Nordic Journal of African Studies 11(2): 266-282 (2002)

WEB FOR/AS CORPUS: A PERSPECTIVE FOR THE AFRICAN LANGUAGES

GILLES-MAURICE DE SCHRYVER
Ghent University, Belgium & University of Pretoria, South Africa

ABSTRACT

In this article the potential of the multilingual Web to function *as* a corpus, in addition to a source *for* corpus creation, is examined. Despite the fact that English dominates the Web, and despite the fact that most work in corpus linguistics revolves around English, it will be argued that African languages do have a place in the bigger picture. Substantial African-language Web corpora can indeed already be compiled (Web for Corpus) and accessed (Web as Corpus), and the list of potential applications grows by the day.

Keywords: Web, corpus, parallel corpora, African languages, spelling and grammar checker, online web-as-corpus query software

INTRODUCTION

Web and corpus, the two key terms of the main title, hardly need any clarification. The World-Wide Web (often WWW or the Web) has been around since 1994, and is that part of the Internet linking documents, pictures and increasingly any combination of multimedia (including audio and video), the world over. To the younger generation browsing the Web by means of a mere sequence of mouse clicks, has become a most natural fact of life.

A (text) corpus in the language sciences has been around longer still. From a lexicographic perspective, Kilgarriff & Tugwell (2002: 125-127) divide the corpus era into four ages. The first age is pre-computer, i.e. linguistic data are collected on many thousands of paper slips, stored in boxes, and subsequently analysed and processed. The second age truly starts with the COBUILD[1] project in the early 1980s, which results in the first fully corpus-based dictionary in 1987 (Sinclair 1987a, 1987b). Corpora from this second age onwards are obviously 'electronic' (text) corpora. While the *COBUILD Main Corpus* contains just 7.3 million words in 1982 (Renouf 1987: 7), the 1995 *British National Corpus* (BNC)[2] already has 100 million words, and today's corpora, such as the *Bank of*

[1] COBUILD is an acronym for the *Collins Birmingham University International Language Database*.
[2] For the *British National Corpus* (BNC), see http://info.ox.ac.uk/bnc/.

English,[3] easily run into hundreds of millions of words. With increasing corpus sizes, the basic corpus-query packages of the second age give way to tools with which the data can be summarised statistically. The most outstanding contribution from this third age is doubtless Church & Hanks's (1989) *pointwise Mutual Information* and the *t-score* (both collocation statistics). During the first half of the 1990s corpora also (finally) enter the field of computational linguistics. While corpus sizes continue to skyrocket, one starts to realise that the data simply become too much to handle for any human being, and the focus today is shifting towards the creation of highly complex software tools for the automatic analysis of corpora. A first tangible and highly useful product of this fourth age is the so-called Word Sketch,[4] being "an automatically-produced summary of a word's behaviour" (Kilgarriff & Tugwell 2002: 136).

"The Web is an immense, multilingual, freely available corpus" reads the first sentence of a Call for Papers for a special issue of the journal *Computational Linguistics* (forthcoming). With currently something like a trillion words – that is one million million words – the Web indeed holds unprecedented possibilities to function 'as' a corpus. More down to earth, the Web can also be used as a source 'for' corpus creation. Although no one will dispute that the majority of the current corpora are corpora *of* English, and although the larger part of the data on the Web is *in* English, the non-English share is growing, including the African-language share. The aim of this article is therefore, as the sub-title indicates, to give a brief African-language perspective on the *Web for/as Corpus*.

1. WEB FOR CORPUS – AN AFRICAN-LANGUAGE PERSPECTIVE

1.1 AFRICAN-LANGUAGE DATA ON THE WEB

The number of corpora that can be accessed online, as well as the (text) sources and tools that are freely available in order to build and query one's own corpora, grows by the day. Attempting to provide an overview here is futile, especially since an excellent launching pad exists on the Web at the aptly named http://devoted.to/corpora. This well-structured site, maintained by David Lee, includes circa 1,000 annotated bookmarks for corpus-based linguists. The English-language bias of Lee's site is obviously a direct result of the English-language bias of the corpora compiled thus far, as well as of the nature of the Web itself. Regarding the latter Grefenstette observes, however, "that the proportion of non-English text to English is growing over time" (2002: 205). He calls this multilingual aspect a 'godsend' as "[n]on-English corpora were extremely rare and difficult to obtain before the WWW" (2002: 200-201).

[3] For the *Bank of English*, see http://titania.cobuild.collins.co.uk/boe_info.html.
[4] For Word Sketches, see http://www.itri.bton.ac.uk/~Adam.Kilgarriff/wordsketches.html.

An overview of the African-language text corpora built in the 'traditional' way – i.e. through scanning and optical character recognition (OCR) of hardcopy sources, transcription of recordings, and transfer of existing electronic files – can be found at ELC for ALL (Electronic Corpora for African-Language Linguistics, http://www.up.ac.za/academic/libarts/afrilang/elcforall.htm). The pioneers in this field are Prinsloo (1991) for Sepedi and Hurskainen (1992b) for Kiswahili. The first published account of the use of Web data 'for' African-language text-corpus creation appeared a decade later (De Schryver & Prinsloo 2000: 102, 106). Over a period of 10 days in mid October 1999, a *Kiswahili Internet Corpus* of 1 million words could be compiled. Since then, enough material has been downloaded to reach at least ten million words, and many more could be collected. Indeed, for the so-called 'big' African languages, the number of online, often daily-updated, sources is becoming substantial. For Kiswahili, for example, extensive coverage of the news can be found at the Tanzanian *IPP Media* (http://ippmedia.com/, several newspapers) or at the Kenyan *Kenya Broadcasting Corporation* (http://www.kbc.co.ke/), news, analysis and service from Germany and Europe at *Deutsche Welle* (http://kleist.dwelle.de/), information about Africa at the South African *Channel Africa* (http://www.channelafrica.org/), business news at the Tanzanian *Business Times* (http://www.bcstimes.com/, several newspapers), wide-ranging compilations of news at *Afrika Leo* (http://www.afrikaleo.com/), etc. etc. Apart from these tens of thousands of extra running words in Kiswahili posted daily, various other sources such as the continuously expanding *Archives of Popular Swahili* (http://www.pscw.uva.nl/lpca/aps/index.html), or more static sources such as the Koran (http://www.quranitukufu.com/) or the New Testament (http://www.ministrymall.com/ministrymall/swa/swa.html), can be visited for building Kiswahili corpora.

In a discussion on available electronic language data for the African languages, Hurskainen (1998) points out that the 'emphasis' is on Kiswahili. This is also true for the data on the Web. Nonetheless, many of the 'big' African languages are also well represented on the Web. The already mentioned *Deutsche Welle*, for instance, also provides news in Amharic and Hausa, while *Channel Africa* includes news in Silozi and Chinyanja. A dynamic source for isiZulu and isiXhosa is for example LitNet (http://www.mweb.co.za/litnet/).

1.2 Downloading Web Material

With all this African-language material available on the Web, and the Save button just one mouse click away, what stops linguists from compiling large-scale text corpora? Two points need some attention. Firstly, the controversial issue of copyright. The wisest and shortest answer has probably been provided by Adam Kilgarriff, President of the *Association for Computational Linguistics* (ACL) *Special Interest Group on the Lexicon* (SIGLEX):

> The hobgoblin of corpus research: copyright. As long as corpora are not being copied on, it is not infringed. (Kilgarriff 2002)

As long as the compiled corpora are thus solely manipulated for research purposes, and are not used or published commercially, linguists should be on the right track.

Secondly, what should be downloaded, and how? Recently, Grefenstette suggested that the comprehensive list of tools "needed to automatically retrieve and package" the material on the Web, would be "web crawlers, language identifiers, domain and genre classifiers, and morphological analysers, parts of speech taggers and shallow parsers" (2002: 205-206). With world languages such as English, German and Spanish in mind, this list is undeniably a feasible one, and current research indeed focuses on how software can automatically cull and analyse material from the Web. For the African languages, where the available amount of data is still relatively small, manually searching, saving and analysing the data still seems to be the best option. Some of the mentioned tools can come in handy though. Grefenstette himself suggests, in a slightly different context however:

> Instead of using a web crawler to collect text, one can also parasite commercial portals such as www.google.com, www.alltheweb.com, www.yahoo.com, www.altavista.com, etc. (Grefenstette 2002: 207)

This is indeed also how we have been hunting for African-language data over the years. One simply types in one or more key words unique to a particular African language, say in Google, and the gross of the returned hits surely constitutes pages of that particular language. Humans then perform the 'language identification', i.e. only those pages that are really couched in the language under scrutiny are kept. Once a page, and often through that page an entire web site, is found, the data must be saved. Experience has shown that it is wise to save *all* the data, i.e. the complete html file, or the complete pdf file, etc. as is. Even if one is, in most cases, only interested in the plain text, this can always be obtained in a second round, by saving the html, pdf, etc. as txt. During analysis one often feels the need to go back to the original file, and then it is ideal if the downloaded document contains all the original multimedia. The sum of the txt files then makes up the sought text corpus, and these can be stored in one folder, while the downloaded documents, stored in another folder, can be considered as a backup.

Grefenstette also suggests language-analysis software to 'automatically package' the corpus data. Unfortunately, to date Kiswahili remains the only African language for which an efficient morphological analyser has been built (cf. Hurskainen 1992a, 1995, 1996, 1999; and compare Hurskainen & Halme 2001). Several human-language technology (HLT) projects are under way however – in Pretoria for isiZulu, isiXhosa and Sepedi, in Harare for ChiShona and SiNdebele – and it is expected that in less than a decade there will be a handful of African-language morphological analysers. For the time being and except for Kiswahili,

human beings should thus still do the analysis. Actually, one of the aims of compiling African-language corpora, is exactly to have enough data to build the needed morphological analysers.

1.3 A Case Study – Web for 11 Parallel Corpora

In order to test the potential of an African-languages' *Web for Corpus* at this point in time, a full-blown case study was set up. Instead of attempting to build a straightforward corpus of a single African language, the aim was to see whether or not large parallel corpora could be brought together. The champion among the 'officially' multilingual African countries is South Africa, so it seemed feasible to hunt the Web for parallel texts in South Africa's nine official African languages, viz. Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, isiNdebele, isiXhosa and isiZulu. As most parallel texts are translations from English or Afrikaans – the other two official languages of South Africa – it seemed sound to keep those as well.

A few test runs quickly revealed, however, that trying to find parallel texts in *all eleven* languages simultaneously would be very hard indeed. For example, the official South African Government Information web site misses out on Tshivenda when listing all language versions of the *Constitution of the Republic of South Africa* (http://www.polity.org.za/govdocs/constitution/). Likewise, the official United Nations' web site with the translations of the *Universal Declaration of Human Rights* (http://www.unhchr.ch/udhr/navigate/alpha.htm), misses out on Tshivenda and Xitsonga. It was therefore decided to accept any parallel texts consisting of at least two of the eleven languages, with at least one in an African language.

Following five days of surfing in mid August 2002 – including saving, conversion to text format and archiving – an eleven-language parallel corpus with a combined count of over two million running words was assembled. Compared to the Kiswahili test of mid October 1999, when 10 days were needed to collect one million words, being able to collect double that amount in half the time, and being restricted to at least two parallel texts, clearly indicates that the African-language Web has truly grown. The results of this test are shown in Appendixes 1 and 2.

From these appendixes it can be seen that 45 sets of parallel texts were downloaded. Appendix 1 shows the exact word count for all parallel language texts in each of those sets, while Appendix 2 lists every topic and the corresponding Web addresses (also known as Universal Resource Locators (URLs)). Except for the *Universal Declaration of Human Rights* (HR), all sets were downloaded from a South African domain (.za) – which is logical. Less evident however is the fact that, out of 45 sets, only ten (A58, anc, cen, CoA, etq, mdd, mil, nqf, nsb, pub) contain parallel texts in *all eleven* official South African languages. With a push two more sets could be added (psb, wpl), yet for these

two, (large) sections of some of the translations are missing. The fact of the matter is thus that in only one quarter of the parallel sets, equal weight is given to all eleven languages. As the larger part of the sets come from official government bodies (with domain names ending in e.g. gov.za, pansalb.org.za, polity.org.za, sahrc.org.za, saqa.org.za, etc.) there is reason for concern. In Section 6 of the Constitution of the Republic of South Africa, that is the section on languages, one reads:

> (1) The official languages of the Republic are Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, Afrikaans, English, isiNdebele, isiXhosa and isiZulu.
> (2) Recognising the historically diminished use and status of the indigenous languages of our people, the state must take practical and positive measures to elevate the status and advance the use of these languages. …
> (4) … Without detracting from the provisions of subsection (2), all official languages must enjoy parity of esteem and must be treated equitably.
> (Constitution of the Republic of South Africa, Section 6)

Although not the purpose of the current investigation, it is clear that South African Web data, including the information distributed through the official government channels, are *not* treated 'equitably' in all eleven languages as stipulated by the Constitution. As the requirement of this case study was to include at least one African language per parallel set, the real picture is obviously many times worse. This finding corroborates Kamwangamalu's (2000), who found similar patterns for the television medium, in education, and in the government and administration at large.

Nonetheless, to return to the aim of the case study, eleven exact parallel corpora *could* effectively be brought together using the Web for parallel corpus creation. The combined size for all eleven is 348,467 words, or thus nearly 32,000 words on average per language. Furthermore, if not all eleven languages are required simultaneously, but for instance only language pairs such as Sepedi versus isiZulu or Tshivenda versus Xitsonga, then up to 32 parallel sets can be compared to one another, constituting up to a quarter million words per language. The value of the availability of such sets of parallel corpora for, say, corpus-based translation studies (CTS) is evident. They can also be used, as shown in Prinsloo & De Schryver (2002), in the design of a measurement instrument for the degree of conjunctivism / disjunctivism of the South African languages.

2. WEB AS CORPUS – AN AFRICAN-LANGUAGE PERSPECTIVE

2.1 WEB AS SPELLING AND GRAMMAR CHECKER

In §1.2 we indicated that it is possible to download the African-language Web data discussed in §1.1, and this was illustrated with a case study in §1.3. Instead of actually downloading and *saving* material to one's own computer 'for' corpus building, one can also simply *access* the data stored on the millions of servers around the world 'as' if they were a giant corpus. In the absence of sophisticated language-dependent tools, which is the case for all African languages, excluding Kiswahili (cf. §1.2), the ideal way to do so is of course to visit commercial portals. Except for http://www.alltheweb.com/ where the language can be set to Kiswahili, no other current commercial portal has the option to search for web pages written in a specific African language. Searches will thus not always be as fine-grained as one would want them to be. However, as long as the search words (or phrases) are unique to a particular African language, the returned material will also most probably be in that particular language. In addition, one can also restrict searches to a particular domain, e.g. .za.

A first straightforward use of the *Web as Corpus* could be to check the spelling of a particular African-language word. Zaenen, in imitation of ideas propounded by Grefenstette and Kilgarriff, formulates it as follows:

> If one doubts about the spelling of a word, one can type in the various versions and, relying on the possibly dubious assumption that a word is more often spelled right than wrong, see which one wins. (Zaenen 2002: 237)

Instead of 'inventing' data, we will illustrate this with a real case. Extract 1 shows the draft of an article abstract in Sepedi. Upon checking the proofs prior to publication (Nong *et al*. 2002), two spelling errors were suspected. Both instances have been highlighted below and concern aspiration, i.e. *bakgatatema* versus *bakgathatema* 'participants' and *bontšhi* versus *bontši* 'amount, number'.

Extract 1. Draft of an article abstract in Sepedi, with two spelling errors highlighted.

Maadingwa ge a bapetšwa le Mantšu a Setlogo go Sesotho sa Leboa – Kgopolo ya Bangwalapukuntšu. Maikemišetšo a taodišwana ye ke go nyakišiša, go ya ka kgopolo ya bangwalapukuntšu, ka fao baboledi ba Sesotho sa Leboa ba dirago kgetho ya mantšu magareng ga maadingwa le mantšu a setlogo polelong ye. Dipoelo tše di hweditšwego go tšwa go bakgatatema ba e lego baboledi ba Sesotho sa Leboa, banna le basadi, ba lekgolo (100) ba mengwaga ya go fapana, maemo a a fapanego a thuto, ba ba dulago mafelong ao a fapafapanego, bj.bj. di tla fetlekwa. Go ipontšha gore le ge dipoelo tša nyakišišo ye di laetša gore bontšhi bja bakgathatema bo kgetha go šomiša mantšu a setlogo go ena le maadingwa, bangwadi ba dipukuntšu ba swanetše go phafošwa mabapi le diphetogo tše di ka bago gona pateroneng ya kgetho ya tšhomišo ya mantšu. Dipoelo tša nyakišišo ye di bapetšwa le ka moo

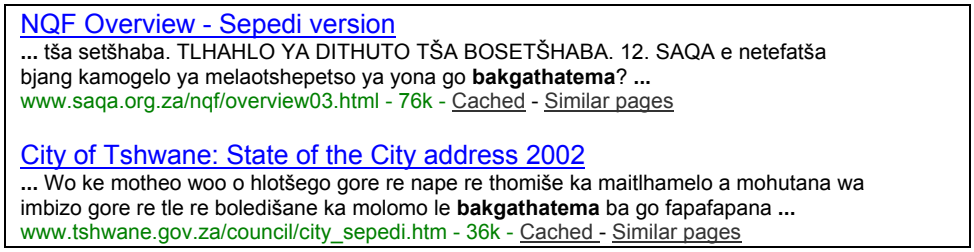
mantšu a tšwelelago kgafetšakgafetša go tšwa khophaseng gammogo le ka fao dipukuntšu tše di šetšego di le gona di šomišitšego mantšu ao ka gona.[5]

When one searches for *bakgatatema* on the Web using Google, the response is as shown in Extract 2.

Extract 2. A Google search for the incorrectly-spelled *bakgatatema* returns no hits.

A search for *bakgathatema*, however, returns 7 hits, all of them on pages in Sepedi. Extract 3 shows two of the hits.

Extract 3. A Google search for the correctly-spelled *bakgathatema* returns 7 hits (2 are shown).



[5] In English: **Loan Words versus Indigenous Words in Northern Sotho – A Lexicographic Perspective.** The aim of this article is to investigate, from a lexicographic perspective, the preferences of Northern Sotho mother-tongue speakers for loan words versus so-called 'traditional' or 'original' counterparts in the language. Results obtained from a survey conducted among 100 randomly selected mother-tongue speakers from different age and gender groups, backgrounds, places of residence, etc. will be analysed. It is shown that although the overwhelming preference of the respondents lies with the use of (more) indigenous words in comparison to loan words, lexicographers should be alerted to possible, even rapid, changes in this preference pattern. The results from the survey are compared throughout with frequency counts derived from a corpus as well as with current dictionary treatment.

One is obviously bound to conclude that *bakgathatema* is the correct spelling – which is true. Similarly, a Google search for *bontšhi* returns 2 Sepedi pages and for *bontši* 5. Although close, the spelling *bontši* 'wins' (in Zaenen's terms) – and this is also the preferred spelling today. Therefore, in the absence of a dictionary, corpus or spellchecker for a particular African language, one can try to use any of the commercial search engines in order to solve spelling issues.

A second use of the *Web as Corpus*, still through a commercial portal, is as an aid in solving grammatical questions. For example, in Sepedi there is a rule saying that the word (group) acting as the complement of the locative particle *go*, may not have the semantic feature [+ locative]. An instance such as **go bathong* 'to the people' is thus ungrammatical, as *batho* 'people' carries the locative suffix *-ng*. If in doubt, however, and if nothing else except for access to the Web is available, Google can again be called upon. In this case the ungrammatical form returns no hits, while a search for the 'exact phrase' *go batho* with as context word *Sepedi*, returns a dozen hits.

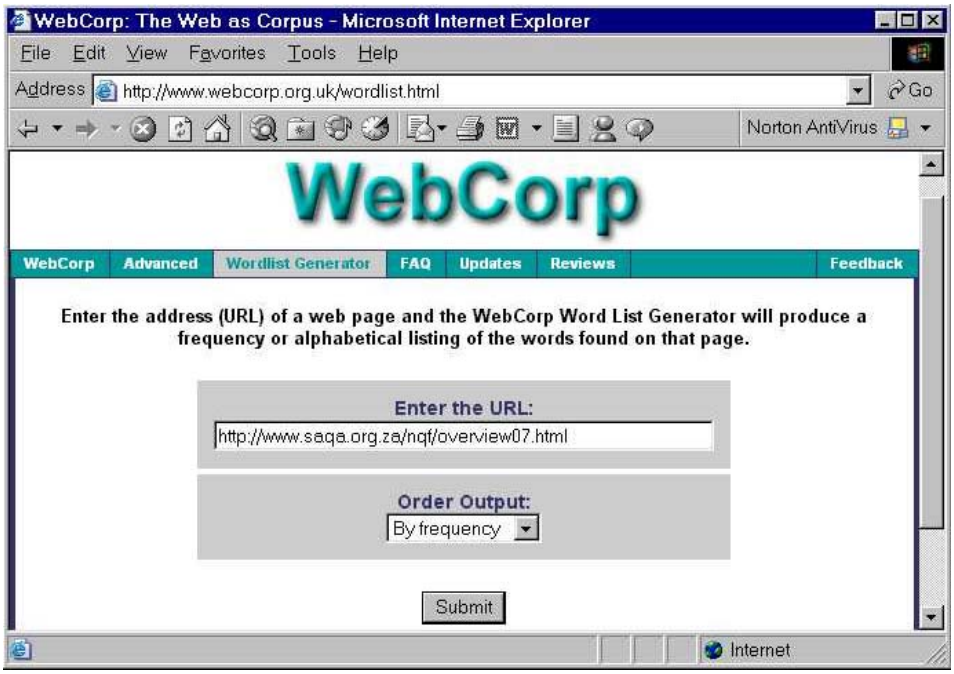
2.2 Online Web-as-Corpus Query Software

One of the exciting developments over the past few years has been the construction of language-independent web-as-corpus query software packages. Examples include *WebCorp* (http://www.webcorp.org.uk/), *KWiCFinder* and *WebKWiC* (the latter two available from http://miniappolis.com/). We will illustrate *WebCorp* for the African languages.

Like offline corpus-query software such as the now-standard *WordSmith Tools*,[6] basic corpus statistics can be calculated with *WebCorp*. To do so, one simply provides the URL of a certain web page, and WebCorp's Word List Generator produces word lists in frequency and alphabetical order. In Extract 4, e.g., the URL of a page in Tshivenda on South Africa's *National Qualifications Framework* is entered (http://www.saqa.org.za/nqf/overview07.html), while samples of the output are shown in Extract 5.

[6] For *WordSmith Tools*, see the Home Page of Mike Scott, the creator of the software: http://www.lexically.net (or else: http://www.liv.ac.uk/~ms2928).

Extract 4. The Word List Generator of WebCorp (the URL points to a page in Tshivenda).



Extract 5. Samples of the word lists generated by WebCorp's Word List Generator.

Word List - Frequency order **For: http://www.saqa.org.za/nqf/overview07.html**		**Word List - Alphabetical order** **For: http://www.saqa.org.za/nqf/overview07.html**	
Word	**Frequency**	**Word**	**Frequency**
u	613	…	…
ya	565	na	558
na	558	naa	1
ha	269	nadzangalelo	1
a	269	naho	1
vha	259	nahone	18
nga	253	nanga	4
dza	248	nangiwa	1
wa	236	nangiwaho	5
…	…	nanguludza	1
madzangano	50	nao	2
…	…	…	…

Of particular interest is the versatile Key Word in Context (KWIC) facility offered by WebCorp. Basically, WebCorp works 'on top of' one of the search engines (Google, AltaVista, MetaCrawler, FAST (AlltheWeb) or Northern Light) – the choice of which can be set by the user. One can search for words or phrases, including wildcards and patterns; context words to be in- or excluded can be provided; and one can optionally also specify the domains to search in (i.e. all of the pages on individual web sites, and/or all sites of particular types, and/or all sites of a number of individual countries). Various options are offered for the output format, as well as for the size of the concordance span. To give a straightforward example of the output, Extract 6 shows part of the output when WebCorp is instructed to search for any *ka * go* on South African servers.

Extract 6. Using WebCorp in search of locative trigrams in Sepedi.

WebCorp output for search term "ka * go"
Domain: ".za"
Producing output...

http://www.cosatu.org.za/samwu/womchartpedi.htm
Document Dated: Mon, 21 Aug 2000 10:28:35 GMT
Plain Text Word List

le gore ditaelo tša bakoloti di lokelwe	**ka mo go**	meputso ya bohle, le go maloko a rena
go godiša tsebo ya basadi le banna	**ka moka go**	phethagatša tekatekano ka mo go
ka moka go phethagatša tekatekano	**ka mo go**	SAMWU gomme go tloga moo, ya išwa
a yunione ka moka. Modiro wa taolo	**ka mo go**	yunione o na le maemo a fase gomme
kgontšha baemedi ba basadi go šoma	**ka mo go**	kgontšhago, bjalo ka senamelwa sa go
ge phesente ya basadi ba ba šomago	**ka mo go**	mmušo-selegae e gola. tsebiša,
ba tla dira lesolo bakeng sa basadi	**ka moka, go**	akaretša le bao ba dulago ka fase
le tlhokego ya methopo ya mehleng	**ka mo go**	ditirelo tša maphelo tša rena. Dikliniki
Moeno wo o tla dira maloko a SAMWU	**ka moka go**	godiša tsebo mo mafelong a mošomo
SAMWU e swanetše go dira ditlhagišo	**ka mo go**	melao ya naga ye e amanago le tlhori
Bjalo ka maloko a SAMWU re tla leka	**ka gona go**	kgobokanya gomme ra ba ra bitša
ba ra bitša maloko, banna le basadi	**ka moka, go**	šomiša dinepo le dinyakwa tša moeno

http://www.up.ac.za/academic/libarts/afrilang/newsletteroct98.htm
Document Dated: Wed, 02 Jan 2002 09:30:29 GMT
Plain Text Word List

hlaloša ditaba ka tsela ya maleba bjale	**ka ge go**	lemogilwe thutong ya go hwetša

The idea of this basic search for just *ka * go* in the South African domain was to see (a) if locative trigrams exist in Sepedi, and, if yes, (b) if these could also be found on the Web. As can be seen from the concordance lines, there are *several* instances of the locative trigram *ka mo go*. As this phenomenon has never been described in any grammar or publication, we intend to investigate this in depth in a forthcoming research article (De Schryver & Taljard forthcoming).

3. CONCLUSION

In this article we have explored the link between the Web and modern electronic corpora. This was done from two angles. Firstly, the Web was considered as a provider of data 'for' the creation of corpora, and secondly the potential of the Web was investigated 'as' a corpus in itself. In both cases African-language implementations and applications were briefly, yet carefully, scrutinised.

In the *Web for Corpus* section it was pointed out that substantial amounts of data can indeed be found on, and downloaded from, the Web today for many an African language. A full-blown case study even showed that the simultaneous compilation of eleven parallel Web corpora, *in casu* corpora for all the South African languages, has become a feasible endeavour.

From the *Web as Corpus* section three issues need to be remembered. Firstly, the Web can be used as a crude spellchecker for the African languages, to be employed as a line of first defence when no other sources or tools are available. Secondly, the Web can be surfed to confirm or even discover African-language grammatical patterns. Thirdly, online and language-independent web-as-corpus query software exists with which word distributions and 'live' concordance lines can be conjured up in the African languages.

REFERENCES

Church, Kenneth and Patrick Hanks. 1989.
Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*, pp. 76-83. Vancouver, CA.

Corréard, Marie-Hélène (ed.). 2002.
Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins. http://www.ims.uni-stuttgart.de/euralex/: EURALEX.

De Schryver, Gilles-Maurice and D.J. Prinsloo. 2000.
The compilation of electronic corpora, with special reference to the African languages. **Southern African Linguistics and Applied Language Studies** 18(1-4): 89-106.

De Schryver, Gilles-Maurice and Elsabé Taljard. forthcoming.
Locative Trigrams in Northern Sotho, Preceded by an Analysis of the Formative Bigrams.

Grefenstette, Gregory. 2002.
The WWW as a Resource for Lexicography. In Marie-Hélène Corréard (ed.), pp. 199-215.

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

1992b *Computer Archives of Swahili Language and Folklore – What is it?* **Nordic Journal of African Studies** 1(1): 123-127.

1995 *Information Retrieval and Two-directional Word Formation*. **Nordic Journal of African Studies** 4(2): 81-92.

1996 Disambiguation of Morphological Analysis in Bantu Languages. In *Proceedings of COLING-96. The 16th International Conference on Computational Linguistics, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pp. 568-573.

1998 *Maximizing the (re)usability of language data*. Workshop paper read at the International Conference on The Future of the Humanities in the Digital Age, September 25-28, 1998. Available at:

http://www.hit.uib.no/AcoHum/abs/hursk.htm.

1999 *SALAMA. Swahili Language Manager*. **Nordic Journal of African Studies** 8(2): 139-157.

Hurskainen, Arvi and Riikka Halme. 2001.
Mapping between Disjoining and Conjoining Writing Systems in Bantu Languages: Implementation on Kwanyama. **Nordic Journal of African Studies** 10(3): 399-414.

Kamwangamalu, Nkonko M. 2000.
A new language policy, old language practices: status planning for African languages in a multilingual South Africa. **South African Journal of African Languages** 20(1): 50-60.

Kilgarriff, Adam. 2002.
Web as Corpus. Poster session at the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13-17, 2002.

Kilgarriff, Adam and David Tugwell. 2002.
Sketching Words. In Marie-Hélène Corréard (ed.), pp. 125-137.

Nong, Salmina, Gilles-Maurice de Schryver and D.J. Prinsloo. 2002.
Loan Words versus Indigenous Words in Northern Sotho – A Lexicographic Perspective. **Lexikos** 12 (AFRILEX-reeks/series 12: 2002): 1-20.

Prinsloo, D.J. 1991.
Towards computer-assisted word frequency studies in Northern Sotho. **South African Journal of African Languages** 11(2): 54-60.

Prinsloo, D.J. and Gilles-Maurice de Schryver. 2002.
Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages. **Nordic Journal of African Studies** 11.

Renouf, Antoinette. 1987.
Corpus Development. In John M. Sinclair (ed.) 1987b, pp. 1-40.

Sinclair, John M. (ed.) 1987a.
Collins COBUILD English Language Dictionary. London: HarperCollins Publishers.

1987b *Looking Up, An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London: Collins ELT.

Zaenen, Annie. 2002.
Musings about the Impossible Electronic Dictionary. In Marie-Hélène Corréard (ed.), pp. 230-244.

Appendix 1. Word counts for 45 sets of 11 South African parallel texts.

Code	Sesotho	Sepedi	Tshivenda	Setswana	Xitsonga	Afrikaans	English	isiZulu	isiXhosa	Siswati	isiNdebele
A58	3622	2927	2812	2821	2678	2490	2547	1788	1749	1616	1747
agr	1148	—	—	1262	—	—	757	647	—	—	—
anc	1505	1389	1491	1490	1764	1320	1256	1074	1046	948	1058
ant	—	—	—	714	—	—	545	—	—	—	—
cen	802	626	670	714	628	527	506	390	475	340	368
cha	1188	—	1798	—	1945	956	927	797	—	—	—
CoA	1124	1038	1080	1113	1218	1397	1584	651	657	629	723
col	1737	1618	—	—	—	1413	1379	1049	980	—	—
Con	51794	51762	—	50157	52580	45112	44812	31579	34073	32218	31890
etq	7781	7187	6927	6949	5237	5837	5864	4444	4070	3999	3918
foo	1645	—	—	—	—	—	1368	—	—	—	—
hiv	11074	11975	—	—	12415	10363	9548	7392	7327	—	—
HR	2224	2411	—	2261	—	1757	1813	1140	1262	1874	1053
hyb	4299	3820	4351	3781	3881	2778	2716	2393	2228	—	—
hyt	1572	1545	1542	1485	1468	—	**X**	1032	1012	1017	944
hyu	2847	2345	3214	2375	2442	1783	—	1476	1407	1353	**X**
iba	4229	—	3318	—	4541	4690	—	2920	—	—	—
kmd	1163	—	1140	—	—	964	829	—	676	675	690
lit	—	—	3486	—	—	—	2563	—	—	—	—
lot	605	—	621	—	625	561	590	411	486	—	—
lra	—	—	2537	—	—	2195	2279	2003	1754	—	—
mdd	1829	1600	1458	1489	2039	1253	1268	1262	1025	879	1137

*Legend: — = no parallel text available; **X** = parallel text available, but a (large) part is missing or does not entirely correspond; **Code** = see Appendix 2.*

Code	Sesotho	Sepedi	Tshivenda	Setswana	Xitsonga	Afrikaans	English	isiZulu	isiXhosa	Siswati	isiNdebele
mil	524	510	500	492	485	448	420	309	350	269	332
nda	—	—	—	1261	—	785	774	646	—	—	—
nqf	13162	11306	11305	10558	9976	8499	8473	6564	6731	6049	6168
nsb	12631	10644	10919	10618	8485	8697	9036	6496	5499	6396	6032
nsi	283	249	259	235	257	206	168	—	147	143	—
Pan	7329	6244	6370	—	5973	5225	5446	—	3261	—	3810
pro	1548	1432	—	1319	1411	1159	1135	—	913	881	904
psb	**X**	**X**	2043	2160	1856	1755	1819	1678	**X**	1294	1399
pub	1521	1489	1441	1291	1374	1401	1366	970	1073	929	879
pw	—	15002	14286	12833	13627	—	—	—	7734	—	8076
R9f	36124	38024	34462	34130	32600	29436	29337	—	—	20528	21029
R9h	42697	42892	40162	39346	41494	33322	33038	—	—	24554	23403
R9o	9945	10125	—	9421	9033	7483	7520	7072	5536	5786	5523
R9s	19235	—	16693	17448	15973	15773	16540	—	—	10072	10335
rab	—	—	—	1047	—	—	815	—	565	—	—
sch	—	2029	—	—	—	1803	1730	1244	1262	—	—
sfm	—	1631	1684	—	—	1345	—	1038	974	—	—
tap	—	—	—	—	—	—	939	—	688	—	—
tic	1726	—	—	1589	—	—	1158	—	—	—	—
Tir	—	10580	—	—	—	7760	7577	—	5396	—	—
tra	329	—	—	—	284	249	248	200	—	175	200
wpl	8938	8814	9504	8942	8889	7743	7488	5990	5664	**X**	6583
zoo	—	—	—	815	—	—	662	—	—	—	—

Appendix 2. Universal Resource Locators (URLs) for 45 sets of 11 South African parallel texts (*all URLs last accessed in mid August 2002*)

Code	Universal Resource Locator (URL)
A58	South African Qualifications Authority Act (Act 58 of 1995) http://www.saqa.org.za/publications/legsregs/index.htm#legs
agr	2001 – Strategic Plan for South African Agriculture http://www.nda.agric.za/publications/publications.asp?category=Policy+documents
anc	What is the African National Congress? http://www.anc.org.za/about/anc.html
ant	Animal health: Anthrax http://www.nda.agric.za/publications/publications.asp?category=Info+Pak
cen	Census at School http://www.censusatschool.org.za/
cha	AIDS & HIV Charter http://www.aidsconsortium.org.za/charter/charterhome.html
CoA	National Coat of Arms http://www.gov.za/symbols/coatofarms.htm
col	co.za Domain Administration http://co.za/
Con	Constitution of the Republic of South Africa http://www.polity.org.za/govdocs/constitution/
etq	Education & Training Quality Assurance Bodies Regulations (Regulation 1127 of 8 September 1998) http://www.saqa.org.za/publications/legsregs/index.htm#legs
foo	Food safety: Food preparation in the home http://www.nda.agric.za/publications/publications.asp?category=Info+Pak
hiv	The HIV/AIDS – Emergency – Guidelines for Educators http://education.pwv.gov.za/HIVAIDS_Folder/Aids_Index.htm
HR	Universal Declaration of Human Rights http://www.unhchr.ch/udhr/navigate/alpha.htm
hyb	The Policy of Basic Household Sanitation Made Easy http://www.dwaf.gov.za/dir_ws/content/lids/sanitation.htm
hyt	Sanitation Technology Options http://www.dwaf.gov.za/dir_ws/content/lids/sanitation.htm
hyu	Understanding South African Sanitation Needs http://www.dwaf.gov.za/dir_ws/content/lids/sanitation.htm
iba	The Independent Broadcasting Authority: Position Paper on South African Content http://iba.org.za/policypg.htm
kmd	Knowledge Management Development Initiative http://www.km-debate.co.za/frequently_asked_questions.htm
lit	Fruit: Cultivating litchis http://www.nda.agric.za/publications/publications.asp?category=Info+Pak
lot	SA's National Lottery http://www.nationallottery.co.za/IE/HTPL/HTPE.htm
lra	Land Redistribution for Agricultural Development http://www.nda.agric.za/publications/publications.asp?category=General+publications
mdd	Media Development and Diversity Agency http://www.gov.za/documents/2000/mdda/
mil	The Code of Conduct – Military Members http://www.mil.za/Articles&Papers/CodeofConduct/
nda	Executive Summary of The Sector Plan for South African Agriculture http://www.nda.agric.za/docs/sectorplan/sectorplan.htm

nqf	An Overview of the National Qualifications Framework http://www.saqa.org.za/nqf/overview.html
nsb	National Standards Bodies Regulations (Regulation 452 of 28 March 1998) http://www.saqa.org.za/publications/legsregs/index.htm#legs
nsi	National Spatial Information Framework http://www.nsif.org.za/Translations.pdf
Pan	PanSALB's Position on the Promotion of Multilingualism in South Africa http://www.pansalb.org.za/pub.htm
pro	Protecting Your Rights http://www.sahrc.org.za/pamphlets.htm
psb	Pan South African Language Board (PanSALB) http://www.pansalb.org.za/about.htm
pub	The Code of Conduct – Public Servants http://www.mil.za/Articles&Papers/CodeofConduct/
pw	Worxnews April 2002 http://www.publicworks.gov.za/docs/newsletter/
R9f	Rev. Nat. Curriculum Statement Grades R-9 (Schools) – First Additional Language http://education.pwv.gov.za/DoE_Sites/Curriculum/Final%20curriculum/policy/policy.htm
R9h	Rev. Nat. Curriculum Statement Grades R-9 (Schools) – Home Language http://education.pwv.gov.za/DoE_Sites/Curriculum/Final%20curriculum/policy/policy.htm
R9o	Rev. Nat. Curriculum Statement Grades R-9 (Schools) – Overview http://education.pwv.gov.za/DoE_Sites/Curriculum/Final%20curriculum/policy/policy.htm
R9s	Rev. Nat. Curriculum Statement Grades R-9 (Schools) – Second Additional Language http://education.pwv.gov.za/DoE_Sites/Curriculum/Final%20curriculum/policy/policy.htm
rab	Animal health: Rabies – a killer disease http://www.nda.agric.za/publications/publications.asp?category=Info+Pak
sch	Rights and Responsibilities of Parents – A Guide to Public School Policy http://education.pwv.gov.za/DoE_Sites/Enab_env_qual_educ/ed_HR_dev.htm
sfm	Chief Directorate: Forestry http://www.dwaf.gov.za/Dir_Forestry/SFM/Default.asp
tap	Animal health: Tapeworm http://www.nda.agric.za/publications/publications.asp?category=Info+Pak
tic	Animal health: Tick-borne diseases in ruminants http://www.nda.agric.za/publications/publications.asp?category=Info+Pak
Tir	Tirisano: Working Together to Build a South African Education and Training System for the 21st Century http://education.pwv.gov.za/Tirisano_Folder/Tirisano_Index.htm
tra	Invitation to be Trained as a Translator or Interpreter http://www.pansalb.org.za/Advertise/Advertise.htm
wpl	A Short Guide to the White Paper on Local Government http://www.local.gov.za/DCD/policydocs/policymn.html
zoo	Animal health: Zoonosis http://www.nda.agric.za/publications/publications.asp?category=Info+Pak