Corpus-based language teaching: an African Language perspective

Elsabé Taljard

Abstract

Studies on corpus-based language teaching are notably absent within the South African educational context; more so with regard to the teaching of African languages. This article explores the possibilities offered by the availability of an electronic corpus to enhance language teaching, and more specifically, the teaching of Northern Sotho as a second additional language at first year university level to first time learners of the language. Particular attention is paid to corpus-based selection and sequencing of learning material, an activity that has hitherto depended on anecdotal evidence and the intuition of the language teacher. A critical evaluation of existing pedagogical material for Northern Sotho reveals that although excellent sources of reference, these works are inadequate for the purpose of teaching Northern Sotho to first time learners. It is indicated that information gleaned from a corpus provides the language teacher with guidance on both micro and macro level with regard to selection and sequencing of learning content.

Introduction

When the compilation of corpora for the South African Bantu languages of South Africa was first started in the late 1990s, the envisaged application of these corpora was mainly of a lexicographic nature. As a result, corpus-based lexicography is currently the norm, rather than the exception in South Africa, especially with regard to the South African Bantu languages. The use of corpora has since been extended to other disciplines, and a number of corpus-based studies on terminology, translation and linguistics have been published. Notably absent, however, are studies on corpora and the teaching of the South African Bantu languages, particularly as first and second additional languages – the availability of corpora has clearly not revolutionized the teaching of these languages. The aim of this article is therefore to explore the possibilities offered by the availability of a corpus to enhance language teaching; more specifically, the teaching of Northern Sotho as a second additional language at first year university level to first time learners of the language.

Contextualization

Teaching of a Bantu language as a first and/or second additional language is particularly relevant within the South African higher education context. It is common knowledge that since the mid-1990s, departments of African languages at South African Universities have been experiencing severe difficulties. Student numbers in these departments have dropped dramatically and, as is wont within the current framework of state subsidizing, many of these departments have been forced to close down or to amalgamate with other language departments and be brought together with other languages and/or language related studies in so-called school structures. It is therefore a matter of survival for those departments and schools still teaching South African Bantu languages to provide for the academic needs of as wide a spectrum as possible of potential students, including those who have no knowledge of a Bantu language, but who would want to become proficient in one or more of these languages, and who may even want to present it as one of their major subjects for degree purposes. The focus of this article is the supporting role that a corpus can play in the design of a module covering one semester (28 academic weeks) on the first year level. This module would represent a first acquaintance with Northern Sotho, but it could also form the basis for further study, should the student choose to continue his or her study of the language.

The selection of linguistic content for any module needs to take cognisance of the academic requirements, as embodied in the outcomes stated for the module. Currently, these are stated as follows:

After having completed this module, students must be able to:

- Properly pronounce Northern Sotho speech sounds, words and sentences
- Perform specific speech acts, such as greeting, giving commands and making statements

- Produce and understand simple sentences in the three basic tenses, in both positive and negative form
- Formulate and answer questions by using selected question words
- Illustrate a thorough knowledge of the structure and meaning of selected linguistic structures.

NOTE: except for the first outcome, all outcomes listed above pertain to both written and oral form

From these required outcomes it is clear that the aim of the module is to endow students with a general, basic competence in Northern Sotho, and that it strives to strike a balance between spoken communicative competence and a basic, but sound theoretical knowledge of the grammar of the language. This aim needs to be the deciding factor in the design of the module. The approach envisaged for this module therefore places a high premium on exposing students to high frequency language usage, not only on the lexical level, but also with regard to grammar structures. A corpus-based selection of linguistic content therefore needs to balance three aspects, i.e. frequency of occurrence of certain linguistic elements (both on lexical and structural level), the communicative functionality of these elements and the grammatical knowledge needed to support the communicative function, specifically with regard to language production. The question that would determine the selection of linguistic content is therefore: Is what is frequent necessarily functional from a communicative perspective, and secondly, how much theoretical knowledge is necessary to internalize and incorporate such elements into functional language usage on a beginner’s level?

Related to the selection of linguistic content, is the sequencing of the content. The order in which linguistic content is presented is a crucial aspect in the design of any language course. It is tacitly assumed by both language teachers and students that grammatical content offered in the beginning of a language course is somehow easier, more basic and more important than whatever follows later in the course. The decision as to which content is easy to master and should therefore be introduced in an early stage, is usually based on anecdotal evidence and the intuition of the language teacher. However, as pointed out by Biber and Conrad (s.d), decisions about the sequencing of material are not served well by intuition and anecdotal evidence. They claim that language teachers' intuition is often inconsistent and that they often do not notice the most typical grammatical features, simply because they are so common. It is only when language teachers use a data-based source of information such as a corpus that these kinds of oversights can be eliminated.

Before one can venture into the possible advantages of a corpus as a resource for language teaching, a few brief remarks on corpora seem in order, especially since corpus linguistics and especially corpus-based language teaching has not yet made serious inroads into the South African language scene.

A brief introduction to language corpora

Starting with what may be regarded as an overly simplified definition, a corpus can be described as a large collection of authentic written and spoken texts, created in genuine communicative situations that have been gathered in electronic format according to a specific set of criteria. A corpus should furthermore as far as possible be representative of the language or language variety for which it is used as a data source (Bowker and Pearson 2002:1, Sinclair 2004a). Apart from representativeness, the notion of balance is also crucial in corpus design. A balanced corpus should contain proportions of different kinds of texts which correspond with informed and intuitive judgements. Sinclair (2004a) states that most general corpora tend to be badly balanced because they do not contain enough spoken data. He suggests that the ratio spoken language : written texts should be 10:1, since most people experience about 90% spoken language to 10% of written material. He does however recognize that having a corpus which is balanced and representative should be regarded as an ideal to be aspired to, rather than an absolute requirement. This is especially true when working with lesser resourced languages such as the South African Bantu languages.

Different types of corpora are distinguished, e.g. monolingual versus multilingual corpora, general versus special corpora, etc., but within the realm of corpus-based language teaching a particularly relevant distinction is the one made between learner corpora and general corpora. Learner corpora contain language or texts produced by non-native speakers during the language aquisition process. This is used in error analysis, language acquisition and interlanguage studies. The use of learner corpora is particularly prevalent in the teaching of English as a second language. See in this regard Pravec (2002) for a detailed discussion of the various learner corpora available for learning English as a second language (ESL). At this stage, no learner corpus is available for Northern Sotho, mainly because of a dearth of texts produced by learners of the language. The corpus utilized for this particular study is the expanded 2009 version of the University of *Pretoria Sepedi Corpus* (PSC), consisting of 7.5 milion words or tokens. It is a general corpus, representing standard written Northern Sotho, covering a variety of genres and text types – time and financial constraints have however thus far prevented the inclusion of transcribed spoken texts. A part-of-speech (POS) tagged version of this corpus has been produced in collaboration with colleagues from the University of Ghent and the University of Antwerp and wherever applicable, the tagged version of the corpus is used. Compare De Schryver and De Pauw (2007) for a detailed description of the annotation procedure that was used to POS tag the PSC.

In order to extract relevant information from electronic corpora, corpus query tools are used. For this particular study WordSmith Tools version 5 (http://www.lexically.net/wordsmith/version5/index.html) is utilized. It is a suite of corpus linguistics tools used for finding patterns in a language. The tools include a concordancer, word-listing facilities, a tool for computing the keywords of a text or genre, and a series of other utilities. These tools are necessary to unlock the linguistic data contained in the corpus and to present it in a user-friendly format to the researcher or language teacher for analysis and interpretation.

Corpus as an aid to language teaching

In the introduction to this article, it was hinted that the availability of corpora was not all that enthusiastically welcomed by teachers of the South African Bantu languages. This reluctance to incorporate electronic language corpora into language teaching is by no means a solely South African phenomenon, as is clear from a statement by Römer (2010:18) who is 'hesitant to say that corpora and corpus tools have … fully "arrived" on the pedagogical landscape'. Sinclair (2004b:271) states that from a classroom perspective, the emergence of corpora may not seem to be good news, since the use of a corpus invariably fails to confirm the consensus view of language that has been considered adequate for most classrooms for many years. The resistance to the use of corpora can only be overcome if teachers (and grammarians) are well-informed about the possible role a corpus can play in the teaching of language.

The availability of a corpus can contribute on various levels to the effective teaching of language, whether it is first or additional language teaching, cf. Gabrielatos (2005), Biber and Conrad (s.d.) and Conrad (2000):

- When queried with the appropriate software, a corpus can provide valuable information on frequency of use of both lexical elements and grammatical structures. The extent to which it can be queried for grammatical structures and larger language patterns depends on the level of sophistication with which the corpus has been tagged.
- A corpus provides information on the often observed discrepancy between intuition and attested use, and thus influences the content and design of syllabi.
- Native speaker corpora provide a source of attested examples, which can be used for the compilation of corpus-based learning materials (and software). This is in contrast with the current practice, especially for the South African Bantu languages, where illustrative examples are mostly fabricated ones, based on the intuition of the language teacher.
- Corpus-based linguistic research can bridge some of the gaps left by introspection-based research, and thus contribute to more accurate language description. This in turn feeds into the compilation of improved pedagogical grammars and dictionaries.
- Corpora allow comparison of language use across different registers and thus provide information on appropriate language use within different discourse situations.

Recski (2006) indicates that a number of studies have observed discrepancies between corpus findings and the selection of linguistic content in ESL (English Second Language) and EFL (English First language) textbooks and curricula. For Northern

Sotho, no attempts have thus far been made to investigate the alignment between corpus findings and the linguistic content of Northern Sotho curricula in general, and even less for specific courses aimed at pre-defined target student bodies.

The above observations are based on the usefulness of corpora for the teaching of English as second or foreign language. English is arguably the best resourced language in the world with regard to the availability of sophisticated large scale electronic corpora and dedicated corpus query software, therefore not all of these observations will be equally applicable to the current scenario regarding the teaching of Northern Sotho, which has far fewer resources. The question that needs to be answered is to what extent the currently available corpus can support the teaching of Northern Sotho to beginner learners.

Lastly, when the incorporation of a corpus into language teaching is contemplated, a decision needs to be taken as to exactly how the corpus will be integrated into the teaching process. Corpora can be integrated into language teaching in one of two ways. The first approach requires only the language teacher/lecturer to have access to the corpus, and the skills to effectively use the software needed to query the corpus and thus extract data relevant for language teaching from it. In this scenario, which is the one we envisage in the current article, the corpus is mainly used for selection of linguistic content, and also for the development and compilation of learning material. The second option calls for the availability of the corpus and computer facilities to learners, in order to enable them to directly access the corpus. This approach is termed data-driven learning (DDL), a term first used by Johns (Bernardini 2004:16). Here, the function of the teacher/lecturer is that of facilitator of learning, whereas the learner is researcher who is guided towards the discovery of facts about the language (s)he is learning. These two approaches are of course not mutually exclusive. Within the context of the present study, and taking the profile of our target learners into consideration, we have decided to utilize the first option, the reason being that the students registered for this particular module find themselves on a zero entry level. Consequently, confrontation with pages of concordances may prove to be rather intimidating. Furthermore, it would make little sense in a module aimed at first time learners of a language to expect them to intelligently interpret the kind of data offered by a corpus, even if the data are presented to them in the form of pre-edited concordance lines. As Mauranen (2004) points out, noticing things in corpus data is an acquired skill, even for sophisticated language learners.

Before making a case for integrating a corpus into the design of this particular module, a critical evaluation of currently available pedagogical material for Northern Sotho needs to be given in order to ascertain whether these resources are not perhaps sufficient in assisting the language teacher to facilitate reaching the stated outcomes of the module. After all, these sources have stood language teachers in good stead for the past number of decades. Only after shortcomings in these resources have been identified, can the supporting role of a corpus be properly evaluated and delineated.

Critical evaluation of currently available pedagogical material for Northern Sotho

Although Northern Sotho is one of South Africa's lesser resourced languages, a number of grammatical descriptions are available for this language. The following are generally regarded as standard works on the grammar of Northern Sotho:

Handboek van Noord-Sotho (1969) by D Ziervogel
Introduction to the Grammar of Northern Sotho (1985) by D P Lombard, E B van Wyk and P C Mokgokong
Northern Sotho for First-years (1992) by E B van Wyk, P S Groenewald, D J Prinsloo, J H M Kock and E Taljard
A linguistic analysis of Northern Sotho (1994) by G Poulos and L J Louwrens
Aspects of Northern Sotho grammar (1994) by L J Louwrens

Although these grammars present the reader with what may be excellent and detailed grammatical descriptions of the language, from a pedagogical perspective they are of little value to a first time learner of Northern Sotho, especially in a learning situation where strong emphasis is placed on the acquisition of not only grammatical knowledge, but also of an oral communicative competence. The stated aim of these textbooks is mostly to provide insights into the grammar of Northern Sotho, with little attention being paid to communicative functionality and / or patterns of actual language usage. Therefore, it is perhaps not wholly fair to evaluate them as pedagogical materials, but since no other materials are available, these works are the only options available to language teachers. From a pedagogical point of view, these grammars are inadequate to be presented as learning material to first time learners of Northern Sotho, even though their value as sources of sound grammatical descriptions cannot be overestimated. Van Wyk et al (1992:foreword) state in the foreword to their *Northern Sotho for First-Years* that "(t)he purpose of this book is to satisfy the longstanding need for a simple yet comprehensive work for the education of the basic principles of Northern Sotho grammar on a first-year level". However, even in this work which professes to have first year students as a target user, the level of metalinguistic abstraction works against its efficiency – a common problem of pedagogic grammars, as pointed out by Bernardini (2004:17). To illustrate: in the very first lesson *Nouns in Northern Sotho* in Van Wyk et al (1992), students are confronted by the following sentence: "The nouns in Northern Sotho consist of a class prefix and a nominal root". To complicate matters further, a footnote has been added to the term 'nominal root', informing students that nominal roots are also known as nominal stems, thus touching on a still debated, highly theoretical issue, which is of little relevance to our specific target user.

They furthermore observe the traditional distinction between grammar and lexis, where grammar (form) is regarded as being of prime importance, and lexis (meaning) as a coincidental extra. Emphasis is therefore on the description of structures, without reference to the lexical choices that realize the elements of structure. This leads to a typical slot-and-filler approach, where the different slots within a structure are presented in a formulaic manner; the underlying assumption being that any lexical item that meets with the minimum criteria that determine its use can be slotted into the text, resulting in a grammatical sentence.

A further shortcoming in these grammars from our pedagogical perspective is that emphasis is placed on the presentation of complete structural paradigms, regardless of the fact that some structures in a given paradigm are purely theoretical possibilities that are in many cases not linked to actual language usage. The preoccupation with the provision of full paradigms seems to be based on the perception that in order to enable language learners to actively produce certain structures and / or utterances, they need to be familiar with the complete paradigm. The presentation by Poulos and

Louwrens (1994:285) of the monoverbal patterns for the aspect prefixes *-fô-*, *-nô-* and *-diô-*, which is reproduced below, is a case in point.

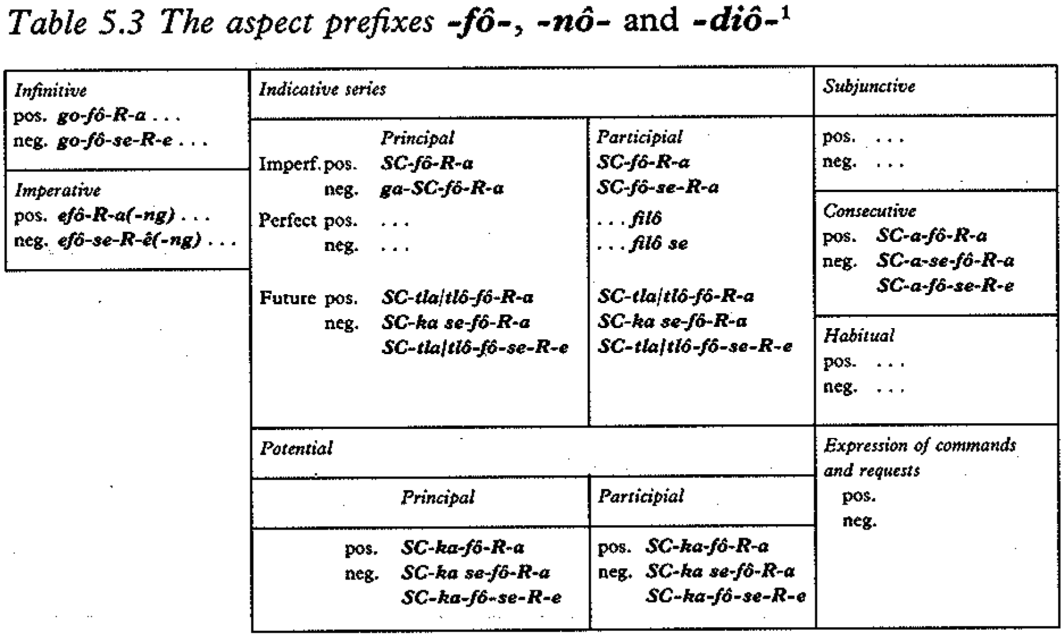


Table 5.3 The aspect prefixes ***-fô-*, *-nô-*** and ***-diô-***[1]

Infinitive pos. ***go-fô-R-a*** . . . neg. ***go-fô-se-R-e*** . . .	*Indicative series*		*Subjunctive*
Imperative pos. ***efô-R-a(-ng)*** . . . neg. ***efô-se-R-ê(-ng)*** . . .	*Principal* Imperf. pos. ***SC-fô-R-a*** neg. ***ga-SC-fô-R-a*** Perfect pos. . . . neg. . . . Future pos. ***SC-tla/tlô-fô-R-a*** neg. ***SC-ka se-fô-R-a*** ***SC-tla/tlô-fô-se-R-e***	*Participial* ***SC-fô-R-a*** ***SC-fô-se-R-a*** . . . ***filô*** . . . ***filô se*** ***SC-tla/tlô-fô-R-a*** ***SC-ka se-fô-R-a*** ***SC-tla/tlô-fô-se-R-e***	pos. . . . neg. . . . *Consecutive* pos. ***SC-a-fô-R-a*** neg. ***SC-a-se-fô-R-a*** ***SC-a-fô-se-R-e*** *Habitual* pos. . . . neg. . . .
	Potential		*Expression of commands and requests* pos. neg.
	Principal	*Participial*	
	pos. ***SC-ka-fô-R-a*** neg. ***SC-ka se-fô-R-a*** ***SC-ka-fô-se-R-e***	pos. ***SC-ka-fô-R-a*** neg. ***SC-ka se-fô-R-a*** ***SC-ka-fô-se-R-e***	

Figure 1: The aspectual prefixes of Northern Sotho as taken from Poulos and Louwrens (1994)

In this figure, a total of 25 grammatical structures in which the aspectual prefixes could potentially appear is given. A search through the tagged version of the PSC reveals that out of these 25 theoretically possible grammatical structures, only 13 actually occur. No instances were for example found of the aspectual prefix *-fô-* appearing in the imperative, i.e. *efô-R-a(ng)* (e.g. *efô-bolêl-a(ng)* 'just speak, all of you', 'all of you, speak freely') or its negative counterpart *efô-se-R-ê(-ng)* (e.g. *efô-se-bolêl-ê(-ng)* 'do not speak freely, all of you'). Obviously, the fact that some structures do not appear in the corpus, does not completely rule out the possibility of their existence, but it does indicate that their occurrence is so low as to be almost negligible. Presenting first time language learners with the full paradigm therefore serves little purpose.

Related to the issue of the emphasis on the provision of full paradigms, is the rule-based nature of the descriptions found in these grammars. This is especially evident in the explanations which are given of linguistic phenomena in which morpho-phonological changes play a role. The notion underlying such a rule-based approach is that knowledge of the rules determining any particular phenomenon enables the learner to generate an unlimited number of expressions or utterances by applying the rules to any given example. To illustrate: the learner should not only be able to express the notion of *ipona* 'see oneself' (< i- + bona), but also *ithata* 'like oneself' (< i- + rata), *itoma* 'bite oneself' (< i- + loma) and *itshwara* 'behave oneself' (<i- + swara). When introduced to reflexive verb forms, learners are therefore invariably confronted with lists of rules on sound changes which need to be learnt by heart in order to be able to actively produce the reflexive form of any given verb. Poulos and Louwrens (1994:186 et seq.) and Van Wyk et al (1992:114) list 15 phonological changes related to reflexive verb forms. Perusal of the frequency list drawn up of the PSC, reveals that amongst the 1 000 most frequent items, only 5 reflexive verb stems are to be found. This makes the inclusion of a full learning unit on reflexive verbs and the phonological rules determining the form of these verbs in our planned module somewhat questionable.

With respect to the issue of vocabulary, the language used to illustrate and exemplify grammatical structures in these grammars falls rather short in terms of the representation of natural language usage. Tognini-Bonelli (2001:40) emphasises the need to constantly expose first time learners to high frequency items. Recski (2006) states that "focusing on words which have a high frequency of occurrence and by concentrating on the usual rather than the exceptional, teachers can help learners acquire the language more efficiently, especially at elementary and intermediate level". It is therefore important that illustrative material should reflect the reality of frequency of use. In order to ascertain whether there is any correlation between the vocabulary used in example sentences in Northern Sotho grammars and lexical items which have a high frequency of use in the PSC, a simple frequency experiment was carried out. All example sentences used in Poulos and Louwrens (1994:206-213) and Van Wyk et al (1992:21-24) in their discussion of the imperfect tense of verbs in the indicative mood were manually analyzed to isolate all words with lexical meaning, i.e. excluding all function words. These words were then compared to a frequency list of the tagged version of the PSC to ascertain which percentage of these items appear amongst the top 500 most frequent words of the PSC. The results of this comparison are presented in Table 1.

Table 1: Top frequency lexical items

Title	**# of lexical words**	**% of lexical words appearing amongst top 500 of the *PSC***
A linguistic analysis of Northern Sotho (Poulos and Louwrens, 1994)	70	43
Northern Sotho for First years (Van Wyk et al 1992)	52	41

These grammars therefore miss out on an opportunity to expose students to high frequency vocabulary items and thus contribute to incidental learning of communicatively functional vocabulary. It once again illustrates an approach where grammar and the teaching thereof are deemed to be the highest priority, with lexis being to a large extent incidental.

It needs to be emphasized once more that the shortcomings identified in the foregoing discussion do not detract from the value of these grammars as reference works, but only serve to illustrate their inadequacy as pedagogical material for our specific target learner. Being aware of these limitations, lecturers often produce their own learning materials in the form of workbooks and class exercises, but even these self-produced materials do not succeed in avoiding the pitfalls of non-corpus based grammars. A KeyWord analysis of a special purpose corpus consisting of a set of class exercises, tests and examination papers compiled for students currently enrolled in the beginners course for Northern Sotho reveals that only 32% of words that are key, i.e. which occur with an uncommonly high frequency in the special purpose corpus appear amongst the top 500 items of the frequency list. In a sense, this result is even more significant than the simple frequency count carried out on the two grammars, since the materials which make up the special corpus were compiled specifically as teaching and learning material for a course aimed at beginner learners. Had these materials been corpus-based, the result would probably have been very different.

It should be clear from the discussion above that existing pedagogical materials are insufficient for the specific purposes of designing and teaching a module for first time learners of Northern Sotho.

Two case studies illustrating the use of a corpus are presented below. The first deals with the design and development of a tool for vocabulary acquisition; the second with selection and sequencing of content on both macro and micro levels, using the word categories ‘demonstrative’, ‘auxiliary verb’ and ‘adjective’ as illustrative examples.

Case study 1 Design and development of a vocabulary learning tool

Traditionally, teaching of vocabulary was viewed as being secondary to the teaching of grammar. Under the influence of Michael Lewis (1993), the idea that vocabulary should be at the centre of language teaching has had an influence in language teaching pedagogy. The key principle of Lewis’ lexical approach is that language is grammaticalized lexis and not lexicalized grammar, therefore one of the central organizing principles in the teaching of language should be lexis.[1] Moudraia (2001:1) states as follows in this regard: “Lexis is misunderstood in language teaching because of the assumption that grammar is the basis of language and that mastery of the grammatical system is a prerequisite for effective communication”. We do not propose that vocabulary teaching should replace grammar teaching, but argue that both should be present in language teaching. A further premise of Lewis’ approach is that language acquisition also includes the acquisition of meaningful ‘chunks’ of language, which include idioms, collocations and other types of formulaic expressions. As will be illustrated below, a corpus provides the language teacher with a veritable goldmine of these multiword lexical chunks, which can be incorporated in any tool used to teach vocabulary.

Language learners may acquire new vocabulary in two ways. The first is through direct instruction and study, a method which should be complemented by the second strategy of incidental acquisition, i.e. through the conscious or unconscious use of context clues during independent reading and listening activities. The second option presupposes that the learner should have sufficient reading and listening skills in the target language in order to enable him/her to use context clues to infer meaning. For first time learners, the initial exposure to the vocabulary of the target language is usually to a large extent through direct instruction, followed and complemented by incidental acquisition, as their proficiency in the target language increases.

Direct exposure to new vocabulary can be facilitated by providing students with a vocabulary list, based on frequency, of Northern Sotho vocabulary items containing at minimum level, the English / Afrikaans equivalents; indirectly, such a list can be used as a core vocabulary during all teaching and learning opportunities in which students are involved. A vocabulary list may seem a very basic tool for language learning, but it needs to be pointed out that no such a tool is available for learners of Northern Sotho, and with the necessary planning and maximal use of corpus data, such a list can be expanded to become a sophisticated, multifunctional tool rather than a mere list, containing much more information than equivalents in Afrikaans or English for randomly selected Northern Sotho words. The basic assumption underlying the usefulness of such a tool is that it should be a user-friendly tool that does not require a high level of referencing skills from the user. It should first of all provide the target user with basic information regarding meaning, supplemented with whatever other information is deemed appropriate for the language learner. The practical applications of these two strategies are discussed in the following paragraphs.

Although the extraction of a frequency list from a corpus is a trivial exercise, it only serves to provide the facilitator with raw data, which need to be intelligently interpreted in order to integrate frequency information with other kinds of linguistically and pedagogically relevant information, resulting in a multipurpose learning tool. Therefore, a certain degree of manual intervention and principled decision making is needed from the course designer / facilitator.

First, a raw frequency list extracted from any corpus lists all tokens included in the corpus, thus including *inter alia* numbers, Roman numerals, personal names and non-Northern Sotho words. The compiler of the vocabulary tool would therefore need to do a manual clean-up of the frequency list thrown up by the corpus search, eliminating as far as possible all 'noise' present in the list.

Secondly, the scope of the vocabulary tool needs to be determined. Here again, the decision should be based on the data provided by the corpus. A cursory look at the frequency list drawn from the PSC, reveals that of the top 100 raw frequencies, 80 are function words, i.e. agreement morphemes, pronouns, aspectual morphemes, particles, etc. Considering that the aim is to compile a tool of which the primary function is to learn vocabulary, function words should not be included in such a list. After deleting all function words from the top 500 frequency list 181 nouns, 112 verbs, 14 conjunctions, 13 adverbs, two interjections and one enumerative remain. Based on these figures and the stated outcomes of the module, a rough target of 332 vocabulary items seems reasonable, especially if it is taken into consideration that frequency wise, these top 500 items cover 70% of the total items in the corpus.

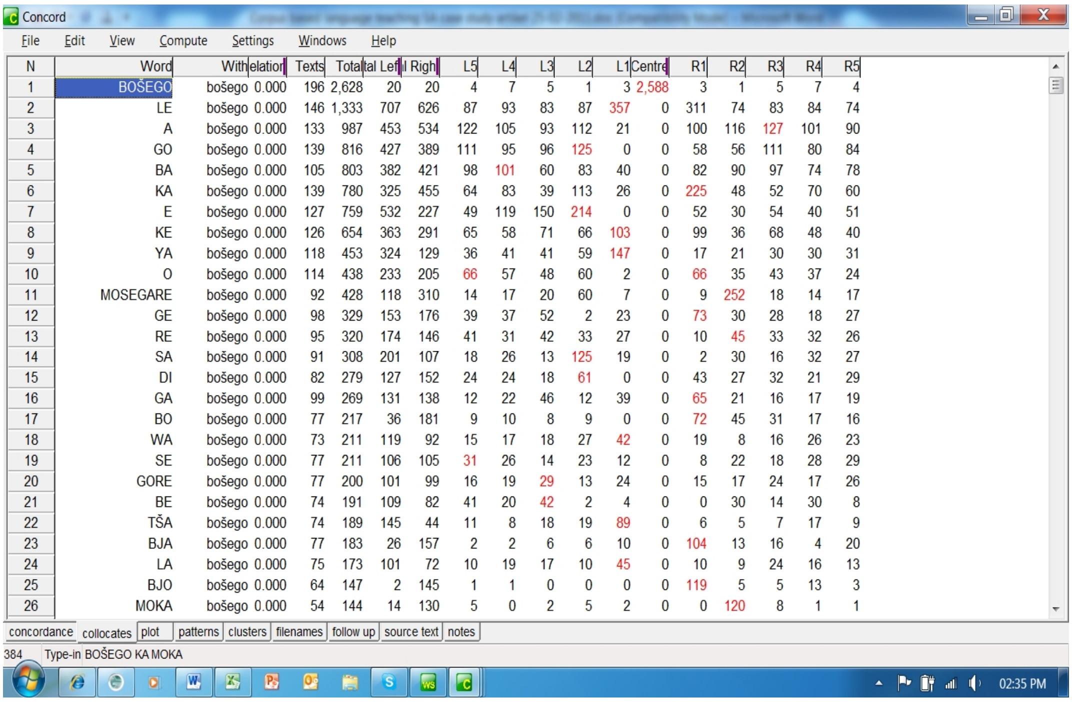
A further decision that has to be made concerns the structure in which the vocabulary tool is to be presented. Should the frequency information be transformed to produce an alphabetized, formally lemmatized list, similar to the one found in most dictionaries, or should frequency alone dictate the structure and contents of the tool? In case of the latter, the actual tokens as they appear in the frequency list will be listed. Should the strictly frequency-based approach be opted for, it would imply, for example, that frequent plural forms of nouns would be listed, but not necessarily the corresponding singular forms; for verbs, a derived form that is frequent may be listed, but not the basic form. It would also imply that the different word categories will appear in a single list. In this particular case, it was decided that a hybrid approach would best serve our purpose. Thus, the lexical items are listed according to frequency, not in alphabetical order, but it was decided to present the different parts of speech in separate lists, since this is more in keeping with the rather more formal approach being used in the presentation of the learning material on the grammar of Northern Sotho.

The availability of a corpus also gives access to two other aspects which contribute to functional vocabulary learning, i.e. real life usage examples, and collocational information. Example sentences need to be selected from the concordance lines provided by the corpus search while bearing the language proficiency of the learners in mind. Collocational information includes idiomatic expressions as well as other collocatory expressions, i.e. words which occur with a higher than normal frequency in each other's company. These lexical chunks can be isolated by making use of the collocation and cluster functions of the Concordance tool of WST. It needs to be decided how this information can be incorporated into a vocabulary learning tool. Due to space constraints, illustrative material is provided for the categories 'noun' and 'verb' only. A prototypical example of the section on nouns appears as Figure 2 in Addendum A.

In the design of the vocabulary tool, the following are taken as guidelines for the section containing nouns:

- Both singular and plural forms are listed, provided that both forms appear in the top 500 frequencies. Whenever a corresponding plural form does not appear amongst the top frequencies, it is not listed, although the class number of the plural form is provided. Compare the item *bošego* 'night' in Figure 2. The same principle is followed with regard to frequent plural forms with infrequent singular forms. Compare *mahlo* 'eyes' in this regard, which has a frequency ranking of 181, occurring 4 457 times in the corpus. Its singular counterpart *leihlo* 'eye' ranks 774 on frequency and appears 722 times in the PSC.
- When both singular and plural forms are listed, only the meaning of the singular form is given.
- If for a singular form, no plural exists (and vice versa), this is indicated by a dash in the appropriate cell.
- In cases where the plural form is significantly more frequent than the singular form, the example sentence reflects the use of the more frequent plural form. Compare the item *ngwana* (freq. 6410) / *bana* (freq. 8708) 'child/children' in this regard.
- Locativized nouns are listed as primary items, and not as derived forms of their underived counterparts. Compare *lapeng* 'in the family, in the homestead'.

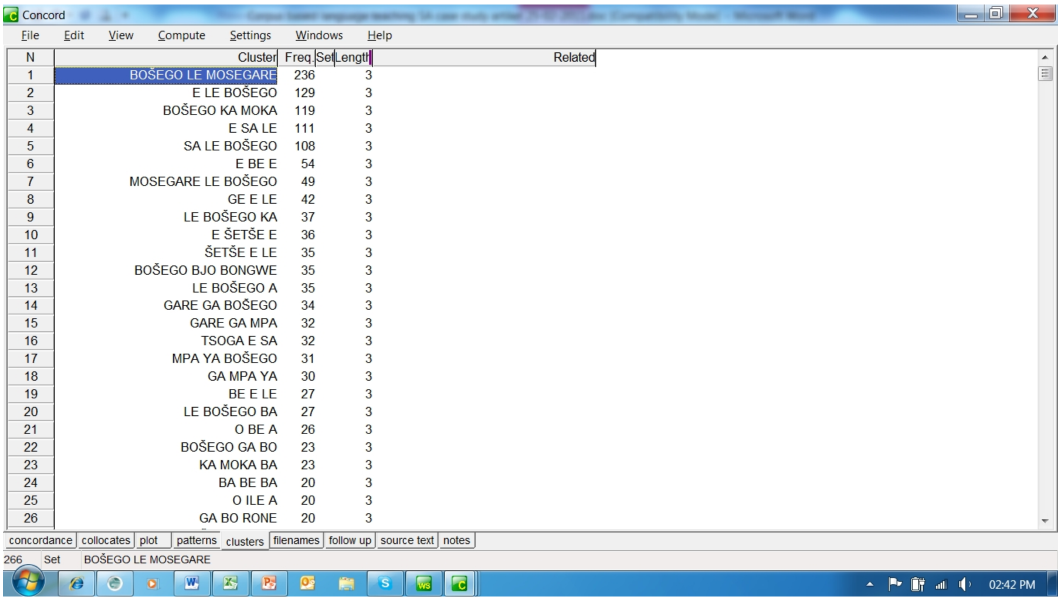
- Where appropriate, usage notes can be provided.
- Collocational information is provided separately under the heading *Company kept*. By including collocational information, learners are exposed to meaningful chunks, the idea being that the acquisition of these chunks takes place simultaneously with that of single lexical items. The first 20 – 30 lines of collocates provided in the Concordance tool are perused for lexical items which frequently appear in the vicinity (5 positions to the left and right) of the search word, supported by the cluster function, which reveals clusters of two, three or more words in which the search word frequently appears. If no significantly frequent collocational information can be found, no information is provided in this column. Compare the screen shots below for the collocational and cluster information for the noun *bošego* 'night' by way of illustration:



N	Word	With	elation	Texts	Total	tal Lef	l Righ	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	BOŠEGO	bošego	0.000	196	2,628	20	20	4	7	5	1	3	2,588	3	1	5	7	4
2	LE	bošego	0.000	146	1,333	707	626	87	93	83	87	357	0	311	74	83	84	74
3	A	bošego	0.000	133	987	453	534	122	105	93	112	21	0	100	116	127	101	90
4	GO	bošego	0.000	139	816	427	389	111	95	96	125	0	0	58	56	111	80	84
5	BA	bošego	0.000	105	803	382	421	98	101	60	83	40	0	82	90	97	74	78
6	KA	bošego	0.000	139	780	325	455	64	83	39	113	26	0	225	48	52	70	60
7	E	bošego	0.000	127	759	532	227	49	119	150	214	0	0	52	30	54	40	51
8	KE	bošego	0.000	126	654	363	291	65	58	71	66	103	0	99	36	68	48	40
9	YA	bošego	0.000	118	453	324	129	36	41	41	59	147	0	17	21	30	30	31
10	O	bošego	0.000	114	438	233	205	66	57	48	60	2	0	66	35	43	37	24
11	MOSEGARE	bošego	0.000	92	428	118	310	14	17	20	60	7	0	9	252	18	14	17
12	GE	bošego	0.000	98	329	153	176	39	37	52	2	23	0	73	30	28	18	27
13	RE	bošego	0.000	95	320	174	146	41	31	42	33	27	0	10	45	33	32	26
14	SA	bošego	0.000	91	308	201	107	18	26	13	125	19	0	2	30	16	32	27
15	DI	bošego	0.000	82	279	127	152	24	24	18	61	0	0	43	27	32	21	29
16	GA	bošego	0.000	99	269	131	138	12	22	46	12	39	0	65	21	16	17	19
17	BO	bošego	0.000	77	217	36	181	9	10	8	9	0	0	72	45	31	17	16
18	WA	bošego	0.000	73	211	119	92	15	17	18	27	42	0	19	8	16	26	23
19	SE	bošego	0.000	77	211	106	105	31	26	14	23	12	0	8	22	18	28	29
20	GORE	bošego	0.000	77	200	101	99	16	19	29	13	24	0	15	17	24	17	26
21	BE	bošego	0.000	74	191	109	82	41	20	42	2	4	0	0	30	14	30	8
22	TŠA	bošego	0.000	74	189	145	44	11	8	18	19	89	0	6	5	7	17	9
23	BJA	bošego	0.000	77	183	26	157	2	2	6	6	10	0	104	13	16	4	20
24	LA	bošego	0.000	75	173	101	72	10	19	17	10	45	0	10	9	24	16	13
25	BJO	bošego	0.000	64	147	2	145	1	1	0	0	0	0	119	5	5	13	3
26	MOKA	bošego	0.000	54	144	14	130	5	0	2	5	2	0	0	120	8	1	1

384 Type-in BOŠEGO KA MOKA

Figure 3: Collocates for *bošego* 'night'



N	Cluster	Freq.	Set	Length	Related
1	BOŠEGO LE MOSEGARE	236		3	
2	E LE BOŠEGO	129		3	
3	BOŠEGO KA MOKA	119		3	
4	E SA LE	111		3	
5	SA LE BOŠEGO	108		3	
6	E BE E	54		3	
7	MOSEGARE LE BOŠEGO	49		3	
8	GE E LE	42		3	
9	LE BOŠEGO KA	37		3	
10	E ŠETŠE E	36		3	
11	ŠETŠE E LE	35		3	
12	BOŠEGO BJO BONGWE	35		3	
13	LE BOŠEGO A	35		3	
14	GARE GA BOŠEGO	34		3	
15	GARE GA MPA	32		3	
16	TSOGA E SA	32		3	
17	MPA YA BOŠEGO	31		3	
18	GA MPA YA	30		3	
19	BE E LE	27		3	
20	LE BOŠEGO BA	27		3	
21	O BE A	26		3	
22	BOŠEGO GA BO	23		3	
23	KA MOKA BA	23		3	
24	BA BE BA	20		3	
25	O ILE A	20		3	
26	GA BO RONE	20		3	

Figure 4: 3 Word clusters of *bošego* ‘night’

Note that the most frequent 3 word cluster in which the item *bošego* ‘night’ appears, is utilized as a usage example in the tool.

Figure 5 in Addendum A represents an excerpt from the verb list. For verbs, the guidelines are as follows:

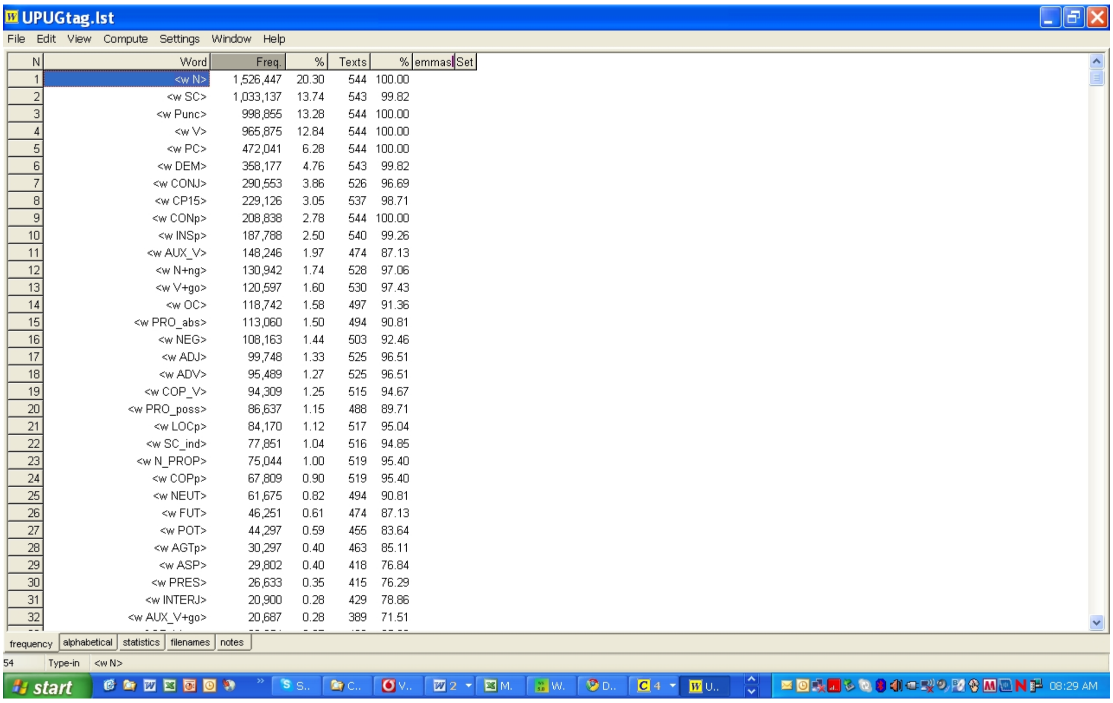
- Verb stems are added to the tool as they appear in the frequency list. Derived forms are therefore entered as such. This includes verbs stems displaying the relative suffix –*go*.
- Verb stems ending in –*e* are split into two entries, since the verbal ending can either be the result of the verb appearing in the subjunctive mood, or it can be because it is preceded by a negative morpheme. This has an influence on the meaning of the verb stem.
- Collocational information is treated as with nouns.

The current version of the vocabulary learning tool is paper-based, which limits its application to some extent and which makes it a rather static tool. Future work includes the design of an e-version, which can be made available online to students via the students’ online learning system. Through a well-designed interface, students would be able to access additional grammatical information through drop-down menus and mouse-overs. Extensive usage notes can be included. The tool can thus evolve into a mini-grammar with lexis as the core element.

Case study 2: selection and sequencing of linguistic content

Until recently, the only available Northern Sotho corpus had been an untagged one, i.e. it contained no information on any linguistic features, e.g. part of speech, noun class information, tense forms, etc. Searches within such an electronic database are therefore mainly limited to the lexical level. Having access to a corpus that is annotated for parts of speech allows much more sophisticated querying of the corpus and does not restrict the usefulness of the corpus to lexical level. Querying on Part of Speech (POS) Tags can reveal the frequency with which certain word categories occur in general language usage, guiding the language teacher in the decision as to the inclusion / exclusion of a particular word class in the curriculum. It needs to be pointed out however, that tagging of the PSC is currently done on the orthographical level. This implies that both full linguistic words e.g. nouns (N), pronouns (PRO) and adverbs (ADV), and morphemes e.g. subject concords (SC), object concords (OC), the future tense morpheme (FUT), etc. are tagged and can thus be retrieved automatically from the corpus. Thus, only in cases where there is a one to one correspondence between a word category and an orthographic unit, will POS frequency show up during a corpus query.

In the selection and sequencing of learning content, frequency information provided by the corpus needs to be carefully interpreted, and frequency considerations need to be weighed up against other factors, such as communicative functionality and the level of grammatical competency that is required for the production and reception of any particular structure. To illustrate: on drawing up a frequency list of the parts of speech in the tagged version of the PSC, demonstratives (DEM) appear in the 6th position on the frequency ranking of the 54 part of speech tags, covering almost 5% of the total corpus, cf. the screenshot below.



UPUGtag.lst

N	Word	Freq.	%	Texts	%	emmas	Set
1	<w N>	1,526,447	20.30	544	100.00		
2	<w SC>	1,033,137	13.74	543	99.82		
3	<w Punc>	998,855	13.28	544	100.00		
4	<w V>	965,875	12.84	544	100.00		
5	<w PC>	472,041	6.28	544	100.00		
6	<w DEM>	358,177	4.76	543	99.82		
7	<w CONJ>	290,553	3.86	526	96.69		
8	<w CP15>	229,126	3.05	537	98.71		
9	<w CONp>	208,838	2.78	544	100.00		
10	<w INSp>	187,788	2.50	540	99.26		
11	<w AUX_V>	148,246	1.97	474	87.13		
12	<w N+ng>	130,942	1.74	528	97.06		
13	<w V+go>	120,597	1.60	530	97.43		
14	<w OC>	118,742	1.58	497	91.36		
15	<w PRO_abs>	113,060	1.50	494	90.81		
16	<w NEG>	108,163	1.44	503	92.46		
17	<w ADJ>	99,748	1.33	525	96.51		
18	<w ADV>	95,489	1.27	525	96.51		
19	<w COP_V>	94,309	1.25	515	94.67		
20	<w PRO_poss>	86,637	1.15	488	89.71		
21	<w LOCp>	84,170	1.12	517	95.04		
22	<w SC_ind>	77,851	1.04	516	94.85		
23	<w N_PROP>	75,044	1.00	519	95.40		
24	<w COPp>	67,809	0.90	519	95.40		
25	<w NEUT>	61,675	0.82	494	90.81		
26	<w FUT>	46,251	0.61	474	87.13		
27	<w POT>	44,297	0.59	455	83.64		
28	<w AGTp>	30,297	0.40	463	85.11		
29	<w ASP>	29,802	0.40	418	76.84		
30	<w PRES>	26,633	0.35	415	76.29		
31	<w INTERJ>	20,900	0.28	429	78.86		
32	<w AUX_V+go>	20,687	0.28	389	71.51		

Figure 6: Frequency ranking for Part of Speech categories

Demonstratives have a high communicative value, their function being a deictic one, indicating the actual position a referent occupies vis-à-vis the speaker and the addressee in a discourse situation. Its secondary function is that of anaphoric pronoun, used to 'refer back' to the antecedent. Demonstratives therefore express the notions of 'this / these (one(s)), that / those (one(s))'. Apart from their communicative functionality, demonstratives form part of quite a number of grammatical constructions, e.g. the adjective, verbal and nominal relative constructions. The frequency with which demonstratives appear in naturally occurring language and the fact that it forms part of larger constructions therefore provide good motivation not only for inclusion of these parts of speech in a curriculum aimed at beginner learners, but also for introducing it at a rather early stage, since knowledge of demonstratives is a prerequisite for mastering other high frequency structures.

Looking at auxiliary verbs (AUX_V), these appear in the 11th position on the frequency ranking of the 54 part of speech tags, covering approximately 2% of the corpus. This could suggest that auxiliary verbs should be considered for inclusion in the curriculum. However, the use of auxiliary verbs requires rather advanced grammatical knowledge, especially of the modal system of Northern Sotho, since the auxiliary determines the grammatical mood in which the complementary verb appears. A corpus-based approach offers a number of possible solutions. In the first instance, a frequency query can be done to ascertain which auxiliary verb stems have the highest frequency of use. A corpus query reveals that the stems -*be*, -*ile* and -*swanetše* account for 36%, 16% and 7% respectively of all incidences of auxiliary verbs, together accounting for almost two thirds of the total usage of auxiliaries. Furthermore, as was pointed out above, apart from frequency considerations, the communicative functionality of any structure needs to be taken into account when a decision as to the inclusion of such a structure is taken. When looking at the function of auxiliaries in Northern Sotho, it is clear that their semantic function is to modify the meaning of the complementary verb with which it co-occurs, thus giving rise to fine semantic nuances being expressed by the verbal element. For a module such as the one under consideration, the production of verbal elements expressing these finely tuned meanings does not seem necessary in order to obtain a basic communicative competence, although nothing prevents the inclusion of these and other frequently used auxiliary verb stems, e.g. -*šetše* 'already', -*fela* 'usually, continuously, regularly' and -*napa* 'then, and then' for reception purposes only.

Lastly, when compared to auxiliaries, Adjectives (ADJ) rank 17th on frequency. In contrast to auxiliaries, adjectives have a high communicative value, with the adjectival stems having a clear lexical meaning. It needs to be pointed out that adjectives in Northern Sotho belong to a closed class, consisting of a limited number (approximately 45) of stems, cf. Taljard (2006) in this regard. Northern Sotho adjectives are used to express notions such as basic colours (black, blue/green, red, white), sizes and shapes (big, small, tall, short, thin, thick, left, right) as well as the numerals 'two' to 'five', making them good candidates for inclusion in the curriculum. On the downside however, adjectives appear in rather complicated grammatical structures, with the form of the adjectival stem, e.g. *–raro* 'three' depending on the noun class of the nominal antecedent, cf. *batho ba* <u>*bararo*</u> 'three people', but *mantšu a* <u>*mararo*</u> 'three words'. When the antecedent belongs to classes 8, 9 and 10, phonological changes occur, cf. *diranta tše* <u>*tharo*</u> 'three rand'. Furthermore, adjectives agree by means of a demonstrative with the nominal antecedent, thus any given adjective can potentially appear in as many as 13 different structures, leading to a total of close to 600 potential adjectival constructions. However, since adjectives are a closed class, frequency queries for each adjectival stem (by means of wildcard searches, if necessary) can reveal which adjectival stems have the highest frequency of use. The notion that there are significant differences in the frequency with which adjective stems appear in real language usage, has never before been mentioned, neither has it been taken into consideration when decisions on the content of learning material are taken. In all existing Northern Sotho textbooks, authors have attempted to give a complete as possible inventory of all adjective stems, which is a sensible approach when writing a grammar. However, from a language learner's perspective differences in frequency of use is a relevant consideration. To illustrate exactly how significant these differences are, the frequency with which ten randomly selected adjective stems appear in the PSC is represented in Table 2:

Table 2: Frequency of selected adjective stems

Adjective stem	**Frequency**	**Adjective stem**	**Frequency**
-*be* 'bad, ugly'	937	-*nyane* 'small, little, few'	728
-*golo* 'big, old'	6811	-*šweu* 'white'	588
-*koto* 'thick(set), deep'	182	-*telele* 'tall, long'	1595
-*ngwe* 'certain, one, some, other'	32082	-*tona* 'right(hand side), male, masculine'	336
-*ntši* 'many, much'	4731	-*sehla* 'yellowish, light(-skinned)	132

Based on frequency considerations, the language teacher can decide on a cut-off point for inclusion / exclusion of certain adjective stems. However, frequency information should once again not override all other considerations. Should it for example turn out that the adjective containing the stem *–ne* 'four' has a frequency which is below the cut-off point, whereas the stems *-bedi* / *-pedi* 'two', *-raro* / *-tharo* 'three' and *–hlano* / *-tlhano* 'five' do qualify for inclusion, common sense should prevail, and 'four' should also be included. Furthermore, collocational information from the corpus reveals which antecedents most often occur with every adjective stem in each noun class, and these can be used as illustrative material when the grammatical structure of the adjective construction is explained, cf. Figure 7 in Addendum A. Once again, such example material contains real usage examples, and does not rely on the intuition of the language teacher.

As far as could be ascertained, querying a POS-tagged corpus for guidance as to which grammatical structures should be included in a language curriculum has not been done for any other language. Corpus-querying seems to be limited to querying on the lexical level, thus the approach described above is a rather novel one and its possibilities should also be investigated for other languages, especially resource rich languages such as English and German which have large POS-tagged corpora at their disposal.

Conclusion

It comes as no surprise that when compared to languages such as English the pedagogy of language teaching for the African languages, specifically for Northern Sotho, lags far behind. It has been shown that pedagogical material used in the teaching of Northern Sotho as an additional language is inadequate in that it is based largely on the structural model of grammatical description, with little attention being paid to aspects such as frequency of use, real language usage and the communicative value of grammatical structures. Selection and sequencing of learning material is largely based on anecdotal evidence or on the intuition of the language teacher, which is often shown to be rather wide off the mark when compared to corpus data. Utilization of corpus data provides the language teacher with guidance on both macro and micro level as regards the content of the curriculum. Furthermore, it stands to reason that whatever information is retrieved from the Northern Sotho corpus, would to a large extent also be applicable to other related languages, especially Tswana and Southern Sotho, and possibly also for the Nguni languages.

In its current format, the prototypical tools described in this article are of a rather rudimentary nature, but the potential of these tools when converted into e-format needs to be further explored and exploited.

Future work regarding the role of corpora in language teaching should furthermore include a detailed analysis of the needs and expectations of language learners, similar to the study done by Römer (2009), in which she investigates the needs and wants of language teachers teaching English as a foreign language. It is often found that first time learners of Northern Sotho have unrealistic expectations as to the level of proficiency that they will reach after having completed one semester course. Once their needs and expectations have been identified, the role of a corpus in satisfying these needs can be determined.

NOTES:

1. Lewis (1993) uses the term 'lexis' to refer to not only single words, but also word combinations that are stored in mental lexicons.

Bibliography

Bernardini S. 2004. *Corpora in the classroom An overview and some reflections on future developments*, in Sinclair, J (ed) (2004b).

Biber D & Conrad S. (s.d.) *Corpus linguistics and Grammar Teaching.* http://www.longmanhomeusa.com/content/pl_biber_conrad_monograph_lo_3.pdf. Last accessed: 13 March 2012.

Bowker L & Pearson J. 2002. *Working with specialized language: a practical guide to using corpora.* London: Routledge.

Conrad S. 2000. Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century? *TESOL Quarterly* 34(3), 548-560.

De Schryver G-M & De Pauw G. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of *TshwaneLex*. *Lexikos* 17, 226-246.

Gabrielatos C. 2005. Corpora and Language Teaching: Just a fling or wedding bells? *Teaching English as a Second Language E-Journal (TESL-EJ)* 8(4).

Lewis M. 1993. *The lexical approach: The state of ELT and the way forward.* England: Language Teaching Publications.

Lombard DP, Van Wyk EB, Mokgokong PC. 1985. *Introduction to the Grammar of Northern Sotho.* Pretoria: J L Van Schaik Ltd.

Louwrens LJ. 1994. *Aspects of Northern Sotho Grammar*. Pretoria: Via Afrika Ltd.

Mauranen A. 2004. *Spoken corpus for an ordinary learner*, in Sinclair, J (ed) (2004b).

Moudraia O. 2001. Lexical approach to second language teaching. *Centre for Applied Linguistics*. Online resources: Digests.
http://www.cal.org/resources/digest/digest_pdfs/0102-moudraia-lexical.pdf.
Last accessed: 12 February 2012.

Poulos G and Louwrens LJ. 1994. *A linguistic analysis of Northern Sotho.* Pretoria: Via Afrika Ltd.

Pravec NA. 2002. Survey of learner corpora. *ICAME Journal 26,* 84 -114.

Recski LJ. 2006. Corpus linguistics at the service of English teachers. *Literatura y lingüística 17*, 303 – 324.

Römer, Ute. 2009. Corpus research and practice: What help do teachers need and what can we offer? In: Aijmer, Karin (ed.). *Corpora and Language Teaching*. Studies in Corpus Linguistics 33. Amsterdam: John Benjamins. 83-98

Römer, Ute. 2010. Using general and specialized corpora in English language teaching: Past, present and future. In: Campoy-Cubillo, Marí Carmen, Begoña Belles-Fortuño & Lluisa Gea-Valor (eds.). *Corpus-based Approaches to English Language Teaching*. London: Continuum. 18-35

Sinclair J. 2004(a). Developing Linguistic Corpora: a Guide to Good Practice Corpus and Text — Basic Principles. *AHDS literature, languages and linguistics.* http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm
Last accessed: 30 November 2011.

Sinclair J. (Ed) 2004(b). *How to use corpora in language teaching*. Studies in Corpus Linguistics. Amsterdam, John Benjamins Publishing Co

Sinclair J. 2004(b). New evidence, new priorities, new attitudes, in Sinclair, J (ed) (2004b).

Taljard E. 2006. Corpus-based linguistic investigation for the South African Bantu languages: a Northern Sotho case study. *South African Journal of African Languages* 26(4).

Tognini-Bonelli E. 2001. *Corpus Linguistics at work*. Amsterdam: John Benjamins Publishing Co.

Van Wyk EB, Groenwald PS, Prinsloo DJ, Kock JHM and Taljard E. 1992. *Northern Sotho for First-years*. Pretoria: J L Van Schaik Ltd.

Ziervogel D. 1969. *Handboek van Noord-Sotho*. Pretoria: J L van Schaik Ltd.

Addendum A

Freq. rank	Singular	Plural	Meaning; (also called sense)	Example	Company kept
56	*mo*tho^{1}	*ba*tho^{2}	person; human being; someone; anyone	*Ke **motho** wa go bolela nnete.* ‘(S)he is a person who speaks the truth.’	***motho*** *wa batho* ‘poor soul’ ***motho*** *yo mongwe* ‘certain person’
73	pele^{N-}	-	first; before; in front of; ahead	*Lapa le agwa* **pele** *ga ngwako.* ‘The verandah is built in front of the house.’ **NOTE:** *pele* is often used as an adverb’	*tšwela* ***pele*** ‘ continue’ ***pele*** *ga gagwe* ‘in front of him/her’
76	ka *mo*ka^{3}	-	all; every	*Batho* **ka moka** *ba ile mošomong.* ‘All the people have gone to work.’	
102	kgoši^{9}	*di*kgoši^{10}	chief; king	*Farao e be e le* **kgoši** *ya Egepeta.* ‘Pharao was king of Egypt.’	
103	*mo*nna^{1}	*ba*nna^{2}	man; male person; husband	*O dula le* **monna** *wa gagwe.* ‘She lives with her husband.’	***monna*** *le mosadi* ‘man and woman, husband and wife’
105	*mo*rena^{1} Morena	*ba*rena^{2} -	Sir; mister Lord	***Morena*** *Matšato ke hlogo ya sekolo.* ‘Mr Matšato is the principal of the school.’	
107	eng^{9}	-	what	*O dira **eng**?* ‘What are you doing?’ **NOTE**: *Eng* is used as a question word.	*lebaka la **eng**?* ‘why?’
109	*ngw*ana^{1}	*ba*na^{2}	child	*Bao ke* **bana ba** *sekolo.* ‘Those are school children.’	
	^{18}				

111	*mo*rago	-	last; after; behind	*Ba fihlile ka* ***morago*** *ga iri. 'They arrived after an hour.'* **NOTE:** *morago* is often used as an adverb	*boela* ***morago*** 'return'
181	^{5}	^{6} *ma*hlo	eyes	*Re bona ka* ***mahlo*** *a rena*. 'We see with (our) eyes.'	
280	^{14} *bo*šego	^{6}	night	*Dipese di feta moo* ***bošego*** *le mosegare.* 'Busses pass here night and day.'	***bošego*** *ka moka* 'the whole night' ***bošego*** *bjo bongwe* 'one night' *gare ga* ***bošego*** 'in the middle of the night' *mosegare le* ***bošego*** 'day and night'
319	^{5} lapeng	^{6}	in; near; at; to the family / courtyard	*Mosadi o boetše* ***lapeng*** *la gagwe.* 'The woman returned to her family.'	

Figure 2: Excerpt from noun list

Freq. rank	Verb stem	Meaning	Example	Company kept
22	*re*	1 say 2 mean	1 *Ba **re** Mampuru ke ngwana wa Malekutu* 'They say Mampuru is the child of Malekutu' 2 *Lentšu leo le **re** eng?* 'What does that word mean?'	
49	*bona*	see	*Ke **bona** gore ba fihlile* 'I see that they have arrived'	
70	*ya*	go	*Ke tlo **ya** sekolong gosasa* 'I'll go to school tomorrow'	
87	*bolela*	speak	*Ke rata go **bolela** le wena* 'I want to speak to you'	***bolela** ka ga* 'talk about'
92	*tseba*	know	*Pheladi o **tseba** diphiri tša bona* 'Pheladi knows their secrets'	*go **tseba** mang?* 'who knows?'
118	*tloga*	depart, leave	*Ba tlo **tloga** lehono* 'They will leave today'	*go **tloga** mathomong* 'from the start' *go **tloga** fao* 'from then / there onwards'
132	*tsebe*	1 must know 2 not know	1 *O **tsebe** gore papa ga a raloke* 'You must know that dad doesn't play' 2 *Ga a **tsebe** polelo ya Setebele* 'She does not know the Ndebele language'	
210	*tlile*	came	*Ba **tlile** ka dikoloyana* 'They came in small cars'	
244	*latelago*	following, who / which follow(s), next	*Bala mafoko a a **latelago*** 'Read the following sentences'	*dipotšišo tše di **latelago*** 'following questions'
270	*diriša*	use	***Diriša** polelo ya maleba* 'Use appropriate language'	*ka go **diriša*** 'by using'
276	*ithuta*	study, learn	*Baithuti ba **ithuta** go bala* 'Learners learn to read'	

Figure 5: Excerpt from verb list

Class # & prefix	*-golo*	*-ngwe*	*-ntši*	*-telele*
1 *mo-*	*motho yo mogolo* 'big, old person'	*motho yo mongwe* 'certain, another person'		*monna yo motelele* ‘tall man’
2 *ba-*	*batho ba bagolo* 'big, old people'	*batho ba bangwe* 'some, other people'	*batho ba bantši* 'many people'	*batho ba batelele* ‘tall people’
3 *mo-*	*mošomo wo mogolo* ‘big job’	*mokgwa wo mongwe* ‘certain manner, another way’		*mosela wo motelele* ‘long tail’
4 *me-*	*mehuta ye megolo* ‘big kinds of’	*mehlala ye mengwe* ‘some, other examples’	*mekgwa ye mentši* 'many ways'	*menwana ye metelele* ‘long fingers’
5 *le-*	*lethabo le legolo* ‘great joy’	*letšatši le lengwe* ‘one day, other day’		*lebaka le letelele* 'long time'
6 *ma-*	*maatla a magolo* ‘big power’	*mantšu a mangwe* ‘some, other words’	*meetse a mantši* 'much, a lot of water'	*maeto a matelele* 'long journeys'
7 *se-*	*setšhaba se segolo* ‘big nation’	*selo se sengwe* 'certain, another thing'		*sebaka se setelele* 'long time'
8/10 *di(N)*	*ditlhaka tše kgolo* 'big (capital) letters'	*dilo tše dingwe* 'some, certain, other people'	*dilo tše (di)ntši* 'many things'	*dipotšišo tše (di)telele* 'long questions'
9 *N,* -ø	*tlhokomelo ye kgolo* 'great care'	*nako ye nngwe* 'sometime, at a certain time'		*nako ye telele* 'long time'
14 *bo-*	*bothata bjo bogolo* ‘big problem’	*bošego bjo bongwe* ‘one night, certain night’	*boya bjo bontši* ‘lots of hair, fur’	*bjang bjo botelele* ‘long, tall grass’

Figure 7: Adjectives