

---

# Semi-automatic Term Extraction for the African Languages, with Special Reference to Northern Sotho \*

Elsabé Taljard, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (e.taljard@freemail.absa.co.za)*, and Gilles-Maurice de Schryver, *Research Assistant of the Fund for Scientific Research — Flanders (Belgium) and Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (gillesmaurice.deschryver@rug.ac.be)*

---

**Abstract:** Worldwide, semi-automatically extracting terms from corpora is becoming the norm for the compilation of terminology lists, term banks or dictionaries for special purposes. If African-language terminologists are willing to take their rightful place in the new millennium, they must not only take cognisance of this trend but also be ready to implement the new technology. In this article it is advocated that the best way to do the latter two at this stage, is to opt for computationally straightforward alternatives (i.e. use 'raw corpora') and to make use of widely available software tools (e.g. WordSmith Tools). The main aim is therefore to discover whether or not the semi-automatic extraction of terminology from untagged and unmarked running text by means of basic corpus query software is feasible for the African languages. In order to answer this question a full-blown case study revolving around Northern Sotho linguistic texts is discussed in great detail. The computational results are compared throughout with the outcome of a manual excerption, and *vice versa*. Attention is given to the concepts 'recall' and 'precision'; different approaches are suggested for the treatment of single-word terms *versus* multi-word terms; and the various findings are summarised in a Linguistics Terminology lexicon presented as an Appendix.

**Keywords:** TERMINOLOGY, TERMINOGRAPHY, MANUAL EXCERPTION, READING AND MARKING, SEMI-AUTOMATIC TERM EXTRACTION, RETRIEVAL, AFRICAN LANGUAGES, NORTHERN SOTHO (SEPEDİ), RAW CORPORA, PRETORIA SEPEDİ CORPUS (PSC), WORDSMITH TOOLS, WEIRDNESS RATIO, KEY WORD, LOG-LIKELIHOOD, RECALL, PRECISION, MOTHER TERM, SINGLE-WORD TERM, MULTI-WORD TERM, STEM, ROOT, KEY-WORD-IN-CONTEXT (KWIC), COLLOCATION, COLLOCATE, LEXICAL GAP, CLUSTER, LINGUISTICS TERMINOLOGY LEXICON

**Senaganwa:** Go ntšhwa ga mareo ka tirišo ya seripa sa semotšhene malebana le maleme a Afrika, šedi ye kgolo e lego Sesotho sa Leboa (Sepedi). Go ntšhwa ga mareo ka tirišo ya seripa sa semotšhene go tšwa ka gare ga dikhophase go thomile go ba

---

\* An earlier version of this article was presented as a paper at the Seventh International Conference of the African Association for Lexicography, organised by the Dictionary Unit for South African English, Rhodes University, Grahamstown, 8–10 July 2002.

setlwaedi go hlangweng ga mananeo a mareo, dipanka tša mareo goba dipukuntšu mererong yeo e itšego lefaseng ka bophara. Ge e le gore boramareo ba maleme a Afrika ba ikemišeditše go tšea madulo a bona mo mileneamong wo mofsa, ga ba swanela go hlokomela fela tsela ye, eupša ba swanetše gape ke go ikemišetša go diriša theknolotši ye mphsa. Mo taodišwaneng ye go hlalošwa gore mo nakong ye, tsela ye kaone ya go dira dilo tše pedi tše go boletšwego ka tšona ke go kgetha ditlhamolo tša thwii tšeo di dirišago khomphutha (se se ra gore tšhomišo ya khophase) le go šomiša ditlabakelo tša *software* (bj.k. *WordSmith Tools*) tšeo di lego gona gohle. Ka fao maikemišetšo a magolo ke go humana ge e ka ba go ntšhwa ga mareo ka seripa sa semotšhene go tšwa ka gare ga khophase yeo e se nago ditlaleletšo tšeo di tseneletšego ka mašakaneng, tša go hlhla, go ka dirišwa malemeng a Afrika goba aowa. Gore re kgone go araba potšišo ye, go hlalošitšwe ka tsinkelo mohlala wa taba ya go nyakišišwa yeo e amanego le diteng tša thutapolelo tša Sesotho sa Leboa. Dipelo tšeo di humanwego ka go diriša khomphutha di bapetšwa ka gohle le dipelo tšeo di humanwego ge go dirišwa kgetho ya mantšu ka matsogo. Šedi e fiwa dikgopolo tša kgakologelo (*recall*) le nepagalo (*precision*); mekgwa yeo e fapafapanego e a akanywa gore e kgone go hlatholla mareo a lentšu le tee ge a bapetšwa le mareo a mantšu a mantši; gomme dikhumano tšeo di fapanego di akaretšwa ka gare ga pukuntšu ya Mareo a Thutapolelo yeo e tšweletšwago bjalo ka Mamatletšo.

**Mantšu a bohlokwa:** MAREO, MONGWALO WA MAREO, KGETHO YA MANTŠU KA MATSOGO, GO BALA LE GO SWAYA, GO NTŠHWA GA MAREO KA SERIPA SA SEMOTŠHENE, GO HWETŠA GAPE, MALEME A AFRIKA, SESOTHO SA LEOBOA (SEPEDI), DIŠEGONTŠU (DIKHOPHASE), KHOPHASE YA SESOTHO SA LEOBOA YA TSHWANE (KST), WORDSMITH TOOLS, WEIRDNESS RATIO, LENTŠU LA BOHLOKWA, LOG-LIKELIHOOD, KGAKOLOGELO, NEPAGALO, LEREO LA MOTHEO, LEREO LA LENTŠU LE TEE, LEREO LA MANTŠU A MANTŠI, KUTU, MODU, LENTŠU LA BOHLOKWA KA GARE GA KAMANO (LBGK), PEAKANYO, BEAKANYA, TLHOKEGO YA LEREO, SEHLOPHA, PUKUNTŠU YA MAREO A THUTAPOLELO

## 1. Semi-automatic Term Extraction — A Brief Theoretical Conspectus

On the international front, the use of electronic corpora for general lexicographical purposes has for the past two decades become a firmly entrenched procedure. According to Ahmad and Rogers (2001: 729) "it is common practice these days in many different types of dictionary to use the systematic evidence of corpora rather than the more *ad hoc* selection of citations by readers more traditionally used in lexicography". De Schryver and Prinsloo (2000: 292) state that "[t]he intensified systematic exploitation of electronic corpora for lexicographic purposes has unmistakably revolutionised dictionary making" and point for example out that during the compilation of all the recent British English learners' dictionaries "electronic corpora were used very actively in order to produce reference works of a standard hitherto simply unimaginable".

The use of electronic corpora for terminological purposes has however been accepted much more slowly on both the theoretical and practical levels. The onomasiological approach in terminology as opposed to the semasiological

bias of general lexicography is offered by Ahmad and Rogers (2001: 729) as one possible reason for the lack of corpora utilisation in terminology management. Terminology as a scientific discipline is concept-driven, the basic objectives of terminological work being: (a) delimiting or identifying the concepts of a subject field or domain, (b) naming these concepts by means of terms, and (c) fixing the referential scope of each term by means of a definition. Thus, on a strictly theoretical level, the concept-oriented nature of terminology logically excludes the possible use of text corpora, since a concept is an abstract entity, not to be found in textual material. If, however, it is accepted that terminology also includes a terminographical dimension, which has the compilation and/or publication of a terminology list, a term bank or a dictionary for special purposes as its final objective, the use of corpora and the subsequent computational management of terminology becomes a relevant issue. In this regard, Ahmad and Rogers (2001: 730) indicate that "there are clear signs that corpus-based terminology management, including the identification of terms and translation-oriented terminology, as well as the whole concept of terminology management, is now being discussed". Sager (1990: 130) is even more emphatic in stating that systematic term compilation is firmly corpus-based, which implies that terms are no longer manually excerpted from previous lists or by individual searches, but from a corpus of material.

The process whereby computer software is used to automatically detect and extract potential terms from electronic corpora, is known as *(semi-)automatic term extraction*. In the great majority of the current approaches, characteristics of a special-language corpus are compared to those of a general-language corpus. In *all* approaches, humans remain the final arbiters, and must decide whether or not the terms suggested by the software do indeed have term status. Broadly speaking, the approaches themselves are either purely statistical, purely linguistic, or they are hybrid, i.e. they combine features of the two extremes. Moreover, different methods are often used to extract single-word terms as compared to the extraction of multi-word terms.

## **2.      Electronic Special-field Corpora for the African Languages and Query Software**

When it comes to the use of electronic corpora for the computational management of terminology for the African languages spoken in South Africa,<sup>1</sup> no headway has hitherto been made, and this for obvious reasons. A corpus is based on available written texts in a specific language. As such, this does not pose a problem for these languages, since they all have a written tradition. Granted, some of these languages have only been reduced to writing during the last decade (e.g. Ndebele), but other languages such as Zulu or Northern Sotho have a relatively long literary tradition. The compilation of an electronic corpus for general lexicographical purposes is therefore not problematic, pro-

vided of course that the necessary technological support is available.<sup>2</sup>

However, if an electronic database is to be compiled for terminological purposes, it presupposes the availability of text material revolving around specific fields. Due to the historically disadvantaged situation of the African languages, even today virtually no subject-specific texts which could be used to build an electronic database are available. As a result of the pre-1994 political and educational system, the vast majority of subject-specific material is written in either English or Afrikaans, with textbooks on literature and grammar of the African languages a possible exception. The African-language terminologist therefore has very little, if any, access to special-field texts which can be used to compile an electronic special-field corpus. This does not only have implications for the compilation of corpora, but also determines the methodology which has hitherto been used by African-language terminologists. Due to the lack of special-field texts, terminologists compiling terminology lists would make use of texts written in English and/or Afrikaans to select the relevant terms. After having isolated the terms which characterise a specific subject field, these terms would then be translated into the African languages, resulting in a multilingual terminology list. Alberts (2000: 236-237) refers in this regard to the compilation of technical dictionaries for the African languages by terminologists from the National Language Service (NLS) of the Department of Arts, Culture, Science and Technology (DACST) on a variety of special-field subjects.

Nonetheless, the dependency of African-language speakers and terminologists on textbooks and other subject-specific sources produced in English and/or Afrikaans will hopefully decrease in the times to come, seeing that the African languages are starting to take their rightful place in the South African educational landscape. As a matter of fact, it is the authors' belief that special-language texts will soon be produced on a *large* scale in the African languages. The evolution on the Internet is a good case in point. Indeed, more and more texts of a technical nature already make their appearance in African languages online. It is this realisation that prompted the current research. Since the trend worldwide is increasingly towards the semi-automatic extraction of terminology from corpora, African-language terminologists must not only be aware of this development, but they must also be fully prepared. The main aim of this article is thus to research the *feasibility* of semi-automatic term extraction for the African languages.

We purposely opted for computationally straightforward alternatives and insisted on using widely available software tools. The rationale behind this is simply that we wish to reach out to as many colleagues as possible. On the corpus level, this implies that we chose to work with *raw corpora*, i.e. just plain running text without any tags or mark-up whatsoever. As far as the software is concerned, we selected WordSmith Tools (WST), "an integrated suite of programs for looking at how words behave in texts" (Scott 1999: WST help).<sup>3</sup> WST is inexpensive, easy to acquire, user-friendly, and already in use at several National Lexicography Units (see e.g. De Schryver and Lepota 2001: 3).

### 3.     **Case Study: Northern Sotho Linguistics Terminology**

In order to evaluate the success rate and usefulness of the suggested approach (i.e. the analysis of raw corpora with WST, with the aim of semi-automatically extracting terminology), the computational result has to be compared to the outcome of the current method of manual term excerption (i.e. the physical reading of texts and marking of relevant terms). A list of manually excerpted terms will therefore serve as a terminological benchmark against which the success rate of the computational extraction of terms will be measured. Conversely, since manual term excerption is of necessity subject to human error, the results of the computational processing will also be compared to the results of the manual excerption in order to ascertain whether the semi-automatic processing might have succeeded in identifying terms which were overlooked during the process of manual excerption.

For the purpose of this investigation, a number of texts on Northern Sotho linguistics were taken as the textual material from which terms are to be retrieved. These texts were kindly provided in electronic format by Prof. L.J. Louwrens and are an integral part of the study material used at the University of South Africa (UNISA). After conversion to a text-only format, a simple count with WST's WordList tool revealed that this special-field corpus contains 74,251 tokens (running words) and 4,744 types (unique words).

### 4.     **Reading and Marking — A Case Study**

Term excerption as a conscious activity forming part of terminology management is influenced by a number of aspects, e.g. the target users and their specific needs, the exact purpose for which the special-field corpus is to be created, the literacy levels of the discourse participants, etc. These aspects do, however, not form the focus of the current investigation. The linguistic texts from which terms are to be retrieved, form part of the study material written for pre-graduate students who are mother-tongue speakers of Northern Sotho. For the sake of the argument, it can be assumed that these students would also be the target users of a basic terminology list containing all the relevant linguistic terms appearing in the texts, and their English equivalents.

The initial phase of the investigation consisted of a manual excerption of terms from the linguistic texts. Manual excerption implies close scrutiny of a text in order to identify terms which are relevant to a specific subject field; in this case, linguistics. This manual reading and marking was performed by a professional terminologist, and the terms were entered into a preliminary term list. Locativised nouns (i.e. nouns displaying the locative suffix **-ng**), as well as relative verbs (i.e. verbs containing the relative suffixes **-go/-ng**), were excluded from this list. This decision was based on the fact that the meaning of these derivations can regularly be inferred from the base forms entered in the list. Furthermore, terms were listed as they appeared in the text, so if a term for

example appeared in its plural form, it was listed as such.

The process of manual term excerption resulted in a term list containing 350 'raw' (i.e. 'unlemmatised') terms. Of these terms, 309 are single-word terms, the rest being made up of 41 multi-word terms. Lemmatisation of the initial list produced a term list containing 285 terms, lemmatised under their singular form in the case of nouns, followed by an indication of gender affiliation. Verbs with verbal extensions were lemmatised as such, and not according to the stem of the verb. Each term was then provided with a translation equivalent in English. The result of this endeavour can be found in the Appendix, where all the articles preceded by ☐ and ☑ constitute the 285 manually excerpted terms after lemmatisation.

## 5. Semi-automatic Term Extraction — A Case Study

As was pointed out in par. 2, our aim is to investigate how well a simple yet powerful and versatile program such as WST fares in semi-automatic term extraction when this software is fed with *untagged* and *unmarked* corpora. This procedure is therefore by definition language-independent and purely statistical. Consequently, the *unlemmatised* list of 350 manually excerpted terms must be used as a benchmark, as also WST processes raw data. As is generally done in the field of semi-automatic term extraction, we will first look into various ways to computationally extract single-word terms, and only then into the extraction of multi-word terms, for which the benchmarks are 309 and 41 terms respectively.

### 5.1. Semi-automatic Extraction of Single-word Terms

Ironically, the only publicised attempt to automatically extract African-language terminology from corpora, is the report by Sewangi (2000, 2001) for Swahili. He (Sewangi 2000: 67-68) states:

[T]he identification of single-word terms in a text corpus is difficult because there are no structural criteria that can be used to separate term-words from non-term-words in the text. [...] the identification of single-word terms should involve subject specialists and language experts. [...] This should be done manually on the basis of the knowledge of the subject-domain and of the language.

Although Sewangi could make free use of the computational tools that have been developed for over a decade by Hurskainen (1992, 1995, 1996, 1999, Hurskainen and Halme 2001), and although he thus had access to corpora with full descriptions of Swahili morphological patterns and constraints, he nonetheless effectively marks single-word terms *manually*, which, in view of the highly-technical corpus annotations at his disposal, is disappointing.

Up to this day Swahili remains the only African language for which an

efficient morphological analyser has been built. Several human-language technology projects are under way however — in Pretoria for Zulu and Northern Sotho, in Harare for Shona and Zimbabwean Ndebele — and it is expected that in less than a decade there will be a handful of African-language morphological analysers. As this is not yet the case, investigating the possibility of analysing *raw* corpora at this stage is defensible.

### 5.1.1. Top Ranks in a Frequency List

An undemanding operation in the computational processing of a corpus is a simple frequency count, in which each type in the corpus can be listed according to its rank, or in order of its frequency. Obviously, since a frequency count lists all types, i.e. unique (orthographic) words, which appear in the text according to their frequency of occurrence, such a procedure will produce a lot of 'noise', in that most of the items appearing amongst, say, the 100 most frequent words, are not terms at all, but other lexical items such as concords, conjunctions, etc. and general vocabulary not related to the subject field.

The linguistic texts were selected in WST and on comparing the frequency list created by WordList with the list of manually excerpted terms, it was found that only 20 items appearing in the top 100, appeared on the list of manually excerpted terms and could thus be regarded as terms. Scanning the frequency list down to rank 500 revealed only an average of 18 terms per 100 items (viz. for the first 100, 20; the second 100, 20; the third 100, 18; the fourth 100, 14; and the fifth 100, 17).

### 5.1.2. Stop-list Constrained Top Ranks in a Frequency List

Much of the 'noise' produced by a simple frequency count can be reduced substantially by filtering out all items which are generally known as function words and closed-class words. In WST, this filtering can be done automatically by making use of stop lists (also known as 'exclude lists'). For Northern Sotho, the following can be regarded as function words, i.e. lexical items with little or no lexical content: all agreement morphemes, demonstratives, particles, conjunctions, copulative verb stems and auxiliary verb stems. The second group of lexical items that can also be included in a stop list are those items that belong to so-called closed classes, i.e. classes which contain a very limited number of items. The following closed classes can be identified for Northern Sotho: adverbs, interrogative words, adjectives (class prefix + adjective stem, which number only a handful in Northern Sotho), locative nouns, pronouns and ordinal numbers.

Even if, in addition, locativised nouns, relative verbs and nouns with diminutive suffixes are also disregarded, the top 100 items on the 'cleansed' frequency list still contain only 39 items (as compared to 20 in par. 5.1.1) that also appear on the manually excerpted term list. Even with a stop list, it is clear that a simple frequency list tends to over-generate, i.e. to identify items which

are not terms relevant to the specific subject field. Also, even though stop lists have the advantage that they can be made once and then be used for numerous extractions, they, conversely, hold the risk that interesting data is nonetheless cut away.

### 5.1.3. Weirdness Ratios, Key Words and Recall and Precision

#### *Weirdness ratios*

Reading through *top-frequency words*, no matter whether these were automatically extracted with or without stop lists from running text, is obviously an unrefined procedure. Ahmad and Rogers (2001: 744-745) suggest that the comparison of the frequency distribution of items in special-language texts to that of items in general-language texts might be a next option worth investigating. When the relative frequency of each item in special-language texts is divided by the relative frequency of the same item in general-language texts, the 'weirdness ratio' of those items, and thus their potential as term candidates, can be measured.

#### *Key words — How they can be extracted semi-automatically*

A much more sophisticated and statistically sounder approach to weirdness ratios is offered by the KeyWord tool of WST. Since 1997 Mike Scott, the creator of WST, has extensively studied — and widely published on — the computational treatment of key words and related concepts (e.g. Scott 1997, 1997a, 2000, 2000a, 2001). He rightly notes that the term 'key word' has remained undefined in linguistics, despite being in common use by non-linguists, and despite the fact that the notion itself features strongly in fields such as Content Analysis, Information Retrieval, and Corpus Linguistics at large (Scott 2000: 51).

In his work on text schemata and stereotypes, Scott (1997: 236) defines the term 'key word' as 'a *word which occurs with unusual frequency in a given text*'. Unusual frequency can be related to outstandingness and implies that a word has an unusually high (or unusually low) frequency in a text (or sub-corpus), in comparison to its occurrence in a *reference corpus* of some kind. In this specific study, Scott's aim was to make certain culturally significant inferences about schemata, i.e. the socially determined networks of links between ideas, by identifying key words which appear across feature stories taken from the *Guardian* newspaper.

The basic procedure for identifying key words in one or more texts is to compare the frequency of every distinct word-type in those texts with the frequency of the same word-type in a reference corpus, the reference corpus being the bigger of the two corpora. As a first step, a wordlist is compiled of the reference corpus by making use of the WordList tool. The wordlist contains all the different types in the reference corpus and lists them according to their frequency of occurrence. Secondly, a similar, but much smaller, wordlist is drawn



up for the specific text(s) for which the key words are to be identified. The third step, which is done by means of the KeyWord tool, consists of comparing the frequency of each item in the smaller of the two wordlists with the frequency of the same item in the reference wordlist. Items which display a great disparity in frequency are identified as key words, since the disparity would imply that that specific item occurs with unusual frequency in the smaller corpus. Note that KeyWord throws up items with outstanding frequencies and not 'top frequencies'. Thus, even if a given item appears with, for example, an extremely high frequency of 5% in the smaller corpus, and if that item would have a similar percentage in the reference corpus, such an item will not turn out to be 'key', even though it might perhaps be the 'most frequent' item. Every word which appears in the smaller corpus is taken into account when 'keyness' is calculated, except if it has been excluded by entering it into a stop list.<sup>4</sup>

For the actual calculation of keyness, WST offers two statistical tests, viz. the classic  $\chi^2$  (chi-square) test of significance with Yates' correction for a 2 x 2 table, and Dunning's (1993) log-likelihood test "which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against [a] reference corpus" (Scott 1999: WST help). We opted for the log-likelihood test, a choice generally made by corpus linguists today (compare Scott 1997a: 238). Two more parameters are important when calculating key words, and must be set: (a) the minimal frequency, and (b) the level of outstandingness. The first parameter specifies the minimum frequency with which a potential key word must occur in the text(s) from which the key words are to be extracted. The value was set at 3, yet values of 2 and 1 resulted in near-identical findings. The second parameter, also known as p value, establishes a minimum probability. The standard  $p \leq 0.000001$  was used, meaning that each key word's appearance has a danger of only 1 in a million of not being statistically significant. Reformulated, in our study an item is said to be a 'key word' if: (a) it occurs in the text(s) at least 3 times, and (b) its frequency in the text(s) when compared with its frequency in a reference corpus is such that the statistical probability as computed by the log-likelihood procedure is smaller than or equal to one in a million.

***Key words — An illustration of how they can be extracted semi-automatically***

It is clear that the identification of key words is a purely mechanical process, based on a comparison of patterns of frequency. To illustrate the procedure, a Northern Sotho text on the new South African coat of arms, taken from the Internet,<sup>5</sup> was randomly selected to serve as an example of a specific text for which key words are to be identified. This text, henceforth CoA, contains 1,038 tokens and 356 types. As a reference corpus, a selection of the Pretoria Sepedi Corpus (PSC), consisting of 5,175,686 tokens and 136,567 types, was used. As far as the reference corpus is concerned, Scott (1997: 244, endnote 9) observes that "as long as the reference corpus is fairly sizeable" — and he suggests at least a million running words — "results are quite similar even if the reference

corpus is altered" (compare also Scott 2000a: 115). This observation was found to be true.<sup>6</sup> Scott (2001: 126, endnote 2) further suggests that the text(s) from which key words are to be extracted can even form a tiny sub-set of the reference corpus. This is also what was done, i.e. CoA is part of PSC. Wordlists were drawn up for CoA and PSC with the WordList tool, and these lists were fed into the KeyWord tool. All the key words suggested by KeyWord are shown in (1).

- (1) Semi-automatic key-word extraction from a text on the new South African coat of arms

N	Key word	Translation	CoA Count	CoA %	PSC Count	PSC %	Keyness
1	sefoka	coat of arms	10	0.96	231		87.3
2	ditirelo	services	8	0.77	114		77.3
3	bontšha	show	11	1.06	640	0.01	76.1
4	seswa <sup>7</sup>	(something) new cl. 7	6	0.58	18		75.2
5	Afrika	Africa(n)	9	0.87	941	0.02	51.9
6	tlhame	secretary-bird	4	0.39	19		46.9
7	barulaganyi	designers	3	0.29	3		42.8
8	Borwa	South	8	0.77	1,137	0.02	41.4
9	badiriši	users	3	0.29	9		37.6
10	setšhaba	nation	10	0.96	3,204	0.06	36.2
11	lebišitšwe	is / are aimed at	3	0.29	14		35.3
12	batho	people	16	1.54	12,232	0.24	33.1
13	se	subj. conc. cl. 7; dem. cl. 7; ...	39	3.76	73,986	1.43	27.6
14	mmušo	government	6	0.58	1,214	0.02	26.9
15	manaka	tusks	3	0.29	72		25.9
16	leswa	(something) new cl. 5	3	0.29	73		25.9
17	emela	represent(s)	4	0.39	308		25.5
18	tshedimošo	information	3	0.29	85		25.0
19	mabapi	with regard to, regarding	5	0.48	830	0.02	24.4
20	<i>a</i>	<i>subj. conc. cl. 6; poss. conc. cl. 6; ...</i>	26	2.50	301,005	5.82	26.1

Columns 2 and 3 in (1) list the key words the KeyWord tool extracted entirely automatically. As stressed at the outset, human beings remain the final arbiter, which is why Column 3 was added here. Columns 4 and 5 show the occurrence (as a count and percentage respectively) of the suggested key words in CoA. The count and percentage of those same items in PSC, the reference corpus, is shown in Columns 6 and 7. The last column, Column 8, lists the keyness values.

From (1), it is clear that **sefoka** 'coat of arms' occurs 10 times in CoA, compared to an occurrence of 231 times in the bigger reference corpus, yet *proportionally* its frequency is many times higher in the smaller corpus than in the 5.2-million-word reference corpus. **Sefoka** is, as a result of this large disparity in frequency, the item with the highest keyness value. In (1) all suggested key words are 'positively key', except for the last item (in italics) **a** 'subject concord of class 6; possessive concord of class 6; ...' which is 'negatively key'. The latter simply means that it occurs less often than would be expected by chance in comparison with the reference corpus. By simply scrolling over the key words in (1) one can deduce that the text in question must provide information

regarding the new South African coat of arms. It also seems as if the designers aimed at showing symbols such as a secretary bird and tusks to represent the government's attempt to provide new services to the nation's people / users. Note that all the relevant<sup>8</sup> key words were used in the previous description, clearly suggesting that key words — which, lest we forget, are proffered *fully automatically* by the KeyWord tool — do indeed pinpoint the 'aboutness' (Scott 2000a: 107-109) of a text.

### *Key words — Semi-automatically extracting single-word terminology*

In his discussion of possible applications of the KeyWord tool, Scott (1997: 243) makes no mention of its potential value for terminological purposes. However, Ahmad and Rogers (2001: 744) claim that "[c]omputing the 'ratio' of word forms in special-language and general-language texts also allows a provisional distinction to be made between general-language open-class words on the one hand, and special-language open-class words on the other, i.e., term candidates". This is exactly what KeyWord does, albeit in a more sophisticated way. In terms of Scott's KeyWord procedure, this will imply that the frequency of items appearing in the linguistic texts (being special-language 'texts') be compared to their frequency in the reference corpus (being general-language 'texts'), in order to identify term candidates. Again, the result of such an investigation will be compared to the outcome of the manual term excerption in order to evaluate the efficacy of the computational procedure.

In carrying out the KeyWord procedure as described above, 654 key words are suggested by KeyWord. From item 586 onwards, though, one moves into negative keyness, where there are obviously no linguistic terms to be found. Reading through the 585 terms that are positively key, one quickly finds out that an amazing 189 of them also appear on the manual list. 61% of the manually excerpted single-word terms (189 out of 309) are thus thrown up *entirely automatically* by the software. The only required human intervention is to read through the suggested list and to decide on term status.

Moreover, in doing so, an *extra* 18 terms that had been missed during the manual excerption, are revealed. Besides the intrinsic advantage of a computational approach with which a large percentage of the terminology can be extracted automatically, the fact that a computational approach also reveals items that are missed during a manual pass, might even be of greater value. Actually, following the various semi-automatic approaches for the extraction of single-word terms (see also par. 5.1.4 below), a total of 33 new single-word terms were revealed. The benchmark that will therefore henceforth be used for comparisons will be  $309 + 33 = 342$  single-word terms.

### *Recall and Precision*

Two concepts now need to be introduced that are central to Information Retrieval, viz. 'recall' and 'precision'. We list two sets of definitions — first Streh-

low's (2001: 428-429) from a document-retrieval perspective, and then Ahmad and Rogers' (2001: 748) from a terminology-retrieval perspective — to show their wide application:

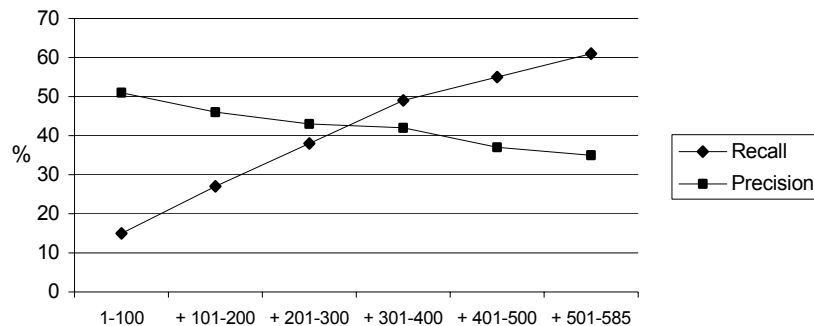
Retrieval effectiveness is usually described by the *precision* and *recall* that are associated with a retrieval operation. *Recall* is the fraction of relevant documents retrieved from a collection, and *precision* is the fraction of documents that are relevant in a retrieved set of documents. [...] In general, recall and precision are found to be inversely related, i.e., an increase of one results in a decrease of the other [...]

"Recall" is the proportion of relevant materials retrieved from a text collection given a set of terms. "Precision" is the proportion of retrieved materials that are relevant.

In other words, given a running text or sub-corpus in which there are a number of terms, *recall* is the percentage of terms actually retrieved as compared to the total number of terms in that text or sub-corpus. When all the retrieved items as a whole, i.e. both the retrieved terms and the 'noise' (i.e. retrieved non-terms) are considered, the actual percentage of terms in that body of retrieved items is called *precision*.

Both recall and precision will now be calculated for the semi-automatic single-word extraction achieved with the KeyWord tool above. In the top 100 KeyWord list, 51 so-called key words are effectively terms, of which 1 is new compared to the manual excerption, i.e. 50 + 1. Analogously, in the second 100, there are 36 + 4 terms; in the third 100, 36 + 3; in the fourth 100, 31 + 5; in the fifth 100, 19 + 2; and in the last stretch up to 585, 17 + 3. As one proceeds through the KeyWord list, recall and precision, expressed in %, are therefore as shown in (2).

(2) Recall and Precision for the semi-automatic extraction of single-word linguistic terms with KeyWord



With a benchmark of 342, the *recall* for the first 100 is 51 divided by 342 or 15%; when the next 100 are added, it is 51 + 40 divided by 342 or 27%; etc. As one proceeds through the KeyWord list, the recall thus *increases*. The *precision* for the first 100 is 51 out of 100 items, or thus 51%; when the next 100 are added it is 51 + 40 out of 200 items, or thus 46%; etc. As one proceeds through the Key-Word list, the precision thus *decreases*. From the graph in (2) one can see that the claimed inverse relationship between recall and precision holds rather well for our case study.

It should be clear from the discussion so far that the KeyWord procedure can be used successfully for the semi-automatic extraction of a large proportion (here over 60%) of the single-word terms in running text. When compared to the use of (stop-list constrained) top ranks in a frequency list, the KeyWord analysis represents a substantial increase as regards the success rate of computational processing. The question now remains, however, as to how the rest of the terms can be extracted in a semi-automatic way.

#### 5.1.4. Mother Terms and the Generation of Single-word (Compound) Terms — Stems / Roots

When one studies the first 25 unique *terms* on the KeyWord list, one sees that the core terminology of the special-field at hand has been identified. To use Ahmad and Rogers' (2001: 742) terminology, one could say that these terms are "the 'mother' terms of a given specialism or the signature terms of a specialist subject". Like mothers, they can generate other terms through *compounding*, or whatever term formation process is valid for the language at hand.

By making use of the Concord tool of WST, single-word (compound) terms built up around the stem or root of any mother term can quickly be identified. For example, if the stem of the key word **tlhalošo** 'meaning', i.e. **tlhaloš-**, is used as the search node in concordance lines, the terms shown in (3) are but a few of the terms generated by the mother term.

- (3) A sample of the single-word compound terms generated by the stem of the mother term **tlhalošo**

swana le theksi, bese, setimela, paesekele, mmotoro, bj. bj. **Tlhaloš-** **-okamanyi** e dira gore mantšu a a amanego ka tlhalošo a lebopi la mathomo mo lentšung, bj. bj. Re hlalošitše gore **tlhaloš-** **-okatološo** ga e kwešišege ge lentšu le eme le le nnoši ka ge lentšu, di bitšwa tlhalošokelello ya lentšu leo. Lemoga ge **tlhaloš-** **-okelello** e fapana go ya ka batho, ka ge e theilwe godimo ga

The first two terms in (3), namely **tlhalošokamanyi** 'associative meaning' and **tlhalošokatološo** 'extended meaning', were also marked manually, but the third term **tlhalošokelello** 'cognitive meaning' was only found computationally. By doing a search with the stem and/or root of each of the top 25 mother terms as search item, all (compound) terms containing those stems and/or roots can be identified. The Search function furthermore allows for automatic exclusion of certain items, thus further refining the search procedure. Although this procedure is less automatic than the KeyWord tool, it is still a rather swift

way to detect terms that were not included in the KeyWord list. Column 5 in (4) lists the number of those (new and extra) single-word terms each mother term generated.

- (4) Generation of *new* and *extra* (compound) terms from the first 25 unique mother terms on the KeyWord list (new = not listed by KeyWord; extra = missed during manual excerption and not listed by KeyWord)

N	Term	Translation	Stem / Root	New + Extra single-word terms	Extra multi-word terms
1	lediri	verb	-dir-	6 + 3 = 9	0 + 1 = 1
2	lefoko	sentence	-fok-	2 + 2 = 4	1 + 2 = 3
3	mantšu	(linguistic) words	-ntšu	3 + 1 = 4	3 + 3 = 6
4	tlhalošo	meaning	tlhaloš-	3 + 1 = 4	0 + 1 = 1
5	serewa	topic	-rew-	0	0
6	popego	morphology, form, structure	-popeg-	1 + 0 = 1	0
7	mmoledi	first person	-boled- / -moled-	0 + 1 = 1	0
8	kgokagano	discourse	kgokagan-	0	1 + 2 = 3
9	legoro	(noun) class	-goro	2 + 2 = 4	0
10	togaganyo	cohesion	togagan- / logagan-	0	0
11	legokedi	agreement morpheme	-kgoked-	0	0
12	tswalano	relationship, association	tswalan-	0	1 + 0 = 1
13	tiro	predicate, action, process	tir-	1 + 1 = 2	0 + 1 = 1
14	mmoledišwa	second person, addressee	bolediš- / -molediš-	0	0 + 3 = 3
15	leina	noun	-ina	3 + 1 = 4	0 + 1 = 1
16	lethuši	auxiliary verb	-thuš-	0	0
17	lešala	pronoun	-šal-	3 + 0 = 3	0
18	lebopi	morpheme	-bop-	1 + 0 = 1	1 + 1 = 2
19	lehlathi	adverb	-hlath-	1 + 0 = 1	0
20	kganetšo	negation, negative	-ganetš-	0	0 + 1 = 1
21	lereo	term	-reo	0	0
22	makopanyi	conjunctions	-kopany-	4 + 1 = 5	0
23	kamano	(inter)relationship	-aman-	0	0
24	matlema	prepositions	-tlem-	1 + 2 = 3	0
25	kgatelelo	emphasis	-kgatelel-	0	0
Σ				31 + 15 = 46	7 + 16 = 23

In all, the top 25 mother terms in (4) generated 31 *new* single-word terms (i.e. terms that were not listed by KeyWord) and 15 *extra* single-word terms (i.e. terms that were missed during the manual excerption, and were not picked up by KeyWord either). For example, the root **-tlem-** of the mother term **matlema** 'prepositions', generated 1 new term (**tlemagano** 'cohesion') and 2 extra terms (**tlemagantšha** 'link, connect' and **tlemaganya** 'link, connect'), as shown in (5).

- (5) New and extra single-word terms generated by the root of the mother term **matlema**

mo temaneng ka ge bobedi bja ona bo hlola kgokagano goba -tlem- -agano gare ga mantšu, dikafoko le mafoko. Ke go re  
 mo nomorong ye: (vii). Le tlemagantšha mantšu a fe? Le -tlem- -agantšha melamo le marumo. Go bonala ka eng ge e le  
 ke lefe? Ngwala nomoro ya maleba ... Le -tlem- -aganya mantšu a fe? ... le ...

As far as the single-word terms are concerned, the KeyWord tool followed by the Concord tool for just the top 25 unique KeyWord terms, throws up as many

as 189 + 18 and 31 + 15 single-word terms respectively, or thus 253 terms in all. With a benchmark of 342 items, this thus means that 74% of the single-word terms were extracted semi-automatically.

## 5.2. Semi-automatic Extraction of Multi-word Terms

Up to this point, the investigation has centred around single-word terms only. Terms often consist of multi-word units — in the list of manually excerpted terms, 41 were multi-word units — and any computational term-extraction process should also be able to isolate multi-word terms. For Northern Sotho, as for any other language, the computational identification of multi-word terms is many times more complex than the identification of single-word terms. Given the fact that we are working with raw corpora, the extraction will still be purely statistical, yet linguistics will have to come into play to make more informed decisions as to the term status of the computational suggestions. The need for this linguistic, and thus language-dependent, support will be apparent in the discussion below.

### 5.2.1. Top Ranks in a Frequency List

With the WordList tool it is possible to make multi-word wordlists, i.e. 2-word wordlists, 3-word wordlists, 4-word wordlists, etc. up to 8-word wordlists. From the manual excerption, one knows that at least three 2-word terms, thirty-five 3-word terms, two 4-word terms, and one 5-word term — totalling 41 multi-word terms in all — are to be extracted semi-automatically. Reading through the first few hundred items in each of those multi-word wordlists, quickly indicates that this process does not produce very significant results. It was thus decided to immediately move to the KeyWord process.

### 5.2.2. Key Words

Using the KeyWord tool on multi-word level is analogous to using it on single-word level. Multi-word wordlists for the linguistic texts were already compiled in the previous phase (par. 5.2.1), so only multi-word reference wordlists (based on the 5.2-million-word reference corpus) had to be compiled in addition. Once done, KeyWord was requested to calculate the 2-word key words, the 3-word key words, etc.

The two manually identified 2-word terms were listed in the 2-word KeyWord list, while no extra terms were found. The recall is thus 100% for the 2-word terms. 18 of the 35 manually excerpted 3-word terms were listed in the 3-word KeyWord list (recall = 51%), together with 5 extra 3-word terms. The two 4-word terms were listed in the 4-word KeyWord list (recall = 100%), together with 1 extra 4-word term. Finally, the sole manually identified 5-word term was not listed in the 5-word KeyWord list (recall = 0%), yet 2 extra 5-word terms were identified instead. Precision values are extremely low in all cases.

Taken together, 22 multi-word terms (out of 41, recall = 54%) were thrown up semi-automatically, and a surprisingly high number of 8 extra multi-word terms were discovered. Together with the extras that were additionally extracted with the methods described below (in the last section of par. 5.2.3 and in par. 5.2.4), the benchmark for the multi-word terms rises to 72. This simply means that over 40% (31 out of 72) of the multi-word terms were missed during the manual pass — confirming not only the value of a computational approach, but also pointing at the difficulty of a manual excerption of especially multi-word terms from running text.

### 5.2.3. Mother Terms and the Generation of Multi-word (Compound) Terms — KWIC Lines

#### *Key-word-in-context (KWIC) searches*

One simple method which could be used to isolate the multi-word terms that were missed with KeyWord, is to use single-word mother terms in key-word-in-context (KWIC) searches. Such searches will reveal the collocations in which these terms are involved, which often turn out to be multi-word terms. When used for terminological purposes, this procedure is however not without its problems. The purpose of a concordance is primarily to identify collocations, i.e. — to use Scott's (1999: WST help) metaphor — to provide 'information on the company words keep'. It cannot simply be assumed that all collocations showing up in concordance lines are multi-word terms. From a multi-word perspective, KWIC lines tend to over-generate, producing false positives as well as true positives. In this regard Heid (2001: 791) states that "[t]he relationship and the borderline between collocations [...] and 'multiword terms' is not easy to describe". The combination of a term and its collocate(s) seems to be a linguistic issue rather than a terminological one. From the existing literature, it would seem that there are two basic principles which provide guidance as to the distinction between collocations and multi-word terms.

#### *Term : Collocate(s) combinations and The denomination of new concepts*

In the first instance, the terminological status of the *term : collocate(s)* combination depends on whether the combination of a term and its collocate(s) can be seen as the denomination of a new concept in its own right. If this is the case, such a collocational combination will qualify as a multi-word term; if not, it would be described as a false positive. In some cases, false positives in a concordance are quite obvious and easily identifiable, whereas others seem to be on the borderline between multi-word terms on the one hand, and simple collocations on the other. The concordance for the word **lediri** 'verb' serves as an example. There are 391 instances of **lediri** in the linguistic texts. With WST's concordance function, Concord, a concordance line can be drawn up for each, and one finds 15 incidences of **lethuši le lediri** 'auxiliary verb and verb'. This



collocation is obviously a false positive and cannot be regarded as a multi-word term, since it does not comply with the basic requirement for multi-word term status, i.e. the combination of the term **lediri** 'verb' and its collocate **lethuši (le)** 'auxiliary verb (and)' does not refer to a single and/or new concept. Other cases are, however, more problematic: 26 incidences of the combination **modirišo wa lediri** 'mood of the verb' were found — the collocate **modirišo (wa)** 'mood (of)' having the highest co-occurrence frequency with the term **lediri**. Even with the guidance of the principle of 'denomination of a new concept', it is difficult to decide whether **modirišo wa lediri** is to be regarded as a multi-word term, or whether it simply indicates the company which **lediri** typically keeps.

*Term : Collocate(s) combinations and Lexical gaps across languages*

A second principle which could be useful for the terminologist trying to decide on the terminological status of collocations, is to ascertain whether the combination of term and collocate(s) in one language fills a lexical gap when compared to a dominant language such as English; in other words, does the term and its collocate(s) have a *term* as translation equivalent in another language? This might lead the terminologist to decide that **modirišo wa lediri** 'mood of the verb' should not be regarded as a multi-word term, since its equivalent in English does not represent a term. On the other hand, a collocation such as **tatelano ya mantšu** 'word order' (where **tatelano (ya)** 'order (of)' was found as collocate for the term **mantšu** 'words') may then be regarded as a multi-word term, since it does: (a) refer to a concept in its own right, and (b) have a term, i.e. 'word order', as an equivalent in English. The downside of this principle is of course that in some cases it simply shifts the decision as regards the terminological status of a multi-word unit from one language to another.

*Language-specific guidelines*

It is clear that the two principles as formulated above are not sufficient as regards the distinction between collocations and multi-word terms. It is therefore understandable that Heid (2001) recognises the need for the establishment of some internal guideline as regards the distinction between collocations and multi-word terms in any given language. He does this by first identifying frequent collocational patterns appearing in the Indo-European languages, and classifies these patterns according to the part-of-speech category to which the two lexical items making up the collocation belong. Using the structure of these recurrent collocational patterns as basis, he attempts to formulate an internal and language-specific guideline in order to distinguish between mere collocational patterns and true multi-word terms. The syntactic categories which he distinguishes for Indo-European languages are as follows: (a) noun + verb, (b) noun + adjective, (c) noun + noun, (d) verb + adverb, and (e) adjective + adverb. For terminological purposes, only the first three categories (a)–(c) seem to

be of importance, since they are much more frequent in special-field languages than (d) and (e).

With specific reference to German, Heid (2001: 791) mentions the possibility of classifying sub-type denoting noun-adjective collocations as multi-word terms, and noun-verb collocations simply as combinatory properties of the nominal term. He uses the following noun-adjective type collocations taken from maintenance literature for automobiles as illustrative examples: (a) **pneumatische Leuchtweitenregulierung**, (b) **elektrische Leuchtweitenregulierung**, and (c) **automatische Leuchtweitenregulierung**. These combinations are regarded as being sub-type denoting noun-adjective collocations and each should therefore be regarded as a multi-word term. The adjectives **pneumatische** 'pneumatic', **elektrische** 'electric' and **automatische** 'automatic' denote sub-types of **Leuchtweitenregulierung** 'light-distance regulation' in that they define or describe different types of light-distance regulation. A noun-verb collocation such as **Parameter festlegen** 'fix a parameter' (Heid 2001: 790) would then not be regarded as a multi-word term — **festlegen** 'fix' would simply be regarded as a combinatory property of the term **Parameter** 'parameter'.

#### *Internal guidelines for Northern Sotho*

The point that needs to be emphasised at this stage is that no attempt has hitherto been made to formulate internal guidelines for Northern Sotho that would enable terminologists to distinguish mere collocations from multi-word terms. Since such internal guidelines are dependent on the structure of frequently occurring collocations, a preliminary investigation was consequently done by making use of Concord in order to identify the typical collocational *patterns* that are found in Northern Sotho concordance lines. (Note that it is only on this level that Sewangi (2000, 2001), see par. 5.1, starts to use the power of the tagged Swahili corpora he had at his disposal. In his approach a 'pattern matching program' semi-automatically extracts potential multi-word terms, using *pre-defined* term-formation patterns.)

Within WST, a concordance can be drawn up from either a KeyWord list or a WordList list. The output of such a query can then be sorted in a multitude of ways so as to see different collocational patterns emerge. In the case of English, sorting the output to the left shows up modifier-head patterns where the mother term is the *head*, e.g. **infinitive verb**, **singular verb**, **transitive verb**, etc., whereas sorting to the right results in modifier-head patterns in which the mother term is the *modifier*, e.g. **verb phrase**, **verb root**, etc. These patterns are language-specific and will have to be formulated for any particular language in which any such investigation is to be done.

For the purpose of the current discussion, the term **mantšu** 'words', which appears high up in both the keyness-ordered KeyWord list and the frequency-ordered WordList list, can be used as a search term. Concord displays 392 concordance lines for **mantšu**. By sorting to the left (primary sort on *L1*, i.e. first position to the left of the search item, and secondary sort on *L2*, i.e. second

position to the left) two typical patterns can be identified. The pattern which occurs most frequently is the noun-determiner pattern, in which the search item **mantšú** is the nucleus of the determiner. Compare the examples shown in (6), taken from Concord.

- (6) Noun-determiner pattern, with the search item the nucleus of the determiner

Thutišo ye e leka go go lemoša bohlokwa bja *tatelano* ya **mantšú** mo lefokong. Go ya ka melao ya tatelanontšú ya Sesotho sa ke gore lefoko le lengwe le le lengwe le na le *tatelanotšeo* ya **mantšú**. Gape re tlišitše taba ya go re lentšú la mathomo mo lefokong, mmadi goba mmoledišwa mo kgokaganong, ka *ge tlhalošo* ya **mantšú** a mohuta wo e ka ba e na le maphakga a mantši. Eupša,

The second most frequent pattern is the verb-noun pattern in which the search item typically appears as the object of a verb. Examples are shown in (7).

- (7) Verb-noun pattern, with the search item typically as the object of a verb

mokgwa woo. Ge batho ba boledišana, ba swanetše *go kgetha* **mantšú** le go bopa mafoko semeetseng. Gantši mmoledi o timelelwa (iii) 3. Bopa mafoko a mahlano ka *go šomiša* **mantšú** a mangwe bakeng sa mantšú a a dirišitšwego mo lefokong le: di a fapano? Na e ka ba ke eng se se dirago gore *ba tswalanye* **mantšú** a a filwego le ditlhalošokelello tše di fapanego? MODIRO 3

When sorting takes place to the right of the search word, the typical pattern which emerges is again a noun-determiner pattern, but in this case the search item forms the nucleus of the nominal part of the combination. This can be seen from the examples in (8).

- (8) Noun-determiner pattern, with the search item the nucleus of the nominal part

e hlolwa ke tlhalošo ya deiktiki. Gore re kwešiše taba ye ya **mantšú** a *deiktiki* gabotse, a re fetelele poledišana gare ga mmoledi go šetša molao wo, gomme a ka se kgone go tswalanya **mantšú** a *lefoko* ka boikgethelo. Ge mmoledi goba mongwadi a ka

A pattern which also frequently occurs, is the noun-verb pattern, but in this instance, the search term appears as the subject of the verb. Compare (9) in this regard.

- (9) Noun-verb pattern, with the search item typically as the subject of a verb

a e šomišago e šašarakane. 2.3.4 Tswalano ka tlhalošo Ka ge **mantšú** a *fapano* ka tlhalošo, ga a šome go swana mo lefokong. Ge re mantšú ka gona ge re bopa dikafoko goba mafoko. Ge **mantšú** a *hlatlamantšhwa* mo lefokong, tthatlamanano (tatelano) yeo e , o tšweletšwa ka mafoko ao a bopilwego ka mantšú, gomme **mantšú** a *latelana* ka tsela ye e itšego mo mafokong. Ge re šetša

With regard to Northern Sotho, the principles regarding the term status of certain collocational patterns as formulated by Heid, would for example imply that the collocations **lediri la modirišopego** 'indicative verb', **lediri la modirišopegotlhaodi** 'participial / situative verb' and **lediri la modirišogore** 'subjunctive verb' should all be regarded as multi-word terms, since the function of the determiners **la modirišopego**, **la modirišopegotlhaodi**, etc. is similar to that of the adjectives cited in Heid's example, i.e. to denote sub-types of **lediri**. A general principle for Northern Sotho could thus be to regard sub-type denoting noun-determiner collocations as multi-word terms, provided of course that these

collocations also meet the requirement of referring to an independent concept. Such an approach seems to make sense from a theoretical point of view, but the practical implications it would have for the full-scale compilation of a terminology list and/or an LSP dictionary remain to be seen.

If the principle to *regard noun-verb collocations as combinatory properties of the noun* is applied to one of the collocations found for **lediri**, viz. **lediri le tšwelela ...** 'a/the verb generates ...', this would imply that this combination should thus not be regarded as a multi-word term, which seems sound. One could however assume that in a dictionary on linguistic terms, the frequent combination of **lediri** 'verb' and **tšwelela** 'generate' should in some way or other be reflected in the article of **lediri**. This argument is further supported by the statistical analysis of collocates, which can be done computationally from the concordance given for any search item. The statistical analysis of collocates gives an indication of the significance of certain items appearing within a specified span or horizon of the node term. A calculation of the collocates of **lediri** 'verb' indicates that the verb **tšwelela** 'generate' is the verb which appears most often (26 times) as a collocate of the noun **lediri**. Therefore, even though the combination **lediri le tšwelela ...** 'a/the verb generates ...' cannot be regarded as a multi-word term, it is clear that it should be treated in a more detailed terminology list and/or LSP dictionary, either as a combinatory property of the noun, or, alternatively, used in an explanatory example in the microstructural treatment of the specific noun.

It should however be kept in mind that the two guidelines formulated above, i.e. to regard sub-type denoting noun-determiner collocations as potential multi-word terms and noun-verb collocations as combinatory properties of the noun in Northern Sotho, should at this stage be seen as suggestions, since the formulation of such guidelines presupposes an in-depth study of the linguistic structure of collocations in Northern Sotho — an endeavour which falls outside the scope of the current article.

#### *Semi-automatic generation of multi-word (compound) terms*

The results of the above discussion can now be used to semi-automatically generate multi-word terms by means of KWIC lines where the node is a single-word term. Recall that for the generation of single-word (compound) terms (par. 5.1.4), the stems / roots of the top 25 unique KeyWord mother terms were used. Those same 25 mother terms can now function as node terms for the KWIC lines. For comparison purposes, the results are also shown in (4) above.

From (4) one sees that, compared to the various multi-word terms thrown up by the multi-word KeyWord lists, the 25 mother terms generated 7 multi-word terms that had also been marked manually, but in addition also an astonishing 16 multi-word terms that were missed manually (and were not picked up by KeyWord either). Together, KeyWord and Concord for the top 25 unique single-word mother terms, thus throw up 22 + 8 and 7 + 16 multi-word terms respectively, or thus 53 terms in all. With a benchmark of 41 + 8 + 16 =

65, this means that 82% of the multi-word terms (known so far) were extracted semi-automatically.

#### 5.2.4. Clusters

If kept in mind that the multi-term words are the hardest to retrieve, a recall of 82% seems exceptional. Nonetheless, besides the unsuccessful study of the multi-word top ranks on the one hand, and the very successful study of KeyWord and Concord on the other, a fourth approach to extract even more multi-words was experimented with, viz. the Cluster function. In the words of Scott (1999: WST help):

Clusters are words which are found repeatedly in each others' company. They represent a tighter relationship than collocates, more like groups or phrases (but I call them clusters because these terms already have uses in grammar).

Clusters can be identified from either a wordlist or a concordance, the difference being that Concord only processes concordance lines, whereas WordList processes whole texts. As we are primarily interested in clusters containing the search item itself, we opted for Clusters in Concord. The 2-, 3-, 4- and 5-word clusters for each of the top 25 unique single-word KeyWord mothers were calculated. The outcome of this Cluster procedure was truly surprising, as yet another extra 7 multi-word terms, shown in (10), were extracted.

- (10) Generation of *extra* multi-word linguistic terms with the Cluster function (extra = not marked manually, and not found with KeyWord nor KWIC lines)

**mmoledišwa ka botee** second person singular  
**peakanyo ya mantšu** arrangement of words, word arrangement  
**popogo ya lediri** morphology of the verb, verbal morphology  
**tlhalošo ya lefoko** sentence meaning  
**tlhalošo ya lentšu** word meaning  
**tlhalošotheo ya lefoko** basic sentence meaning  
**tumanoši ya mafelelo ya lediri** verbal ending (lit. final vowel of the verb)

None of the multi-word terms listed in (10) had been marked in the manual pass, nor were they picked up with KeyWord or the study of KWIC lines. With these 7 extras, the recall for the multi-word terms becomes 60 out of 72, which thus means that as many as 83% of the multi-word terms were extracted semi-automatically!

## 6. Outlook: Manual Excerption *versus* Semi-automatic Extraction of Terminology

A very significant result of the research presented above is the fact that over

half the multi-word terms identified by computational means ( $8 + 16 + 7 = 31$ , out of 60) were overlooked during the process of manual excerption. The manual excerption of multi-word terms therefore merits some further discussion. Identification of multi-word terms is a problematic issue for terminologists, since the terminological status of especially multi-word units is not always clear-cut.<sup>9</sup> The process of manual term excerption is subject to the influence of a number of factors, and rightly so. According to the principles of good terminological practice, term excerption cannot be done without taking into account practical issues such as *inter alia* the potential target user and that user's encyclopaedic knowledge regarding the specific subject field. Taking the target user into account, the terminologist might argue that the conceptual meaning of a multi-word unit can be derived from the meaning of its constituent parts and that it is therefore unnecessary to list the particular unit as a term. For a different target user, the terminologist might deem it necessary to indeed list the multi-word unit as a term. An unexpected, and certainly underestimated, value of computational term extraction therefore seems to lie in its ability to *identify* term candidates. Taking all relevant information as regards target user, etc. into account, the terminologist can then make an informed decision as to whether a particular multi-word unit should be included as a multi-word term or not.

Apart from the research reported on above, experiments were also done with other concepts that were introduced by Scott (1997), viz. *key key words* (i.e. items that are key across numerous texts), *associates* (i.e. items that are key in the same texts as a given key key word), and *clumps* (i.e. co-occurring associates). With these concepts (and their computational implementations) we hoped to be able to automatically group terms into sub-fields. So far, however, the results of these attempts have not been satisfying.

## 7. In Conclusion

The main aim of this article was to discover whether or not the semi-automatic extraction of terminology from *untagged* and *unmarked* running text by means of *basic* corpus query software would be feasible for the African languages. In order to answer this question a full-blown case study revolving around Northern Sotho linguistic texts was undertaken. Upon comparison of the manual outcome with the computational results, it was found that 74% of the single-word linguistic terms, and an astonishing 83% of the multi-word linguistic terms could indeed be extracted semi-automatically. These high figures were obtained with basically just three software tools: WordList, KeyWord and Concord, all part of WordSmith Tools (Scott 1999). Based on this case study one is thus bound to conclude that the semi-automatic extraction of terms for the African languages is indeed a *viable* endeavour.

It was also pointed out that human beings will always remain the final judges in any terminological activity, whether that endeavour be manual or computational. The terms proffered by the software will always need to be

scrutinised by the terminologist. Conversely, however, the research revealed rather surprisingly that the software can *isolate* potential terms and *force* the terminologist to *consider* term status in ways that are less obvious when wading manually through running text. This turned out to be especially valid for multi-word terms, as more than 40% of the multi-word linguistic terms were seemingly missed during manual excerption. Viewed from this angle, the semi-automatic extraction of terms for the African languages is not only viable, but even *crucial* in order to counteract inevitable human errors.

Finally, at the start of this article we pointed out that a terminographical approach to terminology would enable the computational management of terminology. Terminography, whether pursued manually or computationally, always has the creation of a terminology list, a term bank or a dictionary for special purposes as its final objective. The various outcomes of the research presented in this article are therefore summarised in a tiny special-field *lexicon*. This lexicon is the Linguistics Terminology presented in the Appendix, in which all the terms that were retrieved in the current study are listed in alphabetical order and in their *lemmatised* form. In addition, the linguistic terms that were only excerpted manually are preceded by ☞, those that were only extracted computationally are preceded by ☞, and the ones that were retrieved both manually and computationally are marked with ☞. There are respectively 98, 50 and 187 of them. This means that, out of the 335 lemmatised terms, 285 or 85% were excerpted manually, and 237 or 71% were extracted computationally. The difference between the two approaches (14%) is smaller than the number of items not retrieved in either approach. There can thus be no doubt that, when looking at the end product, semi-automatically extracting terminology for and in the African languages is indeed a worthwhile venture.

## Endnotes

1. Since this article is being submitted for publication in a South African journal, necessary sensitivity with regard to the term 'Bantu' languages is exercised in our choice rather to use the term *African* languages. Keep in mind, however, that the latter includes more than just the 'Bantu Language Family'.
2. General-language corpora for all South African languages have indeed been built at the Department of African Languages of the University of Pretoria. The sizes of these corpora are in constant evolution. For the latest developments, we would therefore like to refer the reader to the home page of ELC for ALL (Electronic Corpora for African-Language Linguistics): <http://www.up.ac.za/academic/libarts/afrilang/elcforall.htm>
3. For more information on WordSmith Tools, we would like to refer the reader to the home page of Mike Scott, the creator of the software: <http://www.lexically.net> (or else: <http://www.liv.ac.uk/~ms2928>).
4. The term 'keyness' too remains undefined. Scott and Thompson (2001a: 109) state rather vaguely that "[k]eyness relates to the frequency of particular lexical items within a text as compared with their frequency in a reference corpus". Unfortunately, even under the head-

ing 'definition of keyness' in WST's help section (Scott 1999), 'key words' are discussed and not 'keyness'.

5. The text on the new South African coat of arms can be found at: [http://www.gov.za/symbols/coa\\_sepedi.htm](http://www.gov.za/symbols/coa_sepedi.htm)
6. Compare in this respect Prinsloo and De Schryver (2001) which deals extensively with corpus-stability issues, with special reference to Northern Sotho and Tsonga.
7. The spelling of the adjective **seswa** is incorrect in the CoA text. The correct orthography is **sefsa**, but we have chosen to keep the spelling as it appears on the cited web page. The same goes for **leswa** which should be **lefsa**.
8. Due to the morphology of Northern Sotho the subject concord and/or demonstrative **se**, which refers to **sefoka** 'coat of arms', is also thrown up by KeyWord but does of course not appear as such in the English-language paraphrase.
9. This is of course also true of single-word terms, but in this particular instance the relatively high correspondence between the manually excerpted single-word terms and those extracted by semi-automatic means, indicates that manual identification of single-word terms is less of a problem for terminologists.




## References

- Ahmad, Khurshid and Margaret Rogers.** 2001. Corpus Linguistics and Terminology Extraction. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001: 725-760.
- Alberts, Mariëtta.** 2000. Terminology Management at the National Language Service. *Lexikos* 10: 234-251.
- De Schryver, Gilles-Maurice and B. Lepota.** 2001. The Lexicographic Treatment of Days in Sepedi, or When Mother-Tongue Intuition Fails. *Lexikos* 11: 1-37.
- De Schryver, Gilles-Maurice and D.J. Prinsloo.** 2000. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The Macrostructure. *South African Journal of African Languages* 20(4): 291-309.
- Dunning, Ted.** 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1): 61-74.
- Heid, Ulrich.** 2001. Collocations in Sublanguage Texts: Extraction from Corpora. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001: 788-808.
- Hurskainen, Arvi.** 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology. An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122.
- Hurskainen, Arvi.** 1995. Information Retrieval and Two-directional Word Formation. *Nordic Journal of African Studies* 4(2): 81-92.
- Hurskainen, Arvi.** 1996. Disambiguation of Morphological Analysis in Bantu Languages. *COLING-96. The 16th International Conference on Computational Linguistics, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996. Proceedings, Volume 1: 568-573.*
- Hurskainen, Arvi.** 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.





- Prinsloo, D.J. and Gilles-Maurice de Schryver.** 2001. Monitoring the Stability of a Growing Organic Corpus, with Special Reference to Sepedi and Xitsonga. *Dictionaries: Journal of the Dictionary Society of North America* 22: 85-129.
- Sager, Juan C.** 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.
- Scott, Mike.** 1997. PC Analysis of Key Words — and Key Key Words. *System* 25(1): 233-245.
- Scott, Mike.** 1997a. The Right Word in the Right Place: Key Word Associates in Two Languages. *AAA — Arbeiten aus Anglistik und Amerikanistik* 22(2): 235-248.
- Scott, Mike.** 1999. *WordSmith Tools version 3*. Oxford: Oxford University Press. See for this software also: <http://www.lexically.net/wordsmith/index.html>.
- Scott, Mike.** 2000. Reverberations of an Echo. Lewandowska-Tomaszczyk, Barbara and Patrick J. Melia (Eds.). 2000. *PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Lodz, 15-18 April 1999*: 49-65. Lodz Studies in Language 1. Frankfurt am Main: Peter Lang.
- Scott, Mike.** 2000a. Focusing on the Text and its Key Words. Burnard, Lou and Tony McEnery (Eds.). 2000. *Rethinking Language Pedagogy from a Corpus Perspective. Papers from the Third International Conference on Teaching and Language Corpora*: 103-121. Lodz Studies in Language 2. Frankfurt am Main: Peter Lang.
- Scott, Mike.** 2001. Mapping Key Words to *problem* and *solution*. Scott, Mike and Geoff Thompson (Eds.). 2001: 109-127.
- Scott, Mike and Geoff Thompson (Eds.).** 2001. *Patterns of Text: In Honour of Michael Hoey*. Amsterdam: John Benjamins.
- Scott, Mike and Geoff Thompson.** 2001a. Editors' Introduction. Scott, Mike and Geoff Thompson (Eds.). 2001: 109-110.
- Sewangi, Selemán S.** 2000. Tapping the Neglected Resource in Kiswahili Terminology: Automatic Compilation of the Domain-Specific Terms from Corpus. *Nordic Journal of African Studies* 9(2): 60-84.
- Sewangi, Selemán S.** 2001. *Computer-assisted Extraction of Terms in Specific Domains: The Case of Swahili*. Academic Dissertation, December 2001. Helsinki: University of Helsinki, Faculty of Arts, Institute of Asian and African Studies. Also available at: <http://ethesis.helsinki.fi/julkaisut/hum/aasia/vk/sewangi/>.
- Strehlow, Richard A.** 2001. The Role of Terminology in Retrieving Information. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001: 426-441.
- Wright, Sue Ellen and Gerhard Budin (Eds.).** 2001. *Handbook of Terminology Management. Volume 2. Application-Oriented Terminology Management*. Amsterdam: John Benjamins.

## Appendix: Manually-excerpted and/or Semi-automatically Extracted Linguistics Terminology








- Legend:  = only excerpted manually  
 = only extracted computationally  
 = retrieved both manually and computationally  
x/y = gender (singular and plural class) of noun  
v. = verb  
⇒ = cross-reference to synonym(s) and/or variant(s)

### A


-  **-a- ya lebjaletelelele** 9/- imperfect tense -a-

-  **amana** v. relate



### B

-  **boemo** 14/6 position  
 **boemo bja motheo** 14/6 basic position ⇒ **boemotheo**  
 **boemotheo** 14/6 basic position ⇒ **boemo bja motheo**  
 **boikakanyetšo** 14/6 idea, thought, concept  
 **boinaganelo** 14/6 consciousness, imagination, pre-knowledge  
 **bontši** 14/- plural  
 **botee** 14/- singular


### D

-  **deiktiki** 9/- deictic





### F

-  **fonetiki** 9/- phonetics  
 **fonolotši** 9/- phonology




### G

-  **gatelela** v. emphasise

### H

-  **hlogo** 9/10 prefix  
 **hlogo ya leina** 9/10 nominal prefix, noun prefix ⇒ **hlogoina**  
 **hlogoina** 9/10 nominal prefix, noun prefix ⇒ **hlogo ya leina**  
 **hlopha** v. categorise, arrange

### K

-  **kago** 9/10 structure  
 **kamano** 9/10 (inter)relationship  
 **kamanotlhalošo** 9/10 semantic relationship

-  **karolo ya lefoko** 9/10 sentence part ⇒ **karolofoko**  
 **karolofoko** 9/10 sentence part ⇒ **karolo ya lefoko**  
 **karolopolelo** 9/10 word class, part of speech ⇒ **legorontšu**  
 **kganetšo** 9/10 negation, negative  
 **kgatelelo** 9/10 emphasis  
 **kgetho ya mantšu** 9/10 word choice  
 **kgoeletšo** 9/10 exclamation  
 **kgokagano** 9/10 discourse  
 **kgokagano ka go ngwala** 9/10 written discourse  
 **kgokagano ka molomo** 9/10 spoken / oral discourse  
 **kgokagano ya linkwistiki** 9/10 linguistic discourse / communication  
 **kgokaganya** v. connect, link  
 **kgopolotheo** 9/10 main idea  
 **khuduego** 9/10 excitement, enthusiasm  
 **khutlo** 9/10 full stop  
 **khutsofatša** v. abbreviate  
 **khutsofatšo** 9/10 abbreviation  
 **kopanyo ya mantšu** 9/10 combination of words, word combination  
 **kutu** 9/10 stem  
 **kutu ya lediri** 9/10 verb stem ⇒ **kutudiri**  
 **kutudiri** 9/10 verb stem ⇒ **kutu ya lediri**  
 **kwagatšo** 9/10 sound production  
 **kwana** v. agree  
 **kwano** 9/10 agreement  
 **kwano ka popego** 9/10 morphological agreement

## L

- 📖 **leadingwa** 5/6 adoptive, loan word  
 🗣️ **leamanyi** 5/6 relative  
 🗣️ **leamanyidiri** 5/6 verbal relative, relative verb  
 🗣️ **leamanyi-ina** 5/6 nominal relative  
 📖 **learogi** 5/6 exception  
 📖 **leba** 5/6 copula(tive)  
 📖 **lebadl** 5/6 numeral  
 📖 **lebadiri** 5/6 copulative verb  
 📖 **lebahlogwana** 5/6 copulative prefix  
 🗣️ **lebaka la lefetile** 5/- past tense ⇒ **lefetile**  
 📖 **lebaka la letlago** 5/- future tense ⇒ **letlago**  
 🗣️ **lebjale** 5/- present tense  
 📖 **lebjale le lekopana** 5/- short present tense ⇒ **lebjalekopana**  
 📖 **lebjale le letelele** 5/- long present tense ⇒ **lebjaletelele**  
 📖 **lebjalekopana** 5/- short present tense ⇒ **lebjale le lekopana**  
 🗣️ **lebjaletelele** 5/- long present tense ⇒ **lebjale le letelele**  
 🗣️ **lebopi** 5/6 morpheme  
 📖 **lebopi la kganetšo** 5/6 negative morpheme ⇒ **lebopikganetši**  
 🗣️ **lebopi la lebaka la letlago** 5/6 future tense morpheme  
 📖 **lebopi la lebjaletelele** 5/6 present tense morpheme  
 🗣️ **lebopikganetši** 5/6 negative morpheme ⇒ **lebopi la kganetšo**  
 📖 **lebotšiši** 5/6 interrogative  
 🗣️ **lediri** 5/6 verb  
 📖 **lediri la modirišogore** 5/6 subjunctive verb  
 📖 **lediri la modirišopego** 5/6 indicative verb  
 📖 **lediri la modirišopegotlhaodi** 5/6 participial / situative verb  
 📖 **lediri la modirišotaelo** 5/6 imperative verb  
 🗣️ **ledirifelopedi** 5/6 verb which can combine with a subject and one object, two-place verb  
 🗣️ **ledirifelotharo** 5/6 verb which can combine with a subject and two objects, three-place verb  
 🗣️ **leekiši** 5/6 ideophone  
 🗣️ **lefeledi** 5/6 intransitive verb  
 🗣️ **lefetedi** 5/6 transitive verb  
 🗣️ **lefetile** 5/- past tense ⇒ **lebaka la lefetile**  
 🗣️ **lefoko** 5/6 sentence  
 📖 **lefoko le le feleletšego** 5/6 full sentence  
 🗣️ **lefokofoko** 5/6 complete sentence  
 🗣️ **lefokofokwana** 5/6 complex sentence ⇒ **lefokontši**  
 📖 **lefokofokwanapego** 5/6 complex declarative sentence  
 🗣️ **lefokonolo** 5/6 basic / simple sentence ⇒ **lefokotho**  
 🗣️ **lefokontši** 5/6 complex sentence ⇒ **lefokofokwana**  
 🗣️ **lefokotho** 5/6 basic / simple sentence ⇒ **lefokonolo**  
 📖 **lefokothwi** 5/6 direct speech  
 🗣️ **legoro** 5/6 (noun) class  
 🗣️ **legorofelo** 5/6 locative noun class  
 🗣️ **legoroina** 5/6 noun class  
 🗣️ **legorontšu** 5/6 word class, part of speech ⇒ **karolopolelo**  
 🗣️ **lehlalošagotee** 5/6 synonym ⇒ **lehlalošetšagotee**  
 🗣️ **lehlalošantši** 5/6 polysemous word ⇒ **lehlalošetšagontši**  
 🗣️ **lehlalošetšagontši** 5/6 polysemous word ⇒ **lehlalošantši**  
 🗣️ **lehlalošetšagotee** 5/6 synonym ⇒ **lehlalošagotee**  
 🗣️ **lehlaodi** 5/6 adjective  
 🗣️ **lehlathafelo** 5/6 adverb of place  
 🗣️ **lehlathamokgwa** 5/6 adverb of manner  
 🗣️ **lehlathanako** 5/6 temporal adverb, adverb of time  
 🗣️ **lehlathasedirišwa** 5/6 instrumental adverb  
 🗣️ **lehlathi** 5/6 adverb  
 🗣️ **leina** 5/6 noun  
 🗣️ **leinaina** 5/6 proper name  
 📖 **leinakgopolo** 5/6 abstract noun  
 🗣️ **leinataodi** 5/6 head noun  
 🗣️ **lekgoka-amanyi** 5/6 relative con-

- cord
- 📖 **lekgokahlaodi** 5/6 adjectival concord, qualificative concord ⇒ **lekgokatlhaodi**
  - 📖 **lekgokamong** 5/6 possessive concord ⇒ **lekgokarui**
  - 🕒 **lekgokarui** 5/6 possessive concord ⇒ **lekgokamong**
  - 🕒 **lekgokasediri** 5/6 subject concord
  - 📖 **lekgokasediri la mmoledišwa ka bontši** 5/- subject concord second person plural
  - 📖 **lekgokasediri la mmoledišwa ka botee** 5/- subject concord second person singular
  - 🕒 **lekgokasedirwa** 5/6 object concord
  - 📖 **lekgokatlhaodi** 5/6 adjectival concord, qualificative concord ⇒ **lekgokahlaodi**
  - 🕒 **lekgokedi** 5/6 agreement morpheme
  - 🕒 **lekopanyi** 5/6 conjunction
  - 🕒 **lekopanyibaka** 5/6 causal connector / conjunction
  - 🕒 **lekopanyikoketšo** 5/6 additive connector / conjunction
  - 🕒 **lekopanyinako** 5/6 temporal connector / conjunction
  - 🕒 **lekopanyipeelano** 5/6 conditional connector / conjunction
  - 🕒 **lekopanyipharologanyo** 5/6 adversative connector / conjunction
  - 🕒 **lelahlelwa** 5/6 interjection
  - 🕒 **lelatodi** 5/6 antonym, opposite
  - 🕒 **lentšu** 5/6 (linguistic) word
  - 🕒 **lentšu la deiktiki** 5/6 deictic word
  - 🕒 **lentšugokwa** 5/6 compound word
  - 📖 **lentšutheo** 5/6 head word
  - 🕒 **lereo** 5/6 term
  - 🕒 **lereokakaretšo** 5/6 general term, umbrella term
  - 🕒 **lerui** 5/6 possessive
  - 📖 **leruo** 5/6 possession
  - 🕒 **lešala** 5/6 pronoun
  - 🕒 **lešalagohle** 5/6 quantitative pronoun
  - 🕒 **lešalapadi** 5/6 quantitative pronoun
  - 🕒 **lešalašala** 5/6 absolute pronoun
  - 🕒 **lešalašupi** 5/6 demonstrative pronoun
  - 📖 **lešupakarolo** 5/6 word referring to a part of a whole ⇒ **lešupetšakarolo**, **sešupetšakarolo**
  - 🕒 **lešupetšagotee** 5/6 coreferent
  - 📖 **lešupetšakarolo** 5/6 word referring to a part of a whole ⇒ **lešupakarolo**, **sešupetšakarolo**
  - 📖 **leswaodikga** 5/6 punctuation mark
  - 🕒 **lethekgi** 5/6 stabiliser
  - 🕒 **lethuši** 5/6 auxiliary verb
  - 🕒 **letlago** 5/- future tense ⇒ **lebaka la letlago**
  - 🕒 **letlema** 5/6 preposition
  - 🕒 **letšwalediring** 5/6 deverbative
  - 🕒 **logaganya** *v.* integrate
- ### M
- 📖 **maikemišetšo** -/6 purpose, aim, intention ⇒ **malebiša**
  - 📖 **malebiša** -/6 purpose, aim, intention ⇒ **maikemišetšo**
  - 📖 **malebišatheo** -/6 basic purpose / aim / intention
  - 🕒 **mmoledi** 1/2 (*pl. is baboledi*) first person
  - 🕒 **mmoledišani** 1/2 (*pl. is baboledišani*) interlocutor
  - 🕒 **mmoledišwa** 1/2 (*pl. is baboledišwa*) second person, addressee
  - 📖 **mmoledišwa ka bontši** 1/- second person plural
  - 📖 **mmoledišwa ka botee** 1/- second person singular
  - 📖 **mmolelwa** 1/2 (*pl. is babolelwa*) third person
  - 📖 **moanegwa** 1/2 person who is being described, character
  - 🕒 **modirišo** 3/4 mood
  - 🕒 **modirišogo** 3/- infinitive mood
  - 🕒 **modirišogore** 3/- subjunctive mood
  - 🕒 **modirišokanegelo** 3/- consecutive mood
  - 🕒 **modirišopego** 3/- indicative mood
  - 🕒 **modirišopegotlhaodi** 3/- participial form of indicative mood, situative mood

- 🗨️ **modirišotaelo** 3/- imperative mood
- 🗨️ **modirišotlwaelo** 3/- habitual mood
- 🗨️ **modirišotona** 3/- main mood
- 🗨️ **modiro** 3/4 function
- 🗨️ **modu** 3/4 root
- 🗨️ **modu wa lediri** 3/4 verb root
- 🗨️ **modumo** 3/4 speech sound
- 🗨️ **mofolotši** 3/- morphology ⇒ **popegopolelo**
- 🗨️ **mokgwapolelo** 3/4 language use
- 🗨️ **molaetšatebanyo** 3/4 intended message
- 🗨️ **molaetšatheo** 3/4 main idea
- 🗨️ **molao wa kgokagano** 3/4 discourse rule
- 🗨️ **molao wa polelo** 3/4 linguistic / language rule ⇒ **molaopolelo**
- 🗨️ **molao wa popafoko** 3/4 syntactic rule
- 🗨️ **molao wa popegopolelo** 3/4 morphological rule
- 🗨️ **molao wa tlhalošo** 3/4 semantic rule
- 🗨️ **molao wa tlhatlamano wa tatanontšu** 3/4 word order rule
- 🗨️ **molao wa tšhomišano** 3/4 cooperative principle
- 🗨️ **molaokgokagano** 3/4 discourse / communication rule
- 🗨️ **molaopolelo** 3/4 linguistic / language rule ⇒ **molao wa polelo**
- 🗨️ **molaotheo** 3/4 basic rule
- 🗨️ **momagana** *v.* coalesce
- 🗨️ **momagano** 3/4 coalescence
- 🗨️ **mong** 1/2 (*pl. is beng*) possessor
- 🗨️ **mongwalo wa fonetiki** 3/4 phonetic orthography
- 🗨️ **mosela** 3/4 suffix
- 🗨️ **mosela wa bontši** 3/4 plural suffix
- 🗨️ **motheeletši** 1/2 addressee
- 🗨️ **motheo** 3/4 basic element, basis

## N

- 🗨️ **ngangego** 9/10 disagreement
- 🗨️ **noko** 9/10 syllable
- 🗨️ **nokotee** 9/10 monosyllable
- 🗨️ **NP** noun phrase
- 🗨️ **nyenyefatšo** 9/10 diminution

## P

- 🗨️ **palo** 9/10 number
- 🗨️ **peakanyo ya mantšu** 9/10 arrangement of words, word arrangement
- 🗨️ **peakanyofoko** 9/10 syntactic arrangement (of constituents)
- 🗨️ **peelano** 9/10 condition
- 🗨️ **pego** 9/10 statement
- 🗨️ **peobakeng** 9/10 replaceability; substitution
- 🗨️ **pharologantšho** 9/10 (distinctive) feature
- 🗨️ **pharologantšho ya linkwistiki** 9/10 (distinctive) linguistic feature
- 🗨️ **pharologantšho ya popego** 9/10 (distinctive) morphological feature ⇒ **pharologantšhopopego**
- 🗨️ **pharologantšho ya seemotikologo** 9/10 (distinctive) discourse feature, discourse characteristic
- 🗨️ **pharologantšho ya tlhalošo** 9/10 (distinctive) semantic feature ⇒ **pharologantšhotlhalošo**
- 🗨️ **pharologantšhopopego** 9/10 (distinctive) morphological feature ⇒ **pharologantšho ya popego**
- 🗨️ **pharologantšhotlhalošo** 9/10 (distinctive) semantic feature ⇒ **pharologantšho ya tlhalošo**
- 🗨️ **pharologanyo** 9/10 distinction
- 🗨️ **phetleko** 9/10 analysis
- 🗨️ **phetleko ya kgokagano** 9/10 discourse analysis ⇒ **phetlekokgokagano**
- 🗨️ **phetlekokgokagano** 9/10 discourse analysis ⇒ **phetleko ya kgokagano**
- 🗨️ **phetlekotaodišo** 9/10 discourse analysis
- 🗨️ **poledišano** 9/10 dialogue
- 🗨️ **popafoko** 9/10 syntax
- 🗨️ **popego** 9/10 morphology, form, structure
- 🗨️ **popego ya lediri** 9/10 morphology of the verb, verbal morphology
- 🗨️ **popego ya lefoko** 9/10 sentence structure

- 🗨️ **popegopolelo** 9/10 morphology ⇨ **mofolotši**  
 📖 **popegopolelo ya Sesotho sa Leboa** 9/- Northern Sotho morphology  
**S**
- 🗨️ **sebopego** 7/8 (relating to) form  
 🗨️ **sediri** 7/8 subject  
 🗨️ **sedirwa** 7/8 object  
 📖 **sedirwa sa bobedi** 7/8 direct object (lit. second object)  
 📖 **sedirwa sa pele** 7/8 indirect object (lit. first object)  
 📖 **seema** 7/8 proverb  
 🗨️ **seemotikologo** 7/8 context  
 📖 **seemotikologo sa kgokagano** 7/8 discourse context  
 📖 **segagabo** 7/- (their) mother-tongue  
 📖 **segageno** 7/- (your) mother-tongue  
 📖 **segalo** 7/8 tone  
 📖 **seka** 7/8 idiomatic expression  
 🗨️ **sekafoko** 7/8 phrase  
 🗨️ **sekafokodiri** 7/8 verb phrase  
 🗨️ **sekafokoina** 7/8 noun phrase  
 📖 **senaganwa** 7/8 idea, thought, concept  
 🗨️ **serewa** 7/8 topic  
 📖 **serewa sa poledišano** 7/8 discourse topic  
 🗨️ **serewakgolo** 7/8 main topic  
 🗨️ **serewana** 7/8 sub-topic  
 🗨️ **serewapoeletšo** 7/8 re-introducing (discourse) topic  
 🗨️ **serewatiego** 7/8 delayed discourse topic  
 🗨️ **serewatirišano** 7/8 collaborating (discourse) topic  
 🗨️ **serewatšhielano** 7/8 introducing (discourse) topic  
 🗨️ **serewatswalano** 7/8 incorporating (discourse) topic  
 📖 **sešupetšakarolo** 7/8 word referring to a part of a whole ⇨ **lešupakarolo, lešupetšakarolo**  
 🗨️ **sešupša** 7/8 referent  
 📖 **swantšhiša** *v.* compare  
**T**
- 🗨️ **taelo** 9/10 command  
 📖 **taodišo** 9/10 essay
- 📖 **tatelano** 9/10 succession, sequence  
 🗨️ **tatelano ya mantšu** 9/10 word order ⇨ **tatelanontšu, tthatlamano ya mantšu, tthatlamanontšu**  
 🗨️ **tatelanontšu** 9/10 word order ⇨ **tatelano ya mantšu, tthatlamano ya mantšu, tthatlamanontšu**  
 📖 **tatelanotseo** 9/10 dominant / basic order  
 🗨️ **thabe** 9/10 clause  
 🗨️ **thabekutu** 9/10 main clause  
 🗨️ **thabenyana** 9/10 subordinate clause  
 📖 **thulano ya tlhalošo** 9/10 semantic incompatibility  
 🗨️ **thutamedumo** 9/- study of (speech) sounds, study of phonetics  
 🗨️ **thutapolelo** 9/- study of linguistics  
 🗨️ **thutapopofoko** 9/- study of syntax  
 🗨️ **thutapopontšu** 9/- study of morphology  
 🗨️ **thutatlhalošo** 9/- study of semantics  
 📖 **tirišano** 9/10 cooperation  
 🗨️ **tiro** 9/10 predicate, action, process  
 🗨️ **tirotona** 9/10 main predicate / action / process  
 📖 **tirwa** 9/10 passive  
 🗨️ **tirwana** 9/10 subordinate predicate  
 🗨️ **tlaleletša** *v.* determine, qualify  
 🗨️ **tlaleletšadiri** 9/10 verbal determiner  
 🗨️ **tlaleletšaina** 9/10 nominal determiner  
 📖 **tlaleletšatiro** 9/10 verbal adjunct ⇨ **tlaleletšo ya tiro**  
 🗨️ **tlaleletšo** 9/10 complement, adjunct  
 🗨️ **tlaleletšo ya tiro** 9/10 verbal adjunct ⇨ **tlaleletšatiro**  
 📖 **tlami** 9/10 hyphen  
 🗨️ **tlamo** 9/10 connection  
 🗨️ **tlemagano** 9/10 cohesion ⇨ **togaganyo**  
 📖 **tlemagantšha** *v.* link, connect  
 📖 **tlemaganya** *v.* link, connect  
 📖 **tlhakakgolo** 9/10 capital letter  
 📖 **tlhalošišo** 9/10 definition  
 🗨️ **tlhalošo** 9/10 meaning  
 📖 **tlhalošo ya deiktiki** 9/10 deictic meaning  
 📖 **tlhalošo ya lefoko** 9/10 sentence

- meaning
- 📖 **tlhalošo ya lentšu** 9/10 word meaning
  - 🗨️ **tlhalošo ya medirišo** 9/10 modal meaning
  - 🗨️ **tlhalošofoko** 9/10 sentence meaning
  - 🗨️ **tlhalošokamanyi** 9/10 associative meaning
  - 🗨️ **tlhalošokatološo** 9/10 extended meaning
  - 📖 **tlhalošokelello** 9/10 cognitive meaning
  - 📖 **tlhalošokhuduego** 9/10 emotive meaning
  - 🗨️ **tlhalošontši** 9/- polysemy
  - 📖 **tlhalošotebanyo** 9/10 intended meaning
  - 🗨️ **tlhalošotheo** 9/10 basic meaning
  - 📖 **tlhalošotheo ya lefoko** 9/10 basic sentence meaning
  - 📖 **tlhaodi** 9/10 qualificative, modifier
  - 📖 **tlhatlagano** 9/10 hierarchy
  - 📖 **tlhatlamano** 9/10 succession
  - 📖 **tlhatlamano ya ditiro** 9/10 consecutive actions
  - 🗨️ **tlhatlamano ya mantšu** 9/10 word order ⇒ **tatelano ya mantšu, tatanontšu, tlhatlamanontšu**
  - 🗨️ **tlhatlamanontšu** 9/10 word order ⇒ **tatelano ya mantšu, tatanontšu, tlhatlamano ya mantšu**
  - 🗨️ **tlhatlamanotheo** 9/10 basic (word) order
  - 📖 **tlhatlamanotheo ya mantšu** 9/10 basic / dominant word order
  - 📖 **tlhopho** 9/10 categorization
  - 🗨️ **tlhopollo** 9/10 analysis
  - 🗨️ **tlogelo** 9/10 deletion
  - 🗨️ **togaganyo** 9/10 cohesion ⇒ **tlemagano**
  - 📖 **tsebo** 9/10 knowledge
  - 🗨️ **tsebo ye e feleletšego** 9/10 full knowledge
  - 🗨️ **tšhalafatšo** 9/10 pronominalisation
  - 🗨️ **tshedimošo** 9/10 information
  - 🗨️ **tšhomišo** 9/10 function
  - 🗨️ **tšhomišo ya polelo** 9/10 function of language, language function
  - 🗨️ **tšhupetšogotee** 9/- coreference
  - 🗨️ **tšhupetšokarolo** 9/- part-whole relationship, interreference
  - 📖 **tšhutišo** 9/10 shifting
  - 📖 **tshwantšhišo** 9/10 simile
  - 🗨️ **tswalana** *intransitive v.* relate, associate, link
  - 🗨️ **tswalano** 9/10 relationship, association
  - 🗨️ **tswalanya** *transitive v.* relate, associate, link
  - 🗨️ **tswalanyo** 9/10 association, link, connection
  - 🗨️ **tswalanyo ya mantšu** 9/10 association / linking / connection of words
  - 📖 **tumagwaša** 9/10 fricative
  - 📖 **tumammogo** 9/10 consonant
  - 📖 **tumammogokodu** 9/10 voiced consonant
  - 🗨️ **tumanoši** 9/10 vowel
  - 📖 **tumanoši ya mafelelo** 9/10 final vowel
  - 📖 **tumanoši ya mafelelo ya lediri** 9/10 verbal ending (lit. final vowel of the verb)
  - 📖 **tumathu** 9/10 plosive consonant
  - 📖 **tumatshwano** 9/10 homonym
  - 🗨️ **tumelo** 9/10 affirmative
  - 📖 **tummogotu** 9/10 voiceless consonant
  - 🗨️ **tumotshwano** 9/10 homonymy
- V**
- 📖 **VP** verb phrase