	2601.447	workshop
15	332	5272.879	organisation	51	320	2587.741	project
16	417	5226.934	environment	52	288	2496.432	teacher
17	153	4980.546	ccm	53	192	2476.378	democracy
18	701	4768.595	language	54	426	2414.859	president
19	436	4648.463	village	55	78	2376.756	tanganyika
20	1608	4619.012	people	56	332	2314.727	population
21	625	4496.347	education	57	70	2278.681	mwanza
22	124	4036.521	arusha	58	251	2264.800	disease
23	125	4001.058	parastatal	59	382	2214.556	students
24	583	3918.874	economic	60	250	2194.409	rural
25	494	3748.880	services	61	316	2178.300	science
26	223	3581.035	accused	62	328	2162.157	material
27	370	3544.278	nation	63	188	2152.757	aids
28	132	3498.927	kenya	64	367	2140.892	food
29	444	3497.375	issue	65	474	2112.641	political
30	752	3353.690	women	66	89	2084.169	witchcraft
31	470	3239.524	health	67	64	2083.366	amref
32	629	3161.397	national	68	82	2070.101	pesticides
33	173	3091.301	crops	69	300	2068.824	rights
34	417	3077.689	international	70	438	2040.173	land
35	297	3045.175	institutions	71	420	2005.836	minister
36	510	3031.394	members	72	382	1991.455	united

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
73	138	1987.707	offence	113	380	1451.258	local
74	61	1985.708	tanga	114	43	1399.761	morogoro
75	230	1984.967	sector	115	227	1391.587	management
76	263	1968.879	district	116	306	1383.618	research
77	267	1960.422	resources	117	151	1380.893	independence
78	62	1951.317	zambia	118	226	1372.088	thousand
79	336	1924.340	production	119	367	1367.787	law
80	59	1920.603	dodoma	120	149	1365.350	constitution
81	419	1906.151	university				
82	243	1905.509	teaching				
83	236	1890.494	primary				
84	358	1885.925	process				
85	179	1871.369	code				
86	123	1863.783	participants				
87	209	1861.217	ministry				
88	165	1857.628	marketing				
89	100	1857.547	respondent				
90	600	1779.698	school				
91	54	1757.840	shinyanga				
92	83	1754.036	malaria				
93	119	1753.035	rape				
94	104	1729.638	seminar				
95	355	1688.372	court				
96	191	1673.295	developing				
97	58	1672.252	rodent				
98	296	1670.883	foreign				
99	61	1663.749	mozambique				
100	62	1647.005	nigeria				
101	105	1614.280	registration				
102	49	1595.077	moshi				
103	129	1588.213	gender				
104	465	1560.870	social				
105	314	1535.575	situation				
106	424	1502.572	party				
107	220	1494.976	conference				
108	274	1478.355	trade				
109	362	1478.330	human				
110	170	1477.124	secondary				
111	114	1469.154	republic				
112	381	1467.002	person				

First 120 Keyword Results of the Tanzanian Written Part

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
1	1123	53122.725	tanzania	37	113	2419.626	rape
2	627	16857.743	africa	38	50	2384.972	tanga
3	288	13737.436	kiswahili	39	72	2333.532	witchcraft
4	306	12955.594	appellant	40	201	2305.108	institutions
5	263	12446.302	dar	41	324	2255.325	court
6	233	11113.968	ndugu	42	95	2196.802	cooperative
7	830	7296.396	country	43	142	2175.860	marketing
8	653	6504.015	language	44	147	2134.424	aids
9	130	6200.926	ccm	45	215	2089.278	district
10	347	5464.761	region	46	408	2066.385	national
11	575	5318.429	development	47	206	2065.047	teacher
12	115	5202.031	zanzibar	48	45	2051.307	parastatal
13	352	4761.558	programme	49	102	2045.636	crops
14	505	4663.973	education	50	69	2013.402	malaria
15	203	4628.242	accused	51	196	1971.974	teaching
16	93	4436.047	udsm	52	41	1955.677	mwanza
17	462	4418.012	areas	53	324	1921.375	members
18	78	3720.556	mrema	54	129	1904.926	democracy
19	79	3534.556	swahili	55	187	1892.318	primary
20	97	3270.091	hiv	56	256	1877.597	issue
21	466	3140.834	problem	57	138	1841.167	organisation
22	65	3100.463	arusha	58	179	1820.307	rural
23	64	3052.764	amref	59	38	1812.578	shinyanga
24	138	2978.181	offence	60	38	1812.578	tabora
25	88	2928.029	shillings	61	477	1811.310	school
26	342	2884.219	services	62	49	1796.166	uganda
27	58	2766.567	mwinyi	63	201	1748.506	activities
28	100	2758.877	respondent	64	67	1741.273	switching
29	385	2718.515	economic	65	323	1720.520	law
30	78	2696.715	kenya	66	38	1718.152	chloroquine
31	642	2683.160	government	67	36	1717.180	dodoma
32	257	2657.721	project	68	36	1717.180	moshi
33	59	2585.191	tanganyika	69	35	1669.480	cassava
34	168	2548.703	code	70	174	1667.541	environment
35	58	2456.153	rodent	71	418	1666.702	children
36	242	2441.537	village	72	229	1642.867	trade

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
73	34	1621.781	mtwara	113	25	1192.486	bridewealth
74	58	1607.926	surgical	114	25	1192.486	utlip
75	149	1532.969	ministry	115	36	1191.043	malice
76	161	1527.071	disease	116	29	1171.345	bantu
77	32	1526.382	trachoma	117	204	1163.653	international
78	82	1525.197	slave	118	83	1163.540	ethnic
79	253	1478.325	health	119	80	1162.074	mathematics
80	294	1468.501	university	120	24	1144.786	actus
81	248	1465.790	students				
82	30	1430.983	singida				
83	156	1427.151	technology				
84	141	1419.836	learning				
85	131	1406.521	secondary				
86	295	1397.529	person				
87	52	1389.528	deceased				
88	29	1383.284	kigoma				
89	772	1353.723	people				
90	28	1335.584	kagera				
91	28	1335.584	morogoro				
92	28	1335.584	musoma				
93	28	1335.584	mzee				
94	28	1335.584	sadcc				
95	29	1335.241	aforethought				
96	29	1335.241	passbook				
97	240	1326.853	process				
98	139	1311.513	trial				
99	83	1299.203	republic				
100	27	1287.885	mtera				
101	299	1287.391	political				
102	33	1287.093	immunization				
103	138	1270.742	goods				
104	278	1264.669	land				
105	130	1260.193	developing				
106	26	1240.185	sido				
107	36	1216.291	condoms				
108	34	1208.780	mozambique				
109	29	1208.593	insecticides				
110	218	1208.263	evidence				
111	110	1203.184	constitution				
112	57	1199.535	intercourse				

First 120 Keywords of the Tanzanian Spoken Part

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
1	691	69907.581	tanzania	37	213	3116.518	international
2	210	21320.560	dar	38	29	2972.861	mwanza
3	431	21127.047	africa	39	30	2879.414	unita
4	188	18664.661	zanzibar	40	217	2678.600	health
5	684	11878.455	country	41	86	2580.117	honourable
6	103	9352.398	leprosy	42	25	2562.811	mwinyi
7	488	8270.056	problem	43	25	2562.811	tembi
8	80	7997.069	parastatal	44	71	2506.277	crops
9	76	7587.281	rwanda	45	61	2471.120	seminar
10	194	7418.285	organisation	46	188	2461.734	issue
11	243	7053.184	environment	47	23	2357.786	ccm
12	59	6048.234	arusha	48	23	2357.786	dodoma
13	252	5951.634	programme	49	23	2357.786	sukoma
14	73	5716.648	pesticides	50	28	2285.092	zaire
15	55	5638.184	ndugu	51	78	2281.473	gender
16	836	5036.835	people	52	32	2262.579	chairperson
17	312	4902.604	areas	53	21	2152.761	kampala
18	214	4605.156	nation	54	25	2127.351	agro
19	443	4563.088	women	55	122	2099.518	technology
20	512	4339.683	government	56	133	2097.335	conference
21	280	4204.568	hundred	57	163	2065.919	material
22	41	4203.010	bawata	58	20	2050.249	mkapa
23	52	4177.872	uganda	59	76	2013.111	cotton
24	202	4076.616	thousand	60	27	1951.020	mozambique
25	320	4033.734	development	61	19	1947.736	tpri
26	40	3901.427	zambia	62	53	1947.528	refugees
27	194	3753.977	village	63	177	1921.228	food
28	47	3753.827	nigeria	64	36	1863.134	sustainable
29	84	3614.359	workshop	65	20	1860.228	lusaka
30	54	3608.658	kenya	66	59	1860.045	participants
31	48	3608.608	sudan	67	52	1806.078	kilometres
32	35	3390.210	seychelles	68	275	1796.602	children
33	166	3309.121	region	69	202	1790.467	minister
34	250	3305.554	united	70	75	1766.637	conservation
35	250	3213.551	president	71	19	1758.621	burundi
36	198	3162.251	population	72	17	1742.712	eafod

Rank	Freq.	Keyness	Word	Rank	Freq.	Keyness	Word
73	23	1737.474	mandela	113	13	1332.662	zanzibaris
74	198	1728.040	economic	114	96	1319.836	institutions
75	50	1727.731	cooperative	115	20	1308.586	nairobi
76	20	1701.881	dakar	116	16	1305.767	convocation
77	141	1682.300	science	117	16	1305.767	malawi
78	137	1673.622	rights	118	23	1302.543	zimbabwe
79	53	1652.696	breast	119	41	1280.488	egypt
80	17	1644.006	nyerere	120	44	1279.849	registration
81	16	1640.199	shinyanga				
82	122	1594.303	resources				
83	102	1545.591	sector				
84	121	1545.520	activities				
85	15	1537.687	excellencies				
86	15	1537.687	ivd				
87	15	1537.687	lagos				
88	15	1537.687	morogoro				
89	15	1537.687	tapo				
90	19	1534.051	tanganyika				
91	186	1529.378	members				
92	43	1498.462	mainland				
93	22	1488.885	johannesburg				
94	33	1451.750	paints				
95	221	1450.947	national				
96	37	1448.439	rebels				
97	15	1439.707	pastoralism				
98	31	1437.240	symposium				
99	14	1435.174	fgm				
100	14	1435.174	goma				
101	14	1435.174	subsector				
102	17	1404.298	mutilation				
103	25	1401.648	palestinian				
104	32	1382.459	diabetes				
105	16	1376.172	kilimanjaro				
106	152	1373.754	services				
107	30	1365.361	shillings				
108	143	1336.041	production				
109	13	1332.662	harare				
110	13	1332.662	mdt				
111	13	1332.662	monrovia				
112	13	1332.662	moshi				

Stoplist

Common English Words

aware	give	supposed
beg	helps	sworn
call	hope	talk
called	important	talking
case	issue	telling
cent	kind	things
conjunctions	year	time
continue	lot	today
day	main	told
dear	order	understand
due	point	year
fact	signed	yesterday
find	sincerely	sir

Fictitious Names

Ato	Zsuzsa	Kibonge
Pamela	Mutiso	Gado
Pamella	Erika	Kellana
Waweru	Muasya	Wembe
Karane	Kanini	Isaya
Chipota	Arita	Hanahela
Murugi	Waceera	Gakenia
Mathenge	Abacha	Nduduzi
Kanaya	Nduta	Yurani
Njeri	Kuya	Prospero
Lucy	Caliban	

Year Numbers	Expressions	Abbreviations	Text Markings	Others
eighty	aha	Dr	BHN	es
fifty	coz	Mr	BINT	Salaam
nineteen	etcetera	Nos	BK	franca
ninety	uh	PW	BT	francas
seventy	uhu	Ref	SK	
sixty	yah	XX		
twenty	yeah	XXd		

Lemma list

Africa → African,Africans
appellant → appellants
areas → area
children → child
cooperative → cooperatives
country → countries
crops → crop
disease → diseases
environment → environmental
farmers → farmer
institutions → institution
issue → issues
Kenya → Kenyan,Kenyans
language → languages
learner → learners
matatu → matatus
material → materials
members → member
minister → ministers
nation → nations
officer → officers
organisation → organisations
parastatal → parastatals
pest → pests
problem → problems
programme → programmes
region → regional
rodent → rodents
rwanda → rwandan, ruanda
school → schools
services → service
shillings→ shs,Kshs
stamp → stamps
story → stories
students → student
Tanzania → Tanzanian,Tanzanians
teacher → teachers
village → villages

Zusammenfassung der Magisterarbeit mit dem Titel:

„Typicality of Lexical Items of Kenyan and Tanzanian English“

Die soziolinguistische Situation stellt sich besonders schwierig dar in Ländern wie Kenia und Tansania. Englisch ist kein gängiges Medium der Verständigung, sondern nur auf besondere Bereiche, wie Politik oder Bildung, begrenzt. Um diese spezielle linguistische Situation näher zu analysieren, wurde ein Korpus mit verschiedenen Textkategorien erstellt, der die linguistischen Gegebenheiten in Kenia und Tansania widerspiegeln soll, zumindest zum größten Teil. Dieser Korpus dient als Grundlage für die Studie dieser Arbeit: die Analyse typischer lexikalischer Begriffe, sogenannter Schlüsselwörter, der Englischen Sprache, wie sie in Kenia und Tansania genutzt wird.

Die Arbeit beginnt mit einem theoretischen Teil, der ein spezielles Hintergrundwissen vermitteln soll, welches für die Studie notwendig ist. Zuerst wird auf die Aufgabe von Korpus Linguistik eingegangen und deren Möglichkeiten, aber auch Schwächen dargestellt. Danach werden spezielle Korpusse vorgestellt, die von wichtiger Bedeutung für diese Studie sind. Dies sind zum einen der schon erwähnte Korpus, als auch weitere Korpusse aus verschiedenen englischsprachigen Regionen, mit denen der untersuchte Korpus verglichen wird. Des Weiteren wird auf die Bedeutung von Typikalität und Schlüsselwörtern eingegangen.

Die Analyse des Korpus beginnt mit einem Kapitel, in der die Arbeitsweise mit einem Analyseprogramm näher erläutert wird, d.h. welche Einstellungen notwendig sind, um angemessene Ergebnisse zu erzielen. Die erhaltenen Ergebnisse werden dann im nächsten Kapitel genauer analysiert, ob sie zu recht als Schlüsselwörter festgestellt wurden und wie hoch der Grad der Typikalität ist. Zusätzlich wird erläutert, welche Schlüsselwörter die Politik und Kultur Kenias sowie Tansanias, als auch die Lebensweise deren Volkes, besonders widerspiegeln. Zum Schluss wird noch ein Vergleich zwischen beiden Ländern gezogen, um herauszufinden, welches am besten durch den Korpus repräsentiert wird.