Southern African Linguistics and Applied Language Studies 2000 18: 89–106
Printed in South Africa — All rights reserved

Copyright © 2000 NISC Pty Ltd
SOUTHERN AFRICAN LINGUISTICS
AND APPLIED LANGUAGE STUDIES
ISSN 1607–3614

The compilation of electronic corpora, with special reference to the African Languages[1]

Gilles-Maurice de Schryver and DJ Prinsloo
Department of African Languages, University of Pretoria, Pretoria, 0002, South Africa.
e-mail: schryver@postino.up.ac.za

Abstract: Compiling and querying electronic corpora has become a *sine qua non* as an empirical basis for contemporary linguistic research. As a result, around the world, corpus applications now abound in all fields of linguistics. In this article it is argued that, if African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should become an absolute priority. The article first presents a comprehensive theoretical conspectus of electronic corpora. This theoretical section is followed by a practical exploration for the African languages. To that end, two very different African-language corpus projects are described in detail. The survey of these two projects, combined to inter-African-language comparisons, are deemed to be sufficient proof of the feasibility of establishing a discipline of corpus linguistics for the African languages at present.

Introduction

Worldwide, the compilation, querying and application of electronic corpora has undoubtedly revolutionised studies as well as descriptions of the structure and use of languages. This revolutionary aspect of present-day linguistics is well echoed in the literature:

> 'Contemporary approaches to the investigation of actual language use entail the examination and analysis of collections of different kinds of spoken or written texts, or corpora (the plural of the Latin corpus "body"). The term 'corpus linguistics' is now used increasingly in the literature, and indeed is found in the titles of a number of influential publications in the field of contemporary linguistic enquiry' (James, Davison, Cheung Heung-yeung & Deerwester, 1994:4)
>
> 'The crucial point here is that corpora are used, and are now widely accepted as valuable, arguably essential, resources in serious linguistic description of any kind' (Moon, 1998:347)
>
> 'It has become widely accepted that a well-designed corpus is a prerequisite for study of any language' (Jeffery, 2000:71)

If scholars of African languages are to take their rightful place in the new millennium, it is plain that the active compilation, querying and application of electronic corpora should become an absolute priority. Against the background of this modern linguistic trend, the aim of the present article is to illustrate the feasibility of compiling electronic corpora for all African languages. As such, this article first gives a thorough overview of the main issues one needs to take into account when dealing with corpora. Special attention is given to the need to compile a structured corpus and to the three main steps of corpus compilation itself, namely corpus design, text collection and text encoding. This theoretical section is then followed by a detailed study of a series of African-language corpora that were assembled recently. In order to illustrate the interplay between the South-African African languages and African languages from the same language family, and in order to enable a thorough analysis of the practical issues involved in compiling a structured electronic corpus, two very different projects are described. In the first, the creation of a small-size electronic corpus for Cilubà is contrasted with different phases in the creation of a Sepedi corpus, and in the second the main features in building a large Internet corpus for Kiswahili show the way for similar possibilities for languages such as isiZulu, isiXhosa and Setswana. At the end of this article, it is hoped that scholars will be in a position to instantly start compiling the corpora which suit their specific need(s).

Whilst the present article deals with both theoretical and practical aspects of corpus compilation, a subsequent article will focus on various corpus applications in the broad field of linguistics. Specific lexicographic applications will be treated in two ensuing articles, one looking at the macrostructural level and one at the microstructural level. Finally, the corpora-articles series will be concluded with a study devoted to corpus integrity and stability issues.

Electronic corpora – A theoretical conspectus

Some basic definitions

Although the term 'corpus' has quite a distinguished pedigree, James *et al.* point out that:

> 'in recent usage [the term corpus] has tended to refer to a comprehensively documented and structured collection of complete texts, or extracts from larger texts, whose components are generally separately accessible' (James *et al.*, 1994:4)

Of utmost importance here is the fact that the different texts making up a corpus are a structured collection which is comprehensively documented. Such a set of texts should not be confused with an 'archive' which is largely unstructured since the latter is collected more or less opportunistically (Leech, 1991:10). Also, the terms corpus and archive should not be confused with 'database':

> 'By contrast, text database, a term often confusingly used for corpus, refers not to the product, but to the process: the method of storage of the material collected. Although not necessarily so, it is probably true that nowadays corpora and text archives can be assumed to be stored in machine-readable form' (James *et al.*, 1994:5)

Of course, it is true that archive material can be incorporated into an electronic text corpus which in turn is processed in a database. Yet it is precisely as a result of the 'machine-readable form' that current corpora are referred to as 'electronic corpora' (often shortened to just 'corpora', however). 'Machine readable' does not necessarily imply that the texts are stored in a computerised 'database', even though they are, in many cases. The fact that they often are has led to a new research paradigm called 'corpus linguistics':

> 'a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description. Over the last three decades the compilation and analysis of corpora stored in computerized databases has led to a new scholarly enterprise known as corpus linguistics. [...] Corpus linguistics is not an end in itself but is one source of evidence for improving descriptions of the structure and use of languages [...] Some of the largest corpus projects have been undertaken for commercial purposes, by dictionary publishers. [...] It is the probabilistic aspect of corpus-based descriptive linguistic studies which especially distinguishes them from conventional descriptive fieldwork in linguistics or lexicography' (Kennedy, 1998:1, 5, 9)

However, long before the advent of electronic, machine-readable corpora in the early 1960s, linguistic projects did already utilise corpora (cf. e.g. Francis, 1992) – albeit without the 'incredible speed, total accountability, accurate replicability, statistical reliability and the ability to handle huge amounts of data' (Kennedy, 1998:5) which characterise *electronic* corpora nowadays.

Major (English) electronic corpora in historical perspective

The earliest major electronic corpus for linguistic research was the pioneering Brown University Standard Corpus of Present-Day American English (known as the Brown Corpus), a synchronic corpus of roughly one million words (Francis & Kucera, 1964). From 1964 onwards until the late 1980s the unofficial standard size, or in terms of Leech (1991:22), the 'going rate' for electronic corpora remained at roughly one million running words. Still, one early electronic corpus built specifically for lexicographic purposes during that period, the American Heritage Intermediate (AHI) Corpus, was a large commercial corpus of 5.09 million words of text (Carroll, Davies & Richman, 1971).

Beginning with the first major lexicographical mega-corpus project, the *Collins*

Birmingham University International Language Database (Cobuild), corpus sizes went inexorably upwards. Indeed, while the *Cobuild Main Corpus* contained 7.3 million words in 1982, by 1987 a 13-million-word *Cobuild Reserve Corpus* had been assembled in addition (Renouf, 1987:7, 10). Meanwhile, the *Longman Corpus Network*, a commercial project consisting of three major corpora, was under construction. One of them, the *Longman Lancaster English Language Corpus*, eventually contained 30 million words (Summers, 1993:184, 201). Between 1991 and 1995 the *British National Corpus (BNC)*, with about 100 million words, was undoubtedly the most ambitious corpus project up to then (Kennedy, 1998:50). Finally, the *Bank of English*, initiated in 1991 at the University of Birmingham, stood at over 320 million words in 1998 (Hartmann & James, 1998:12).

In Figure 1 the sizes of the major English electronic corpora are plotted in historical perspective.

Notwithstanding this skyrocketing growth in corpus size, Kennedy points out that 'the very notion of what constitutes a valid corpus can still be controversial' (1998:2).

Balanced versus representative, and organic versus structured corpora

Anyone leafing through the literature on corpus linguistics will quickly find out that the endeavour to compile 'valid corpora' (Kennedy's term) revolves around two concepts, viz. 'balanced corpora' versus 'representative corpora'.

> 'A general corpus is typically designed to be balanced, by containing texts from different genres and domains of use including spoken and written, private and public [...] For a corpus to be "representative" there must be a clearly analysed and defined population to take the sample from. But because we cannot be confident we know all the possible text types nor their proportions of use in the population, a "representative" sample is at best a rough approximation to representativeness, given the vast universe of discourse. [...] the issue is really 'representative of what?' In light of the perspectives on variation offered by several decades of research in discourse analysis and sociolinguistics, it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre' (Kennedy, 1998:20, 52, 62)

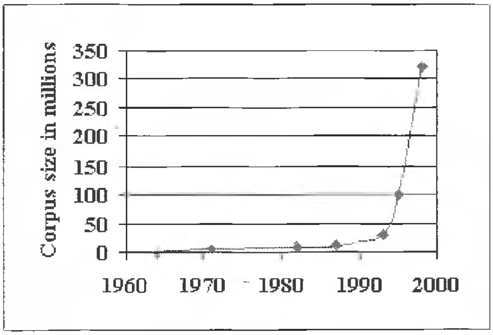

Figure 1: Corpus sizes of the major English electronic corpora in historical perspective

A completely different viewpoint is that of, for instance, Leech (1991). He calls a corpus 'representative' in that findings based on an analysis of it can be generalised, if not to the language as a whole, at least to a specified part of it. Summers also believes in a 'representative' corpus, as she claims:

> 'The idea of representativeness has been central to our thinking about the structure of the corpus. We believe that unless the corpus is representative, it is ipso facto unreliable as a means of acquiring lexical knowledge. Our answer to the question: "Representative of what?" would be "Representative of the standard language in a very general sense, not restricted to a regional variety [...] or a narrow range of text types" [...] What we mean by representative is covering what we judge to be the typical and central aspects of the language, and providing enough occurrences of words and phrases for the lexicographers [...] to believe that they have sufficient evidence from the corpus to make accurate statements about lexical behaviour' (Summers, 1993:186, 190)

> 'to be representative of general language. This is a bold ambition – some say one that is impossible to fulfil' (Summers, s.d. [1996-1998]:6)

Still other scholars simply fuse the two terms, as in:

> 'COBUILD have always insisted that it is impossible to create a corpus that is truly representative of the language, and have focused on size of corpus rather than balance' (Kilgarriff, 1997: 150)

> 'Lexicographers traditionally aim at a "representative" or "balanced" corpus, that is, the corpus should be appropriate as the basis for generalizations concerning the language as a whole' (Kruyt & Dutilh, 1997:230)

From the above it is clear that linguists disagree whether a corpus should try to be balanced or representative. It seems as if a corpus will never be balanced because there are too many parameters, and it seems as if a corpus will never be truly representative of all language usage, either, as it is impossible to define the population. Yet the corpus compiler can strive to come as close to the ideal situation as possible. This agrees with Kennedy's observation in that 'The notions of representativeness and balance are, of course, in the final analysis, matters of judgement and can only be approximate' (1998:62).

Nonetheless, probably one of the most interesting approaches is the one by Atkins, Clearv and Ostler. They introduce the concept of **'organic corpora'**:

> 'a corpus may be thought of as organic, and must be allowed to grow and live if it is to reflect a growing, living language. [...] In order to approach a "balanced" corpus, it is practical to adopt a method of successive approximations. First, the corpus builder attempts to create a representative corpus. Then this corpus is used and analysed and its strengths and weaknesses identified and reported. In the light of this experience and feedback the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually. [...] In our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus – however "unbalanced" – to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts' (Atkins, Clear & Ostler, 1992:1, 4, 6)

Formulated differently, it is any corpus compiler's task to attempt to assemble a representative corpus for his/her specific need(s). Subsequent additions and deletions of sections should be seen as a balancing activity to rectify initial weaknesses, but more importantly, also to take account of and track a growing, living language. As such, there is no such thing as 'the' corpus of a certain language (variety). Rather, at any point in time one selects a certain number of texts from the range of available electronic texts (which might or might not be grouped together into sub-corpora), and uses 'a' corpus for the specific research one wishes to pursue. The minimum requirement for any organic corpus is thus that the corpus compiler(s) will have attempted to put some structure in assembling the range of electronic texts. Within this framework, any first attempt at compiling an organic corpus will at least result in a structured corpus. The series of corpora for the African languages that will be discussed below, should therefore be seen as 'structured corpora', and hence as stepping stones to true organic corpora.

Corpus compilation

Once one recognises the need for compiling a structured corpus which, in due time, should take the form of an organic corpus, one must turn to the issue of 'corpus compilation' itself. There are three steps to be considered in such a compilation: a) corpus design, b) text collection, and c) text encoding. Each of these steps will now be discussed briefly (where the focus will be on issues relevant to the African corpora that will be discussed below).

Corpus design

In the literature, several 'corpora type dichotomies' can be distinguished. One can, for instance, differentiate between 'general or core corpora' versus 'specialised corpora' (such as training and test corpora (Leech, 1992:112), dialect corpora, learner's corpora, etc.), or 'written corpora' versus 'spoken corpora', 'full-text corpora' versus 'sample-text corpora', 'synchronic corpora' versus 'diachronic corpora', or even 'static corpora' versus 'dynamic or monitor corpora' (Kennedy, 1998:19-23). A 'dynamic or monitor corpus' is an open-ended language bank in which new text opportunistically replaces material which was in the corpus earlier. Though such a corpus is constantly growing and changing, Sinclair points out that the huge number (hundreds of millions) of words would gradually 'get too large for any practicable handling and will be effectively discarded' (1991:25). A dynamic or monitor corpus should thus not be confused with an organic corpus, for, in the first, vast numbers of running words simply replace careful planning of sampling as the main design criterion.

With this spectrum of possibilities for 'corpus design' – where different poles of various dichotomies can of course be combined – one should keep in mind that 'The optimal design of a corpus is highly dependent on the purpose for which it is intended to be used' (Kennedy 1998:70). No general statements can therefore be made. Yet, in order to illustrate possible design methods, we can begin by looking at how the first major lexicographical mega-corpus was conceived. Some of the principles that were used in the creation of the 7.3-million-word Cobuild Main Corpus have been enumerated by Renouf (1987:2–5): 25% spoken text, broadly general rather than technical language, from 1960 onwards, preferably 'naturally occurring' text, writing and speech produced by adults aged 16 or over, etc. Spoken texts came from transcripts of radio broadcasts, university archives of oral interviews and lectures, etc. Written texts were chosen from widely read works (excluding poetry) and authorship was 25% female. Newspaper and journalistic texts were also thrown in.

As a second illustration, we can consider the design of the Longman Lancaster English Language Corpus, where texts were selected in two ways. A 'selective half' was chosen through a mixture of pragmatic measures to

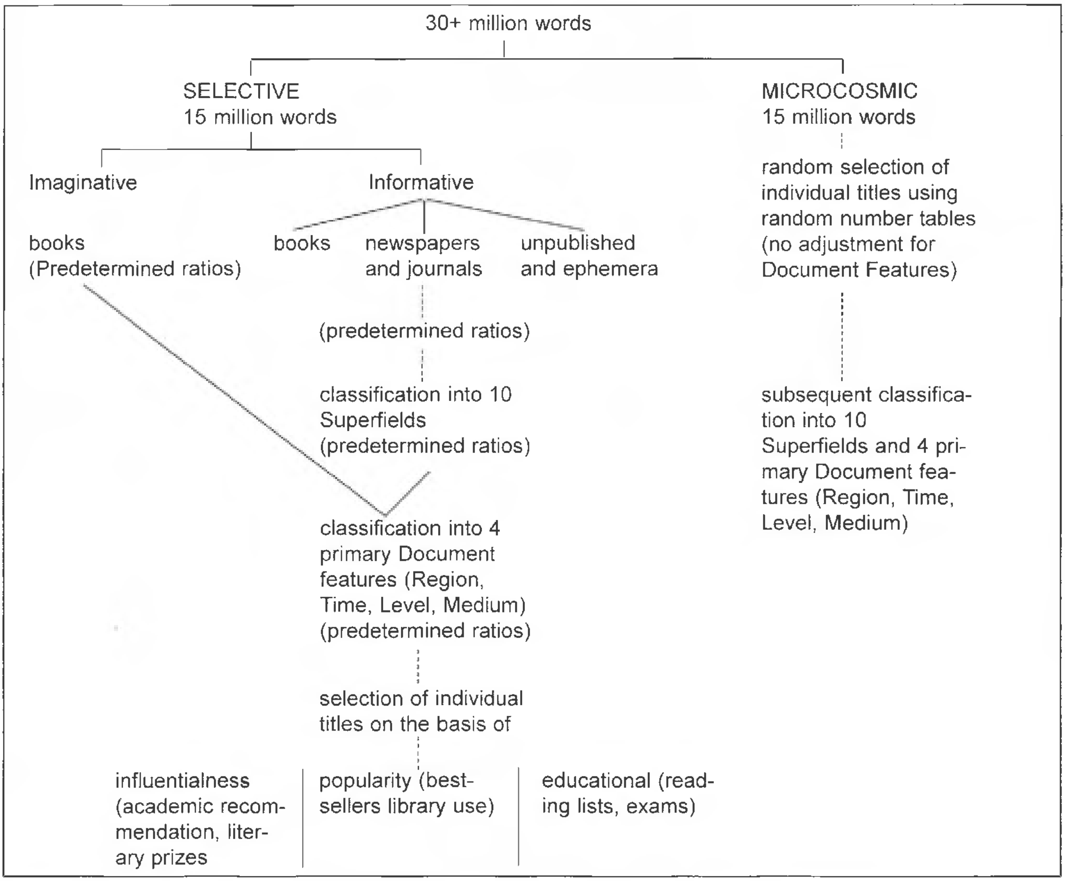
30+ million words
SELECTIVE
15 million words
MICROCOSMIC
15 million words
Imaginative
Informative
books
(Predetermined ratios)
books
newspapers and journals
unpublished and ephemera
(predetermined ratios)
classification into 10 Superfields (predetermined ratios)
classification into 4 primary Document features (Region, Time, Level, Medium) (predetermined ratios)
selection of individual titles on the basis of
influentialness (academic recommendation, literary prizes
popularity (best-sellers library use)
educational (reading lists, exams)
random selection of individual titles using random number tables (no adjustment for Document Features)
subsequent classification into 10 Superfields and 4 primary Document features (Region, Time, Level, Medium)

Figure 2: Design of the *Longman Lancaster English Language Corpus* (Summers, 1993:201)

gather a broad range of objectively defined 'document types', and a 'microcosmic half' was brought together by randomly selecting books. The result is shown in Figure 2.

The idea of using 'document types' was introduced by Michael Rundell, and has been defined as 'text from a particular subject area, together with a cluster of relatively objectively identifiable features such as time, region, medium, and level' (Summers, 1993:192). Ten such broad subject areas were adopted, namely natural and pure science (6.0%), applied science (4.3%), social science (14.1%), world affairs (10.4%), commerce and finance (4.4%), arts (7.9%), belief and thought (4.7%), leisure (5.7%), fiction (40.0%, a discretionary percentage), and poetry, drama, humour (2.3%) (Summers, 1993:192-193). As a result, the corpus is essentially topic-driven rather than genre-driven. In the summary diagram presented in Figure 2, one notes the prominent absence of spoken sources. The corpus team argued that 'although the importance of the spoken language could hardly be overestimated, the logistical problems of capturing a sufficient large body of speech electronically were very real, unlike with written material' (Summers, 1993:184).

There has indeed been a lot of debate among scholars about the value of including spoken data – especially transcriptions of unscripted conversations – in a corpus. Again, the exact purpose for which the corpus is intended to be used should be kept in mind. In the field of lexicography, for instance, Moon suggests that 'the constraints of the conventional dictionary [...] make it difficult if not impossible to deal with the distinguishing features of spoken language properly and fully' (1998: 353). Although we agree in principle with Moon's point of view, it is also true that the study of oral data can pinpoint words which tend to be used more frequently in oral versus written communication. Nonetheless, corpora built by dictionary projects tend to focus on written texts. In other fields, such as speech-processing technology, corpora exclusively consist of oral material, and this 'in spite of the huge practical, technical, financial, and ethical problems associated with the acquisition of spoken data' (Moon, 1998:348).

Text collection

Following the design of a corpus, the next step is 'text collection'. If the design indicates that an oral (sub-)corpus is required, or if one is dealing with a language for which virtually no handwritten or printed, let alone electronic, material is available, there is no way around it: one has to record a large variety of spoken speech from as many different genre/topic areas as possible. These recordings will then have to be transcribed on a word processor using the available (or created) orthography, and instantly saved in computer files. Still, one will do well to keep Kennedy's observation in mind:

> 'A transcription is an imperfect written approximation of a speech event which exists initially as a dance of air molecules. The level of delicacy or amount of detail in a transcription is [...] related to the use to which the transcription will be put' (Kennedy, 1998:82)

In addition, not all 'speech events' are equivalent, with transcriptions of unscripted conversations at one end of the scale, and monologues, lectures, screenplays and semiscripted broadcast journalism at the other end (Moon, 1998:349). While the latter end of the scale might be fairly easy to 'transcribe', the former definitely is not.

When it comes to existing written material, there seem to be three ways of entering them into computer files: (1) 'electronic transfer', e.g. downloading a variety of well-selected documents from the Internet or retrieving texts which already exist on computer disk; (2) '(re)keyboarding', i.e. typing of handwritten documents or even printed matter into computer files; and (3) 'scanning' of printed matter into computer files by means of the so-called OCR (Optical Character Recognition) process using computer software such as OmniPage or Recognita.

Text encoding

Text collection is often (at least for corpora of major European languages) followed by 'text encoding', i.e. the raw text is supplemented by a series of so-called 'standard corpus pre-processing' annotations. Text encoding can consist of any combination of the following: a) word tokenisation, b) part-of-speech tagging, c) lemmatisation, d) syntactic parsing and e) markup.

a) Although we tend to take it for granted, word boundaries are not delimited in all languages by spaces and punctuation marks. Indeed, many African languages, for instance, contain few word delimiters, as they have a conjunctive orthography. Segmenting a text containing conjunctively written words into free-standing words is known as 'word tokenisation' (cf. also Mills, 1998:213, 215).

b) Assigning a word class to all the words in a corpus is called 'part-of-speech tagging' (POS-tagging): 'The original "raw" text can also be annotated or pre-processed linguistically to show the word class of each word in the text by means of a grammatical tag or label which is attached to each word' (Kennedy, 1998: 21). Among others, part-of-speech tagging provides crucial data for lemmatisation, parsing, or advanced concordancing.

c) Part-of-speech tagging, together with detailed morphological information, enables the lemmatisation of a corpus. According to Hartmann & James, 'lemmatisation' should be understood as 'The reduction of a paradigm of variant word forms to a canonical form, e.g. the inflected forms (-s, -ed, -ing etc.) of English verbs to the infinitive' (1998:83). Hence, a 'lemmatiser' (also called 'morphological analyser') merges a certain paradigm of variants into a single canonical form. Lemmatisation is especially useful for lexicographic purposes.

d) If tagging deals with words, parsing deals with entire sentences. A structural analysis of sentences is known as 'syntactic parsing': 'Corpora can also be parsed to show the sentence structure and the function in the sentences of the different word classes' (Kennedy, 1998:21). Syntactic parsing constitutes an important stage in, for example, machine-aided translation or dialogue systems.

Today these four standard pre-processing annotations, (a)–(d), can be automated:

> 'Whilst initially linguistic tagging was effected with minimal computational support, programs are now available which can assign this information automatically, and also offer a full syntactic analysis, or parse, for each sentence, i.e. the complete constituent structure, and for each constituent, its linguistic category and its function in the overall construction' (James *et al.*, 1994:27)

Yet, James *et al.* are referring to tagging and parsing English corpora, as automatic taggers and parsers, as well as automatic tokenisers and lemmatisers, for the African languages are still very much in their infancy. One notable exception is the tools developed by Hurskainen's team at the University of Helsinki (cf. e.g. Hurskainen, 1992).

e)The four standard pre-processing annotations discussed so far, (a) – (d), are used to encode 'detailed factual or interpretive data about the original text for the purpose of information retrieval' (James *et al.*, 1994:26). Existing electronic texts, however, are often already annotated. Yet such annotation, better known as 'markup', operates on a different level, as here 'text features' are encoded. The popular way of encoding various electronic text features such as typefaces, line breaks, word breaks, sections, paragraphs, headings, and all other aspects of page layout, is the Standard Generalised Markup Language, for short SGML (Bryan, 1988; Goldfarb, 1990). SGML is powerful, flexible, independent of particular software systems and 'enables electronic texts originally keyboarded on different word processors to be edited, searched, analysed or typeset consistently' (Kennedy, 1998:83)[2]. Intimately linked with markup languages is the Text Encoding Initiative, for short TEI (Sperberg-McQueen & Burnard, 1994; Ide & Véronis, 1995). TEI is an international group designing standard encoding formats for various types of text. Their objective is to specify a minimum encoding level that will make texts transportable, i.e. they indicate what types of labels each text should have (Rundell, 1998a:16/1.15). In addition, for any particular type of text, a Document Type Definition, for short DTD, can be used as a framework. A DTD not only indicates which labels a text must contain, but also what sequence these labels must come in.

We are convinced that the level of detail in encoding an (entire) electronic corpus has to be related to the potential use that will be made of that corpus. Considering the huge cost and huge manual effort, pre-analysing and marking up African-language corpora right from the start (through word tokenisation, part-of-speech tagging, lemmatisation, syntactic parsing and markup) does not seem justifiable. As such, the corpora that will be discussed below give heed to Sinclair's advice: 'The safest policy is to keep the text as it is, unprocessed and clear of any other codes' (1991: 21).

Corpus query tools and query terminology

Before concluding this theoretical conspectus of electronic corpora, we must point out that corpora are of no use without powerful 'corpus query tools'. As a minimum requirement, such tools must be able to: (1) deal with huge numbers of text files, (2) handle files stored in plain texts as well as in markup format, (3) calculate basic statistics, (4) present alphabetical and frequency wordlists, and (5) provide concordance lines. There are quite a number of software packages available to perform these tasks, like Corpus Bench from Denmark, MonoConc from the US, WordSmith Tools from England, or an Access-based program developed at the University of Pretoria (cf. also Rundell, 1998b:16/3.17-19; Kennedy, 1998:259-267). Testing several, we singled out WordSmith Tools, for short WST (Scott, 2000), for its versatility in handling corpus queries.

This versatility also manifests itself in WST's ability to assemble, for any imaginable query, 'a' corpus through a selection of any number of the available electronic files. In other words, WST does not query 'the' corpus of a certain language (variety); rather, it queries any combination of sources which scholars deem useful at a certain point in time for their particular need(s). As such, given that the available range of electronic files is a large and structured collection, WST enables the near-instant move from a structured to an organic corpus.

It is important to note that a large 'query terminology' is connected with corpus query tools, of which the terms 'tokens' or 'running words', 'types', 'hapax legomena' and 'KWIC concordance' are absolutely basic. In corpus linguistics the terms 'tokens' or 'running words' stand for the total number of items in a corpus, and the term 'types' stands for the total number of different items in a corpus. The term 'hapax legomena' refers to those items which appear only once in a corpus[3]. Finally, 'KWIC concordance' is the acronym for 'keyword-in-context concordance' and is 'A word or phrase extracted from a text and listed in alphabetical, frequency or other order, together with the words occurring in its immediate environment' (Hartmann & James, 1998:79). When listing KWIC concordance lines (often shortened to just 'concordance lines', however) advanced corpus query tools such as WST even permit the use of wild cards as part of the 'keyword(s)'.

Electronic corpora – An exploration for the African languages

African corpora – Chimera or reality?

It is beyond doubt that any first approach to corpora for the African languages cannot even come close either to the size or thoroughness that characterises today's major English corpora. Nor do we have the necessary corpus traditions, nor the necessary linguistic descriptions, nor the necessary theoretical frameworks, nor the necessary human resources, nor the necessary funds, nor the necessary demand – to name but a few – to warrant such a tremendous effort. Nevertheless, an electronic corpus being a crucial aspect in modern linguistics, African corpora must be – and are being – built. The present endeavours in the South African corpus field of which we are aware are summarised in Table 1.

From Table 1 it is clear that most corpus work in South Africa revolves around the University of Pretoria, and to a lesser extent around the University of Port Elizabeth (in cooperation with the PE Technikon). At present, all the corpora listed in Table 1 are structured corpora, with the Pretoria Sepedi Corpus (PSC) starting to show characteristics of an organic corpus. Even though size and thoroughness of the South African corpora so far cannot compare with those of today's major English corpora, the point is that they cannot be compiled in isolation since important aspects from other comparable corpus projects must be taken into consideration. Indeed, from a theoretical perspective we can observe that all the latest relevant developments in corpus linguistics must be incorporated, and from a practical perspective we note that a watchful eye is to be kept on the compilation of African-language corpora outside South Africa. A summary of the latter corpora of which we are aware is shown in Table 2.

In order to illustrate the interplay between the

Table 1: Corpora of South Africa's eleven official languages

Language	Name	Acronym	Place(s)	Size
Afrikaans	Pretoria Afrikaans Corpus	PAfC	Pretoria	0.8 million
Afrikaans	Die Pharos-korpus van hedendaagse Afrikaans	PAK	Cape Town	17.6 million
English	Corpus of South African English	CoSAE	Port Elizabeth	>2 million
isiNdebele	Pretoria Ndebele Corpus	PNC	Pretoria	0.4 million
isiXhosa	Pretoria Xhosa Corpus	PXhC	Pretoria	1.4 million
isiXhosa	(pilot study)	–	Port Elizabeth	–
isZulu	Pretoria Zulu Corpus	PZC	Pretoria	0.7 million
isiZulu	(pilot study)	–	Durban	–
Sepedi	Pretoria Sepedi Corpus	PSC	Pretoria	4 million
Sesotho	Pretoria Sesotho Corpus	PSSC	Pretoria	0.2 million
Setswana	Pretoria Setswana Corpus	PSTC	Pretoria	1.2 million
siSwati	Pretoria Swati Corpus	PSwC	Pretoria	0.1 million
Tshivenda	Pretoria Tshivenda Corpus	PTC	Pretoria	0.2 million
Xitsonga	Pretoria Xitsonga Corpus	PXiC	Pretoria	1 million

Table 2: Corpora of African languages (excluding the South African ones)

Language	Name	Acronym	Place(s)	Size
ChiShona	ALLEX – Shona Corpus	–	Harare/Oslo/Gothenburg	2.2 million
Cilubà	Recall's Cilubà Corpus	RCC	Ghent	0.3 million
isiNdebele	ALLEX – Ndebele Corpus	–	Harare/Oslo/Gothenburg	0.7 million
Kiswahili	Kiswahili Internet Corpus	KIC	Pretoria/Ghent	1.7 million
Kiswahili	Helsinki Corpus of Swahili	HCT	Helsinki	1 million

South African African languages and African languages from the same language family, and in order to enable a thorough analysis of the practical issues involved in compiling a structured electronic corpus, languages from both Table 1 and Table 2 will be compared with regard to two different projects. On the one hand the creation of a small-size electronic corpus for Cilubà, namely Recall's Cilubà Corpus (RCC)[4], will be described thoroughly and appropriate reference will be made to different phases in the creation of the Pretoria Sepedi Corpus (PSC1, PSC2, etc.). On the other hand the main innovative features in building a large Internet corpus, namely the Kiswahili Internet Corpus (KIC), will be outlined and compared to similar possibilities for languages such as isiZulu, isiXhosa and Setswana.

The value of electronic corpora

In dictionary compilation, for example, the data provided by electronic corpora assist the lexicographer in several ways on both the macrostructural and microstructural levels. These issues will form the basis of detailed discussions in forthcoming publications and the discussion in this article will therefore be limited to a few basic examples.

The lexicographer is interested in at least two basic outputs of the electronic corpus, namely word-frequency counts and concordance lines. Word-frequency counts can be used in order to decide which data to include and how to include those data. When utilising word-frequency counts, the lexicographer should consider: (1) the rank or position of items in ordered frequency lists, (2) overall counts, being the total number of occurrences of items in the entire corpus, and (3) the distribution of those items across the different sub-corpora or sources.

From such frequency counts the lemma-sign list can be derived. A word with a substantial total count which also has a reasonable spreading across the different sub- texts will for example be a candidate for inclusion in the dictionary. The value of frequency counts is twofold. Firstly, to ensure that frequently used words are not accidentally omitted as in a typical shortcoming of the traditional way according to which lexicographers added words to the dictionaries 'as they crossed the compiler's way'. Secondly, to ensure that precious dictionary space is not utilized for words 'unlikely to be looked up' by the target user. Frequency counts obtained from the corpus thus assist the lexicographer in solving one of the basic problems in dictionary compilation namely, what to include and what to exclude from the dictionary. Such frequency counts can even be indicated in the dictionary itself. Different conventions can be used such as 'frequency bands', e.g. five filled diamonds indicating that the lemma sign occurs within, say, the 1000 most frequently used words in the language, four filled diamonds, for a lemma sign within the 2000 most frequently used words, etc. Apart from the issue inclusion versus omission, frequency counts also assist the lexicographer in avoiding quite a number of other inconsistencies such as unequal treatment of verbs and nouns in respect of derivations. Numerous examples are found in existing dictionaries where, say, the applicative or reflexive forms of certain verbs are lemmatized but for other verbs, sometimes even more frequently used ones, it is not done.

On the microstructural level, for the purpose of this article, it will suffice to mention that concordance lines culled from living-language sources supplement and support the lexicographer's (native-speaker) intuition. They take him/her to the heart of the actual usage of words through the display of the word(s) in context, allowing the lexicographer to see up to several dozens of contexts at a glance.

Such corpus lines assist the lexicographer in respect of sense distinctions, decisions on translation equivalents, the retrieval of typical collocations, the pinpointing of frequent clusters and the selection of representative, authentic examples to be included in the dictionary.Without a corpus the lexicographer is always in doubt whether he/she has covered all the relevant senses of a lemma sign in the definition or in setting up a translation equivalent paradigm.

Creating a small-size electronic corpus (RCC) – A traditional approach

From the onset one would do well to keep in mind that, for the African languages, the corpus compiler will in most cases be forced simply to add to the corpus what is available in print for the specific language(s). The following comparison will amply prove this point. If one brings together all the works in Cilubà or dealing with Cilubà one can possibly find, one obtains 100-odd written sources. Yet, during the creation of the Longman Lancaster English Language Corpus 'The help of over 100 academics was enlisted to make the actual recommendations for specific texts' (Summers, 1993:199). In other words, there were as many academics recommending specific texts at Longman, as the total available number of written sources in the Lubà language. This fact is of crucial importance, for it simply means that, eventually, every possible piece of text that has ever been written in Cilubà will have to be incorporated. In addition, if a Lubà corpus is ever to run into several millions of words, a very substantial amount of the 'hardest section' of a corpus, the spoken one, will have to be included.

Determining a 'useful' corpus size for RCC

Obviously, one cannot include all available sources at once, so one needs to determine a 'useful' size for a first electronic corpus. As will be discussed in forthcoming articles, the main purpose for creating Recall's Cilubà Corpus (RCC) was to provide data for the compilation of a projected pocket learner's dictionary containing 3000 lemma signs, and in a second phase we wished to use RCC for a small study in distributional phonetics. We therefore opted for the following straightforward approach. Basically, the Collins Cobuild English Language Dictionary (Sinclair, 1987), known as Cobuild1, was based on the earliest electronic mega-corpus, the 7.3-million-word Cobuild Main Corpus. As Cobuild1 contained roughly 70,000 'references', this meant that approximately 100 running words in the corpus were required for each reference in the dictionary. Our idea was to apply this same ratio between the size of RCC and the projected 3 000 lemma signs. Applying the ratio derived from Cobuild1 meant that we had to create a corpus consisting of 300,000 running words.

A. Small-size corpora and stability tests on PSC1, PSC2, etc.

A corpus of just 300,000 words does not even come close to the going rate of one million words. Yet, there are at least two good reasons to support the compilation of small-size corpora as a first approach. Firstly, from a theoretical perspective Kennedy notes:

> 'The new generation of mega-corpora is likely to dominate work in corpus linguistics for a considerable period. However, while definitive lexical descriptions and large commercial projects will be crucially dependent on them, such mega-corpora may not be appropriate or easily accessible for individual academic researchers [...] It is important therefore not to overlook the value of working with small corpora of one million words or less in the period ahead [...] There is nothing magic about one million words and no advantage in compiling a corpus of that size if a corpus of 200,000 words will do the job intended' (Kennedy, 1998:56-57, 73)

Secondly, from a practical African-language perspective we can draw conclusions from different phases in the creation of the Pretoria Sepedi Corpus (PSC1, PSC2, etc.). At different stages 'stability tests' were carried out on presumably highly used items on the one hand and seldom used items on the other. The aim of those stability tests was to determine whether conclusions drawn from a relatively small-size corpus are actually reliable. At present PSC stands at 4 million words, but for the purpose of this article we will only consider the first three consecutive stages, up to one million running words:

Phase 1 (PSC1): 109 322 words;
Phase 2 (PSC2): increased to 225 099 words;
Phase 3 (PSC3): increased to 1 160 550 words.

In order to determine the degree of stability of presumably highly used Sepedi vocabulary, the 100 items with the highest frequency in Phase 1 were selected. This was also done for Phase 2 and Phase 3. Upon comparing the rank numbers of the first 100 items in Phase 1 with the rank numbers of the first 100 items in Phase 2, it was noted that an amazing 90% of the items occurring within the first hundred positions in Phase 1 remained within the positions 1 to 100 in Phase 2. The ousted 10%, although having dropped out of the top 100, still retained very high rankings in Phase 2. This can be seen from the rank numbers listed in Section I of Table 5.

When these same first 100 items from Phase 1 were compared with the first 100 items of Phase 3, thus with the 100 most frequently occurring items in the one-million-word corpus, still as many as 84% of the Phase 1 top 100 items were represented in the Phase 3 top 100. Here, too, the ousted 16% still retained relatively high rankings in Phase 3. Moreover, if the occurrences of the names Mamahlo, Motšheletšhele, Lukas, Lesibana, Lebowa and the abbreviation mna. 'Mr.' are ignored, the per-

Table 5: Comparisons between the top 100 items in three consecutive stages of PSC

Section I			Section II			Section III		
Rank PSC1 versus PSC2			Rank PSC1 versus PSC3			Rank PSC2 versus PSC3		
Item	Phase 1	Phase 2	Item	Phase 1	Phase 3	Item	Phase 2	Phase 3
baka	96	149	Mamahlo	70	557	Lukas	75	316
Manahlo	70	146	Motšheletšhele	55	473	Lesibana	76	308
nama	81	140	Lukas	37	316	Lebowa	90	294
Motšheletšhele	55	123	Lesibana	38	308	mna.	89	271
fase	95	116	Lebowa	69	294	sekolo	66	172
dijo	100	113	mna.	57	271	gopola	100	146
gago	78	112	dijo	100	190	bjo	96	142
wena	89	111	baka	96	154	fa	78	131
tšwa	91	104	nama	81	151	mosadi	91	120
tsena	93	101	bjo	99	142	kudu	70	116
			fase	95	136	gae	77	113
			fa	61	131	nnete	80	112
			mosadi	83	120	nyaka	79	110
			kudu	74	116	tle	92	107
			gae	60	113	gabotse	99	105
			nyaka	92	110	mokgwa	81	103

centage of ousted items is only 10% and these items still occur within the top 200 in Phase 3. This can be seen from Section II of Table 5.

The stable pattern is confirmed when the Phase 2 top 100 is compared with the Phase 3 top 100. Once again, if the names and the abbreviation mna. are ignored, only 12% of the items were ousted from the Phase 2 top 100 and they still occur within the top 200 of Phase 3. This is evident from the rank-numbers presented in Section III of Table 5.

Phase 3 is more than five times larger than Phase 2, yet the occurrence of highly used items could have been accurately predicted. Therefore, as far as 'high-frequency items' are concerned, it can be pι edicted from these comparisons that increasing the size of a small-size corpus does not substantially influence the 'stability of the corpus'.

Conversely, one must also examine the 'stability of the corpus' when it comes to 'low-frequency items', i.e. vocabulary that seems to be rarely used. In other words, is it reasonable to expect that hapax legomena (hence, items which only occurred once) in Phase 2 will probably not occur more than, say, five times if the size of the corpus is also increased five times? This expectation is confirmed, since 58% of the items which occurred only once in Phase 2 still occur only once in Phase 3, and as many as 84% of the Phase 2 hapax legomena occur five times or less in Phase 3. Thus the frequency counts of only 16% of the Phase 2 hapax legomena were increased to more than the expected five. In Table 6 a number of randomly selected Phase 2 hapax legomena are listed together with their frequency counts in Phase 1 and Phase 3.

Table 6: A random selection of hapax legomena in PSC2, together with their frequencies in PSC1 and PSC3

Item	PSC1	PSC2	PSC3
baagši	–	1	2
baakaretša	–	1	1
baanegi	–	1	1
baanegwa	–	1	5
baapei	1	1	4
babagolo	–	1	1
babalele	–	1	5
babapadi	1	1	14
babinakudu	1	1	1
babinanoko	1	1	1

From these randomly selected examples it is clear that the count for babapadi 'players' increased substantially, but that the rest can still be regarded as low frequency items. Hence, it is apparent that items which had a low frequency in Phase 2 did not gain real ground when the corpus was enlarged more than five times. The outcome of this experiment shows that less-frequently-used items detected in a small-size corpus retain a low frequency value even if the corpus is substantially enlarged.

B. Small-size corpora and conjunctively versus disjunctively written African languages

Compared to a corpus of one million running words, both frequent and infrequent words detected in a small-size corpus seem to be stable. Yet, what does 'one million words' really mean? If one considers the African languages as a whole, one sees that – and now we overly simplify – some have been given a conjunctive orthography, while others have been given a disjunctive orthography. This is why the number of words for a simple sentence like 'I know you' (three words in English) might be anything from one to four words in an African language:

isiZulu	*ngiyakwazi*	*ngi-*	*-ya-*	*-ku--azi*
	I know you	I	[pres.]	you know
Cilubà	*ndi mukumanyè*	*ndi*	*mu-*	*-ku--manyè*
	I know you	I amI		you know
Sepedi	*ke a go tseba*	*ke*	*a*	*go tseba*
	I know you	I	[pres.]	you know

In isiZulu and Cilubà 'words' are written conjunctively, while in Sepedi 'words' are written disjunctively. But even among conjunctive orthographies there is a 'degree of conjunctiveness', where isiZulu is more conjunctive than for instance Cilubà. Therefore, a Lubà corpus of 300,000 words is not just 30% of a 1,000,000-word-large Sepedi corpus. Rather, the size of RCC (300,000 conjunctively written words) is roughly equivalent to twice the size of PSC2 (2 x 220,000 disjunctively written words). Hence, the fact that occurrences of both highly and rarely used words could have been predicted in a corpus like PSC2, becomes extremely important, for this implies that such predictions are also possible in a corpus like RCC.

Compiling RCC

From the above, it is clear that the creation of RCC had to result in a structured small-size electronic corpus of 300,000 running words. The three compilation steps will now be reviewed for the creation of RCC.

A. Designing RCC

As far as the overall design of RCC is concerned, we decided to follow a combined genre/topic stratification. The resulting corpus can be characterised as a general corpus built from both written and spoken sections, consisting of both sample and full texts chosen from diachronic sources. This categorisation is a result of endeavouring to create a structured corpus. Due to the scarcity of sources, RCC was bound to be general and diachronic. The strategy of including both sample texts and full texts, and both written and spoken material, was a structural decision. RCC itself consists of seven sub-corpora (i.e. Magazines, Traditional Stories, Informal Literature, Textbooks, Scientific Works, Religious Works and Miscellaneous Sources) of circa 40 000 words each and one sub-corpus (i.e. Poetry & Proverbs) of circa 20 000 words. The outline of RCC is thus straightforward. Each genre/topic area is given equal weight, except for the block Poetry and Proverbs, which is only half the size of the other blocks.

B. Collecting the texts of RCC

An exhaustive listing of all the material included in RCC is presented in Appendix A. It goes without saying that this enumeration, and especially the pinpointing as well as number of specific sources needed to reach the required total of words within each sub-corpus, was only finalised following the actual collection of the data. Actually, every time the quota of a particular block was reached, we could move to the next block.

As can be seen from Appendix A, we only used existing transcriptions and existing written material. As far as the oral section of RCC is concerned, all Traditional Stories (#10-13) are transcriptions of oral traditions, and the first Miscellaneous Source (#_26) is a transcription of an unscripted conversation. All these different spoken sections together amount to 17.4% of the corpus, while the other sources listed in Appendix A constitute the 82.6%-large written section.

The three ways mentioned of entering such existing sources into computer files (namely: a) electronic transfer, b) (re)keyboarding, and c) scanning) were used.

(a)We were fortunate to be able to make free use of a series of computerised files. Indeed, sources #11, #14-16, #19, #26-27, #31-33 were received in electronic format. None of these files, however, could be added to the corpus just like that. Actually, they all required considerable editing. Indeed, as a multitude of non-ASCII signs had been used to indicate the tonal dimensions, especially for the rising tones, all of them had to be standardised. In this process, rising tones were replaced with two dots over the respective vowels. The hardest aspect we had to deal with were the many manifold different overstrikes that had been created in WordPerfect. Each and every single low-toned nasal had to be searched and replaced manually. During this unrewarding task, all low-toned nasals were replaced with a backslash in front of the respective nasal. Eventually, the computerised files turned out to represent one third of the words in RCC.

(b)The remaining two thirds were either (re)keyboarded or scanned. Basically, all books and magazines were scanned. Unfortunately, quite a number of the old books and especially large sections of the magazines had to be completely (re)keyboarded. One kind of data that was (re)keyboarded right from the start were the texts accompanying strip cartoons in magazines, as well as all the lègends clarifying figures, drawings, pictures, maps, etc. in both books and magazines. The number of (re)keyboarded words ran into several tens of thousands.

(c)The largest part of RCC was scanned. In creating RCC, we did our utmost to actually read consistently through all the scanned and recognised data.

C. Encoding RCC

The 35 files making up RCC were all stored in a fixed-point typeface, in plain text. Hence, as announced, the text was kept unprocessed and clear of any other codes.

Querying RCC and the first applications of RCC

As projected, querying the small-size RCC proved to be invaluable during the compilation of a corpus-aided pocket learner's dictionary, the Beknopt woordenboek Cilubà-Nederlands (De Schryver & Kabuta, 1998). To the naked eye, the most obvious innovation in that dictionary is probably the fact that the different lemma have been provided with a frequency-annotation slot. As far as we know, this was the first attempt ever to include frequency-of-usage information in an African-language dictionary, and as such a small-size corpus enabled us to follow the trend set by such revolutionary dictionaries as the *Collins Cobuild English Dictionary* (*Cobuild2*) (Sinclair, 1995^2) or the *Longman Dictionary of Contemporary English, Third Edition* (*LDOCE3*) (Summers, 1995^3).

The same small-size RCC was also queried to provide the data for a small study in distributional phonetics. This study eventually led to proposals for a new approach to pursue phonetic research (De Schryver, 1999a).

The experiences with RCC show that it is beyond any doubt that even small-scale African-language corpus projects are a worthwhile venture.

Compiling a large Internet corpus (KIC) – An innovative approach

Many African languages are in a very similar position as Cilubà as far as available sources are concerned. For those languages it is advisable to start with the creation of small-size corpora, much as was described for RCC – the traditional approach. At the same time however, we must look ahead. Looking ahead means looking into the new millennium which has just started, and hence means realising that we have fully entered the Electronic Information Age. For African languages spoken in South Africa, like isiZulu, isiXhosa or Setswana, just as is the case for South African English and Afrikaans, the electronic master files of the texts of magazines and newspapers can already be obtained on disk under certain conditions. Yet, the Electronic Information Age opens even more promising doors, for it won't take too long before all South African languages will be used on the Internet. Hence, compiling corpora in the future will be much aided by simply downloading a variety of well-selected files from the Internet.

Actually, for at least one African language, namely Kiswahili, this innovative approach is already a reality. In order to test the potential of such a corpus, we kick-started the Kiswahili Internet Corpus (KIC) in mid October 1999. In just 10 days we were able to reach one million running words, and by adding daily news, this corpus stood at 1.3 million words before the start of the new millennium. From then onwards, we only spent a few minutes per week surfing the Internet in search of interesting additional sources. This way, and so far, the full KIC has grown to 1.7 million words – 'full' KIC, as there is no need to use all the files for every query. With well over two thousand files to choose from, there is quite a bit of room for experimenting with the concept of an 'organic corpus'. At present, the genre/topic stratification of KIC is as shown in Appendix B.

As a first application, KIC was queried in view of a thorough study about the lemmatisation of the verb in African languages. This study led to the introduction of a new lexicographic device, which we coined 'frequency-based tail slots' (De Schryver & Prinsloo, 2000).

Conclusion

We have shown clearly that electronic corpora are not just mere collections of texts which are assembled opportunistically and which are then stored in machine-readable form. Rather, a good corpus is a 'structured corpus' which, in due time, should take the form of an 'organic corpus'. We have also pointed out that issues such as size, design, method of text collection and level of annotation are very much dependent on the exact purpose for which the corpus is intended to be used. The different options described in the theoretical section should therefore be seen as potentially sound building blocks for the compilation of exactly that type of corpus which is needed for a particular type of research.

Even though Recall's Cilubà Corpus and the Kiswahili Internet Corpus are two very different approaches to the actual compilation of electronic corpora – with the first being a traditional and the second an innovative approach – we must conclude that they are both very relevant to the African context. Indeed, all existing corpora for the African languages were compiled using one, or a combination, of these approaches. The fact that both the creation of a small-size corpus and the building of an Internet corpus yielded several applications, and this in different linguistic fields, is encouraging. Similar applications for other African languages are bound to follow soon. In fact, some applications have already resulted in tangible products, such as the first Sepedi dictionary with frequency annotations (Prinsloo & De Schryver, 2000), or spellcheckers for isiXhosa, isiZulu, Sepedi and Setswana. The discipline of corpus linguistics for the African languages is definitely on the way up.

Notes

[1] This article is based on a paper read by the authors at the *First International Conference on Linguistics in Southern Africa*, held at the University of Cape Town, 12-14 January 2000. It combines sections of De Schryver's MA dissertation (1999b) and additional research. Since this article is being submitted for publication in a South African journal, necessary sensitivity with regard to the term 'Bantu' languages is exercised in our choice rather to use the term 'African' languages. Keep in mind, however, that the latter includes more than just the 'Bantu Language family'.

[2] It should be stressed that SGML is not only used for marking up texts in corpora. Any other kind of text in electronic form (e.g. books, magazines and newspapers), or even the computerised files of entire dictionaries, can be marked up in SGML. Yet, for Internet applications, SGML is too difficult to implement, so e.g. *HTML* (*HyperText Markup Language*) and *XML* (*Extensible Markup Language*) were developed. For more information, see as a launching pad Walsh (2000).

[3] The Greek *hapax legomenon* (singular of 'hapax legomrna') translates literally as 'something said only once'.

[4] Recall = Research Centre of African Languages and Literatures (Ghent University, Belgium).

References

Atkins BTS, Clear J & Ostler N. 1992. Corpus Design Criteria. *Journal of Literary and lLnguistic Computing* 7(1):1-16

Atkins BTS, Rundell MI & Gouws R. 1998. Afrilex-Salex '98, A training course in the compilation of bilingual dictionaries. (Unpublished course material of a tutorial held at the University of Pretoria, 7-18 September 1998.)

Bryan M. 1988. *SGML, An Author's Guide to the Standard Generalized Markup Language*. Wokingham: Addison-Wesley.

Carroll JB, Davies P & Richman B. eds. 1971. *The American Heritage Word Frequency Book*. New York: American Heritage Publishing Co.

Daenekindt M. 1999. *Dikke Spraak van Dale, Op zoek naar 10 miljoen woorden*, Gent Universiteit, 13(9):17-19.

De Schryver G-M & Kabuta NS. 1998. Beknopt woordenboek Cilubà-Nederlands & Kalombodi-mfùndilu kàà Cilubà (Spellingsgids Cilubà), Een op gebruiksfrequentie gebaseerd vertalend aanleerderslexicon met decodeerfunctie bestaande uit circa 3.000 strikt alfabetisch geordende lemma's & Mfùndilu wa myakù ìdì ìtàmba kumwèneka (De orthografie van de meest gangbare woorden). Ghent: Recall.

De Schryver G-M. 1999a. *Cilubà Phonetics, Proposals for a 'corpus-based phonetics from below'-approach.* Ghent: Recall.

De Schryver G-M. 1999b. Bantu Lexicography and the Concept of Simultaneous Feedback, Some preliminary observations on the introduction of a new methodology for the compilation of dictionaries with special reference to a bilingual learner's dictionary Cilubà-Dutch. (Unpublished MA dissertation, Ghent University.)

De Schryver G-M & Prinsloo DJ. 2000. Towards a Sound Lemmatisation Strategy for the Bantu Verb through the Use of Frequency-based Tail Slots – with special reference to Cilubà, Sepedi and Kiswahili. In Proceedings of the International Colloquium on Kiswahili in 2000, Dar es Salaam, 20-23 March 2000.

Fontenelle T *et al.* (eds). 1998. Actes Euralex'98 Proceedings. Liège: University of Liège.

Francis WN & Kucera H. 1964. *Manual of Information to Accompany 'A Standard Sample of Present-Day Edited American English, for Use with Digital Computers'.* Providence: Brown University.

Francis WN. 1992. Language Corpora BC. In Svartvik, Jan. ed. 1992: pp. 17-32.

Goldfarb CF. 1990. *The SGML Handbook*. Oxford: Clarendon Press.

Hartmann RRK & James G. 1998. *Dictionary of Lexicography*. London: Routledge.

Hurskainen A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology. An Application to Swahili, *Nordic Journal of African Studies* 1(1):87-122.

Ide N & Véronis J eds. 1995. *The Text Encoding Initiative, Background and Context*. Dordrecht: Kluwer Academic Publishers.

James G, Davison R, Cheung Hung-yeung A & Deerwester S. 1994. *English In Computer Science, A Corpus-Based Lexical Analysis*. Hong Kong: Longman Asia Limited.

Jeffery C. 2000. Projected Corpora of South Africa's Official Languages. In Programme & Abstracts of the First International Conference on Linguistics in Southern Africa, 12-14 January 2000, University of Cape Town: 71.

Kennedy G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Kilgarriff A. 1997. Putting Frequencies in the Dictionary, *International Journal of Lexicography*, **10**(2):135-155.

Kruyt JG & Dutilh MWF. 1997. A 38 Million Words Dutch Text Corpus and its Users, *Lexikos* **7**:229-244.

Leech GN. 1991. The State of the Art in Corpus Linguistics. In Aijmer K and Altenberg B (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*: London: Longman. pp 8-29.

Leech GN. 1992. Corpora and Theories of Linguistic Performance. In Svartvik J (ed) pp. 105-122.

Mills J. 1998. Lexicon Based Critical Tokenisation: An Algorithm. In Fontenelle, Thierry et al. (eds) pp. 213-220.

Moon R. 1998. On using spoken data in corpus lexicography. In Thierry *et al.* (eds) *Fontenelle*, pp. 347-355.

Prinsloo DJ & De Schryver G-M (eds). 2000. SeDiPro 1.0, *First Parallel Dictionary Sepêdi–English*. Pretoria: University of Pretoria.

Renouf A. 1987. Corpus Development. In Sinclair JM (ed). *Looking Up, An account of the Cobuild Project in lexical computing and the development of the Collins Cobuild English Language Dictionary*. London: Collins ELT. pp. 1-40.

Rundell M. 1998a. The Computer in Lexicography, An overview. In Atkins BT Sue *et al.* **16/1**:1-19.

Rundell M. 1998b. Corpus Exploration, Extracting lexicographic data from a corpus. In Atkins BT Sue *et al.* **16/3**: 1-19.

Scott M. Home page. 22 May 2000 <http://www.liv.ac.uk/~ms2928/wordsmith/screenshots>

Sinclair JM (ed). 1987. *Collins Cobuild English Language Dictionary*. London: HarperCollins Publishers.

Sinclair JM. 1991. *Corpus, Concordance. Collocation*. London: Oxford University Press.

Sinclair JM (ed.). 1995^{2}. *Collins Cobuild English Dictionary*. London: HarperCollins Publishers.

Sperberg-McQueen CM and Burnard L (eds). 1994. *Guidelines for Electronic Text Encoding and Interchange*. Chicago: Text Encoding Initiative.

Summers D. 1993. Longman/Lancaster English Language Corpus – Criteria and Design, *International Journal of Lexicography* **6(3)**:181-208.

Summers D (director). 1995^{3}. *Longman Dictionary of Contemporary English*, Third Edition. Harlow: Longman Dictionaries.

Summers D (s.d). [1996-1998]. Corpus Lexicography – The Importance of Representativeness in Relation to Frequency, Longman *Language Review* **3**:6-9.

Svartvik J (ed). 1992. Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin: Mouton de Gruyter.

Walsh N. "What is XML?" 22 May 2000 <http://www.xml.com/pub/98/10/guide0.html>

Appendix A: Structure of Recall's Cilubà Corpus (RCC), including the actual sources

Recall's Cilubà Corpus (RCC) (c. 300,000 words)

Magazines (c. 40,000 words)

1.Kàsayì Wetù 55/1. January 1973. Kananga, Zaïre.

2.Sangalayi 1998. December 1997. Brussels, Belgium.

3.Tabala ubale 3. s.d. [1995]. Kananga, Zaïre.

4.Tekemenayi 77. 1993. Kananga, Zaïre.

5.Tekemenayi 87. November 1994. Kananga, Zaïre.

6.Tekemenayi 90. March 1995. Kananga, Zaïre.

7.Tekemenayi 107. June 1997. Kananga, Congo.

8.Tekemenayi 117. June-July 1998. Kananga, Congo.

9.Tekemenayi 118. August 1998. Kananga, Congo.

Traditional Stories (c. 40,000 words)

10.Kazadi, Ntole, Ifwanga wa Pindi, Clémentine Faïk-Nzuji M. and Jean-Luc Sibertin-Blanc. 1984. Contes luba du Zaïre. In Ntole Kazadi et al. 1984. Contes luba et kongo du Zaïre: 13-111. Paris: Fleuve et Flamme.

11.Nzèmbèlà, Kayèmbe B. 1998a. Mikombo wa Kalèwo. (Unpublished manuscript.)

12.Sangalayi. 1996. Ngoma yà Tukwà Matèmbwà, Lusùmwìnù / Les tambours des guêpes, Conte luba. Brussels: Sangalayi.

13.Van Caeneghem, R. 1938. Kabundi sprookjes. Brussels: Vromant & Co.

Informal Literature (c. 40,000 words)

14.Kabuta, Ngo S. 1998b. Ngooyi Mèdar nè miyuukì mikwàbò. Ghent: Recall.

15.Kabuta, Ngo S. 1998. Mpooyi Ìzìdor. In Ngo S. Kabuta and Lufùlwàbò Ndaayà. 1998. Mpooyi Ìzìdor nè Mêyì àà Maamà Sâlà: 1-9, 15-32. Ghent: Recall.

16.Ndaayà, Lufùlwàbò et al. 1998. Mikàndà. In Ngo S. Kabuta. 1998a. Mfùndilu wa Cilubà nè ìmwè mikàndà: 71-175. Ghent: Recall.

Textbooks (c. 40,000 words)

17.(Anon.). 1982. Mukanda wa tshiluba 4. Kananga: Éditions de l'Archidiocèse.

18.Frères de la Charité. 1930. Mabela a mankenda a mubidi bwa bena nkongo. Tessenderloo: Imprimerie du Sacré-Cœur.

19. Kabuta, Ngo S. 1998. Mfùndilu wa Cilubà. In Ngo S. Kabuta. 1998a. Mfùndilu wa Cilubà nè ìmwè mikàndà: 1-69, 177-81. Ghent: Recall.

Scientific Works (c. 40,000 words)

20.(Anon.). 1932. Meyi ne mikandu ya mu Congo Belge. Hemptinne: Imprimerie Hemptinne St. Benoit.

21.Brock, Betsy. 1976. Dilongesha dia mianda ya bu muntu, 1 Mibelu ani malongesha bua mua kadimunda. Kinshasa: Éditions St. Paul Afrique.

22.Kabongo-Kanundowi, E. and Bilolo-Mubabinge. 1994. Kiipàcìlà kàà difùnda mikàndà mu Cilubà. In E. Kabongo-Kanundowi and Bilolo-Mubabinge. 1994. Mwandà wà Bumfùmù: 155-8. München: African University Studies.

23.Scheut. 1952. Mukanda wa pa bukwa bintu, Malongesha a mu kalasa kitanu. Leverville-Kikwit: Bibliothèque de l'étoile.

Religious Works (c. 40,000 words)

24.(Anon.). 1996. Difila bimanyinu bya lupandu (Sacrements) ne bisambukilu (Sacrementaux). Kinshasa: Rituel ad experimentum.

25.Bakole wa Ilunga (Ed.). 1994. Tshibangidilu. In Bakole wa Ilunga (Ed.). 1994. Mukanda wa Mvidi Mukulu, Muaku udi Mvidi Mukulu muambile bantu mu Tshiovo tshia Kale ne mu Tshiovo Tshipiatshipia: 1-61. Kinshasa: Verbum Bible.

Miscellaneous Sources (c. 40,000 words)

26.Ndaayà, Lufùlwàbò. 1998. Mêyì àà Maamà Sâlà. In Ngo S. Kabuta and Lufùlwàbò Ndaayà. 1998. Mpooyi Ìzìdor nè Mêyì àà Maamà Sâlà: 9-13, 33-67. Ghent: Recall.

27.Nzèmbèlà, Kayèmbe B. 1998b. Myanda misobolola mu Ngenzelu. (Unpublished manuscript.)

28.Walker, Alice. 1996. Mamu Zelia, Mukaji wa cipanda. Mbujimayi: CIAM.

Poetry & Proverbs (c. 20,000 words)

29.Bulanda, Nkèsè. 1995. Lubìlà lwà Mwikàlèèbwè. Kwetu Kundela 10: 9.

30.Dipumba, B. 1953. Disambila dyà baapàgaanò. Kongo-Overzee 19/5–30: 464-8.

31.Kabuta, Ngo S. 1998c. Le proverbe luba. (Unpublished manuscript.)

32.Kabuta, Ngo S. 1998d. Mwâ-Bâna. (Unpublished manuscript.)

33.Kabuta, Ngo S. 1998e. Zone L. (Unpublished manuscript.)

34.Kabwe, B. 1952. Dîba. Kongo-Overzee 18/2-3–7: 108.

35.Makolo, Muswaswa. 1990. Munanga wanyi. Kinshasa: Mwanza-Nkongolo.

Appendix B: Structure of the *Kiswahili* Internet Corpus (KIC)

Genre/Topic	Number of files	Number of tokens
Qur'an Main Text	114	154 110
Qur'an Comments	114	292 594
Bible: New Testament	27	160 163
Informal Literature		
• *pieces of Kiswahili satire and 'Uswahilini' short stories*	38	59 364
• *transcriptions of unscripted conversations*		
Newscasts (Radio and TV)		
• *Radio Deutsche Welle*	242	166 598
• *ITV Evening News*		
Newspapers & Magazines		
• *JUA - A swahili literary and general journal, RAI – Gazeti la kila wika, Majira – Gazeti huru la kila siiku, ...*	1 518	857 394
• *Radio Deutsche Welle*		
• *Alasiri, Nipashe, Jumapili, Komeshe, Taifa Letu, Lete Raha,...*		
Poetry		
• Kanga writings	7	5 338
• Methali za Kiswahili		
Various		
• advertisements, home pages, ...	22	25 384
Total	2 082	1 720 945