Semi-Automatic Retrieval of Definitional Information: A Northern Sotho Case Study[*]

Elsabé Taljard, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (etaljard@postino.up.ac.za)*

Abstract: Corpus-based terminology is currently gaining ground on the international front. It is therefore important that terminologists working on the South African Bantu languages not only take note of this development, but that they should also follow this trend, even if they do not have the same measure of access to highly sophisticated software. The aim of this article is therefore to establish whether it is possible to retrieve definitional information on key concepts from untagged, running text by making use of affordable and easily accessible software such as *WordSmith Tools*. In order to answer this question, a case study is done in Northern Sotho, using textual material on linguistics as basis for a special field corpus. Syntactic and lexical patterns serving as textual markers of definitional information are identified and the success rate of the computational retrieval of definitional information is analysed and evaluated. Attention is also paid to the retrieval of specifically conceptual information, which turned out to be a fortunate by-product of semi-automatic retrieval of definitional information. Finally, it is illustrated how definitional information retrieved can be utilised in the writing of a formal terminological definition.

Keywords: TERMINOLOGY, SOUTH AFRICAN BANTU LANGUAGES, DEFINITIONAL INFORMATION, SEMI-AUTOMATIC INFORMATION RETRIEVAL, TERMINOLOGICAL DEFINITIONS, CONCEPTUAL RELATIONSHIPS, LEXICAL PATTERNS, SYNTACTIC PATTERNS, TEXTUAL MARKERS, KEYWORD-IN-CONTEXT (KWIC), WORDSMITH TOOLS

Opsomming: Semi-outomatiese herwinning van definisie-inligting: 'n Noord-Sothogevallestudie. Korpus-gebaseerde terminologie is tans besig om veld te wen op die internasionale front. Dit is daarom belangrik dat terminoloë wat binne die Suid-Afrikaanse Bantoetale werk, nie net sal kennis neem van hierdie ontwikkeling nie, maar dat hulle ook hierdie neiging sal volg, selfs al het hulle nie dieselfde mate van toegang tot gesofistikeerde rekenaarprogrammatuur nie. Die doel van hierdie artikel is daarom om vas te stel of dit moontlik is om definisie-inligting oor sleutelkonsepte uit ongemerkte, lopende teks te herwin deur bekostigbare en toeganklike sagteware soos *WordSmith Tools* te gebruik. Ten einde hierdie vraag te beantwoord, is 'n gevallestudie in Noord-Sotho gedoen, met gebruikmaking van teksmateriaal oor die linguistiek as basis vir 'n gespesialiseerde korpus. Sintaktiese en leksikale patrone wat as tekstuele merkers van defini-

[*] An earlier version of this article was presented as a paper at the Eighth International Conference of the African Association for Lexicography, organised by the Department of Germanic and Romance Languages, University of Namibia, Windhoek, Namibia, 7–9 July 2003.

sie-inligting dien, word geïdentifiseer en die suksesratio van rekenaarmatige herwinning van definisie-inligting word ontleed en beoordeel. Aandag word ook gegee aan die herwinning van spesifiek konseptuele inligting, wat 'n onverwagse byproduk van die semi-outomatiese herwinning van definisie-inligting is. Ten slotte word geïllustreer hoe definisie-inligting aangewend kan word by die skryf van 'n formele terminologiese definisie.

Sleutelwoorde: TERMINOLOGIE, SUID-AFRIKAANSE BANTOETALE, DEFINISIE-INLIGTING, SEMI-OUTOMATIESE INLIGTINGSHERWINNING, TERMINOLOGIESE DEFINISIES, KONSEPTUELE VERHOUDINGE, LEKSIKALE PATRONE, SINTAKTIESE PATRONE, TEKSTUELE MERKERS, KEYWORD-IN-CONTEXT (KWIC), WORDSMITH TOOLS

1. Electronic corpora and terminology — an overview of the current international and national scenario

The use of electronic or machine-readable corpora in general lexicography is a well-established practice, not only on the international front, but also within the South African set-up, where the nine National Lexicography Units for the South African Bantu languages[1] are all to a greater or lesser extent using a corpus-based approach for the compilation of their various dictionaries. According to Ahmad and Rogers (2001: 728), however, the use of corpora for terminological or LSP purposes has been accepted much more slowly, with the use of corpora largely being restricted to general lexicography, one reason being the more prescriptive orientation of terminology, and the other its onomasiological orientation, i.e. concept-based approach. With reference to the international scene, Pearson (1998: 1) identifies three possible reasons for the seeming reluctance of terminologists and terminographers to recognize the vital role that electronic corpora can and should play in terminology work.

In the first instance she points out that in the past it was indeed difficult to get hold of specialized corpora needed for terminological purposes. However, the increasing availability of especially electronic texts has now made it possible for terminologists to build their own specialized corpora with relatively little effort. The unavailability of suitable textual material that can be used in the compilation of electronic corpora is indeed also relevant for the South African context, and specifically for the South African Bantu languages. Due to the political and educational dispensation of the past two decades, very little has been produced in the South African Bantu languages in the line of special field texts. Subject material is written largely in English and/or Afrikaans, thus denying terminologists the opportunity to base their terminological work on authentic special field texts. Fortunately, during recent years, more and more technical material written in the South African Bantu languages has become available, mainly on the Internet, often as translations of mostly English source texts. Furthermore, although the system of 'Bantu education' as implemented by the previous regime surely has no redeeming qualities, one of the by-products of this system was indeed the publication of school textbooks on special subject fields in at least some of the Bantu languages. Due to the fact that the medium of instruction in former black schools is mostly English, these textbooks are no longer used, but can with a little effort be sourced from school storerooms and archives of libraries. Despite their political baggage, these textbooks can be of great value for the building of special field corpora. South African terminologists, especially those working in the Bantu languages, should therefore be encouraged to make use of this material to compile their own specialized corpora.

A second reason why corpora have not been utilised for specialized lexicography and/or terminology is the perception that terms are context independent — a notion that has dominated terminological work for quite some time. This issue is raised by both Pearson (1998: 1) and Sager (2001: 761), who point out that it is only recently that the emphasis has shifted to studying terms in their communicative, i.e. linguistic context. Within the 'traditional' approach, terms are regarded as separate items, "forming part of a semi-artificial language, deliberately devoid of any of the functions of other lexical items" (Sager 2001: 761). Within the modern, corpus-based approach, terms are viewed as lexical items accorded term status based on the communicative setting in which they appear. Real texts are therefore now regarded as primary sources of terminological data, providing information on the meaning, usage and appropriateness of a term, as conditioned by the textual environment in which it appears. Moreover, as Shreve (2001: 773) indicates, when terms co-occur in a text, they establish conceptual relationships. Thus, by analysing the text, important information on the larger conceptual structure of a specific knowledge domain can be retrieved. In South Africa, there still seems to be some reluctance on the part of terminologists to rely on authentic text for terminological data. Currently, terminological work done by terminologists working in the Bantu languages is mainly of a translational nature, consisting of the finding of translation equivalents for English/Afrikaans terms, based mainly on mother-tongue intuition and consultation with subject-field specialists. Even so, this should not exclude the possibility of utilising corpora — as Bowker and Pearson (2002: 14, 20) point out, corpora are useful complements to all other types of resources and should not be viewed as replacements for these. Making use of especially comparable and parallel corpora within this particular terminological environment could contribute much to the quality of the terminological effort.

In the third instance, it is generally accepted that the input of special field experts is indispensable in the identification and definition of terms. The basic premise here is the "conviction that terms are different from words and can only be defined by suitably qualified subject specialists" (Pearson 1998: 1). 'Traditional' terminologists therefore rely to a large extent on subject experts for the identification and definition of terms. Again, having access to electronic corpora does not imply disregarding the input of experts — the information gleaned from corpora can be used to supplement, support and validate the judgement made by experts, which in practice often represents one person's opinion. Within the South African situation, consultation with experts is still the preferred methodology for terminology work. Unfortunately, there seems to be a lack of commitment on the part of special field experts who are mother-tongue speakers of the South African Bantu languages to develop terminology in these languages, the basic assumption being that English is the language of special subject fields. Furthermore, ongoing consultation with subject-field specialists throughout the terminological process is a time-consuming and labour-intensive exercise, and thus makes the seeming reluctance of subject-field specialists to co-operate understandable, although not defendable.

If South African terminologists, especially those specializing in the Bantu languages wish to keep abreast of terminological developments on the international scene, corpus-based terminology is no longer an option, but an imperative. As was pointed out above, the increasing availability of subject-field texts written in these languages — many of them in electronic format — now enables terminologists to build their own corpora for special purposes. Furthermore, access to user-friendly and affordable software such as *WordSmith Tools* opens the door for terminologists to query and analyse these corpora automatically or at least semi-automatically[2]. It has already been illustrated by Taljard and De Schryver (2002) that it is indeed possible to extract terms semi-automatically from corpora based on subject-field texts, thus reducing (but of course not eliminating) the dependence of the terminologist on the co-operation of the subject-field specialist.

2. Rationale

In this study, the feasibility of retrieving definitional information semi-automatically from special field corpora is investigated. For the purpose of this study, the term 'definitional information' is used to refer to any information to be found in an electronic special field corpus regarding the meaning and usage of a term, as well as the conceptual relationships it has with other terms. In this regard, two issues will be addressed. In the first instance, Pearson (1998: 5) states that authors writing within certain specified communicative settings are likely to provide explanations of at least some of the terms they use. This hypothesis is tested with regard to Northern Sotho, using a special purpose corpus consisting of a collection of texts on linguistics as authentic data. Secondly, the possibility of retrieving definitional information in a semi-automatic way from the corpus is investigated. By way of conclusion, an example will be given as to how the definitional information retrieved from the corpus can be used for the writing of a terminological definition. Before these issues are addressed, however, current methodological options open to South African Bantu language terminologists with regard to the generating of definitional information are investigated.

3. Generating definitional information: the current South African scenario

It has already been pointed out that it is indeed possible for South African terminologists to compile their own special field corpora, and, by following the methodology suggested by Taljard and De Schryver (2002), to semi-automatically extract terms from electronic texts. Should the terminologist now want to add definitions to the extracted terms, he/she is currently left with two options: (a) to formulate definitions with the help of a subject-field expert, or (b) to provide translational equivalents for the terms in English/Afrikaans, then search for definitions in either an LSP dictionary or existing term lists, and as a last step, translate the definitions from English/Afrikaans into the appropriate Bantu language. Both these options have inherent pitfalls that often impact negatively on the quality of the terminological activity, and therefore warrant some discussion.

A number of potential problems regarding the process of consultation with experts have already been touched upon: it is time-consuming, it is labour-intensive and the results often represent the view of only one person. The ideal is of course that multiple experts be involved, but this is in practice not always a realistic aim. Furthermore, terminologists need a very specific skill in order to elicit the correct and appropriate information from subject-field specialists. As Bowker and Pearson (2002: 17) point out, terminologists must be careful not to ask leading questions, which would result in obtaining distorted information. Furthermore, not being a subject-field expert, the terminologist is often not equipped to distinguish between relevant and irrelevant information, or to judge whether the proffered information is merely a personal opinion expressed by the expert. It is therefore clear that terminologists should receive at least some measure of training in the conducting of interviews with subject-field specialists. With terminology training being in its infancy in South Africa, it cannot be taken for granted that terminologists do indeed get this kind of training.

The second option is also not without its disadvantages. According to Bowker and Pearson (2002: 15), one of the inherent problems with LSP dictionaries is their incompleteness. Printed dictionaries tend to become out-dated rather quickly, especially in the fields of science and technology, which are characterised by rapid development. Consulting an LSP dictionary therefore does not guarantee retrieval of the current state of knowledge in a particular subject field. Furthermore, LSP dictionaries often do not provide adequate contextual or usage information — another point raised by Bowker and Pearson (2002: 16). Even if the terminologist succeeds in finding definitions in existing LSP dictionaries, there is always the danger of incompatibility of the target group served by the dictionary and the target group which the terminologist has in mind. A definition sourced from an LSP dictionary with postgraduate chemistry students and chemistry experts as target users will for example not be suitable if the target audience for whom the terminologist is compiling a terminological reference source is senior secondary school learners. A mechanistic translation of terminological definitions would in this instance serve no purpose. Translation of definitions is another area fraught with problems. Technical translation requires a high level of translation skills. It is a well-known fact that there is a shortage of well-trained and qualified translators for the South African Bantu languages, a shortage that is even more serious when it comes to translators specializing in technical translation. Making use of a translator who does not have the necessary background in technical translation will undoubtedly result in terminological definitions of poor quality.

Utilizing an electronic special field corpus as a source of definitional information is not the solution to all of the problems identified above, but it does have distinct advantages. In the first instance, being in electronic format corpora can be updated much more easily than for example, a paper LSP dictionary. By querying a corpus that is regularly updated, the terminologist is more likely to find the most recent information on a particular term. Secondly, the corpus provides the terminologist with a wealth of usage and textual information. Should the terminologist wish to include examples of use as part of the definition of a term, authentic examples can be sourced directly from the corpus. Thirdly, the terminologist can obtain a higher degree of compatibility between the target user of the textual material incorporated in the electronic corpus and the target reader of the final terminological product. This can be done by making sure that the texts that are selected for inclusion in the special field corpus have the same target user as the terminological end product. Lastly, since the textual material used for the compilation of the corpus is already in the target language, the input of a technical translator is no longer necessary. Also, less time needs to be spent on consultation with special field experts, since they are only required to verify, supplement or reject the information gleaned from the corpus.

4. Semi-automatic retrieval of definitional information — a case study

4.1 Compilation of an electronic special field corpus

For the purpose of this investigation, a special field corpus was compiled, based on a number of texts on Northern Sotho linguistics, which were kindly provided in electronic format by Prof. L.J. Louwrens of the Department of African Languages of the University of South Africa (UNISA). These texts were reverted to text-only format in order to make it compatible with the software *WordSmith Tools* (WST), which was to be used as a corpus-querying tool. After a simple count by making use of WST's *WordList* function, it was revealed that this special field corpus contains 74 251 tokens (running words) and 4 744 types (unique words). It has to be borne in mind that this particular corpus is unmarked and untagged, as are all the corpora which are currently available for Northern Sotho. Even so, making use of raw corpora, i.e. running texts, does have certain advantages, as has been indicated by Sager (2001: 764, 765). In the first instance, he points out that "terminology extracted from running text or discourse offers a greater guarantee of thematic completeness and coherence". The linguistic behaviour of terms can be deduced from suitably selected contexts. Secondly, terminology is dynamic, in that new terms are continuously coined and added to the body of knowledge of a particular subject field, sometimes even replacing existing terms. Terms also become obsolete. Running text therefore gives a good indication of the actual existence and currency of terms. Thirdly, running text can assist the terminologist in checking the correctness of previously entered terminology, and to update the terminological database if necessary. However, even though utilising running text is a fruitful exercise, the ideal should always be to work towards establishing a more sophisticated tool in the form of a marked and tagged corpus.

4.2 Identification of 50 single word test terms

As a next step, the *KeyWord* tool was used to identify the 50 raw, i.e. unlemmatised single word terms which have the highest frequency in the corpus. This is done by comparing the frequency of each item in the *WordList* of the special purpose corpus with its frequency in a second, much larger reference corpus. As a reference corpus, the *Pretoria Sepedi Corpus* (PSC) of the University of Pretoria was used. The PSC is an organic corpus and the version used for this study contains roughly 5.9 million words. All items that display a great disparity in frequency are identified as keywords, since the disparity would imply that that specific item occurs with unusual frequency in the smaller corpus. This whole process of keyword identification is done automatically; the only required human intervention being to read through the suggested keyword list and to decide on term status[3]. The 50 test terms identified in this manner are listed in (1).

(1) 50 unlemmatised test terms, extracted semi-automatically from the special field corpus

N	Term	Translation	N	Term	Translation
1	baboledišani	*interlocutors*	26	mantšu	*(linguistic) words*
2	deiktiki	*deictic*	27	mašala	*pronouns*
3	dikafoko	*phrases*	28	medirišo	*moods*
4	direwa	*topics*	29	mmoledišwa	*addressee, 2nd pers.*
5	ditlaleletšo	*complements*	30	modirišo	*mood*
6	kamano	(*inter*)*relationship*	31	modirišogore	*subjunctive mood*
7	kgatelelo	*emphasis*	32	modirišokanegelo	*consecutive mood*
8	kgokagano	*discourse, communication*	33	modirišopego	*indicative mood*
9	lebopikganetši	*negative morpheme*	34	modirišopegotlhaodi	*situative mood*

10	lediri	*verb*	35	modirišotaelo	*imperative mood*
11	lefoko	*sentence*	36	modiro	*function*
12	legoro	*(noun) class*	37	phetlekokgokagano	*discourse analysis*
13	lehlathi	*adverb*	38	poledišano	*dialogue*
14	leina	*noun*	39	sediri	*subject*
15	leinataodi	*head noun*	40	sedirwa	*object*
16	lekgokasediri	*subject concord*	41	seemotikologo	*discourse context*
17	lekgokedi	*agreement morpheme*	42	serewa	*topic*
18	lereo	*term*	43	tatelanontšu	*word order*
19	lešala	*pronoun*	44	tiro	*predicate*
20	lethuši	*auxiliary verb*	45	tlhalošo	*meaning*
21	mabopi	*morphemes*	46	tlhalošotheo	*basic meaning*
22	madiri	*verbs*	47	togaganyo	*cohesion*
23	mafoko	*sentences*	48	tšhupetšogotee	*coreference*
24	mafokofokwana	*complex sentences*	49	tswalane	*being related to*
25	mafokotheo	*basic sentences*	50	tswalano	*relationship*

4.3 Semi-automatic retrieval of definitional information — a case study

4.3.1 Isolating Concordance lines for the 50 test terms

As was stated earlier, the first aim of this investigation is to establish whether definitional information is indeed provided in the text. For this purpose, the *Concord* tool of WST is used. When the selected term is entered as a search node, a list of concordance lines or KWIC *(keywords-in-context)* lines in which that specific term appears, is automatically thrown up, thus placing the term within its textual context. Compare the example in (2) where the term **lediri** 'verb' has been used as a search node.

(2) Concordance lines (KWIC lines) for the term **lediri** 'verb'

Leina le o tšogo le ngwalolla le tlaleletša	**lediri**	a rwala ka tsela efe? Ke tlaleletšo
mantšu a mabedi, e lego lethuši o hlwa le	**lediri**	a bala. Mehlala ye mengwe ke
a tlaleletšago modiro wo o hlalošwago ke	**lediri**	a bitšwa didirwa, gomme mo
la Legoro 1 -thuš-: modu wa	**lediri**	-a : mosela woo lediri le felelago
a mararo, e lego (i) sediri baithuti, (ii)	**lediri**	ba rekile; le (iii) sedirwa dipuku
o ngwalolle lediri. Ge tiro e le lethuši le	**lediri**	go nyakega gore o ngwalolle
Le gona re boletše gore	**lediri**	le ka tlaleletšwa ka ditlaleletši tša

The number of concordance lines differs from one term to the next, depending on how many times the specific term appears in the text. By studying the KWIC lines thrown up for every term, one can then establish whether definitional information is provided within the text for that specific term. For these 50 test terms, the total number of KWIC lines that had to be perused, is 4 246. Definitional information was found in 292 of these KWIC lines, covering 45 of the selected 50 terms. These findings therefore support Pearson's hypothesis that definitional information is provided by writers of technical texts. The authors of these specific technical texts do indeed, whether advertently or inadvertently, provide definitional information on the terms they use.

With regard to the terms on which no information was found, a few remarks can be made. The first term on which no information was found, is **baboledišani**. Although this term has the specialised meaning of 'interlocutors' within the linguistic context, its meaning is to a large extent self-evident, possibly because of its morphological structure. It is a deverbative noun, derived from the verbal root **bolel-** 'speak', affixed with a causative suffix **-iš-** and a reciprocal **-an-**. The prefix **ba-** of course indicates that the referents of this term are human beings. Due to its morphological make-up, it is therefore quite easy to derive the meaning of this term and explains why the author of this particular text did not deem it necessary to provide definitional information on the term. The next two terms on which no information was found, are **modiro** 'function' and **lereo** 'term'. These lexical items belong to a category which is known as non subject-specific-specialised vocabulary or subtechnical terms. Pearson (1998: 19) describes subtechnical words as general language words that have taken on specialised meanings in more than one domain, and due to their concomitant high frequency of use could be assumed as being known. With regard to the term **mabopi** 'morphemes' for which no information could be found, it could be argued that this is such a basic concept within the field of linguistics, that it is not surprising that no definitional information was provided, especially when it is borne in mind that the texts used in this particular study target senior university students. The last term for which no definitional information could be retrieved, is **tswalane** 'being related to'. This was the only verb that was found amongst the 50 test terms. In the study done by Taljard and De Schryver (2002), a total number of 350 terms were extracted from the same texts on which the current study is based. Of these, only 14, i.e. 4% were verbs, of which **tswalane** was one. A concordance search on the other 13 verbs revealed that definitional information is not provided for any of these terms. It therefore seems that retrieving definitional information on verbal terms does pose a problem.

4.3.2 Identification of textual markers of definitional information

Studying KWIC lines in search of definitional information obviously represents an improvement on the more traditional method of physically reading through all the textual material in search of definitional information. Ahmad and Rogers (2001: 740) point out that reading a text to extract any kind of terminological data is labour-intensive and potentially repetitive, because of the need to recover different kinds of terminological data. They furthermore indicate that the computational processing of relatively large quantities of text may allow patterns to emerge which are unlikely to be detected by manual scanning. The reference to the possible existence of patterns immediately suggests the possibility of semi-automatic recognition of definitional information. Pearson (1998: 103) indicates that definitional information can be signalled either by syntactic or lexical devices. It therefore follows that if the strategies or patterns which are used to signal definitional information in texts can be identified, it would enable the terminologist to semi-automatically retrieve at least some definitional information from these texts. It should be kept in mind that the corpus utilized for the purpose of this case study is an untagged one. The initial aim was simply to identify lexical markers of definitional information, similar to the way in which Pearson (1998: 136 et seq.) identifies fillers for the slots **X**, **Y** and **=**, in what she terms a formal terminological definition. She identifies two variant formulae that characterise a formal definition:

(3)
Formula 1: **X = Y + distinguishing characteristic, whereby X is subordinate to Y**
Example: an adverb (X) is (=) a linguistic word (Y) which qualifies a verb (distinguishing characteristic). ['Adverb' being a subordinate to the superordinate 'linguistic word'.]

Formula 2: **Y + distinguishing feature = X, whereby X is subordinate to Y**
Example: a linguistic word (**Y**) which qualifies a verb (**distinguishing characteristic**) is (called) (=) an adverb **(X)**. ['Adverb' being a subordinate to the superordinate 'linguistic word'.]

In these formulae, **X** is obviously the term being identified, **Y** the superordinate, and **=** the linguistic device which links the term to its superordinate. She then proceeds by identifying, for English, typical lexical fillers for the slots **Y** and **=**. She indicates that **Y** must either refer to a term (which is the case in the illustrative example above) or to a class word. Typical class words which are to be found in this slot are: 'technique, method, process, function, property, system, class, device', whereas the **=** slot is filled by connectives such as 'comprise(s), consist(s) of, define(s), denote(s), is/are, is/are called, is/are known as', etc. In her study, Pearson focuses mainly on definitional information that is provided in the format of these formal definitions. However, with regard to the Northern Sotho case study, the objective is to retrieve as much definitional information as possible on each term, and not only information provided in the form of a formal definition. Furthermore, although the initial aim was to identify lexical markers that signal definitional information, it quickly became clear that even in an untagged corpus, certain syntactic patterns can be identified as strategies by means of which definitional information is marked. Both these strategies are of value to the terminologist, the main advantage of lexical markers being that definitional information signalled by these items, can be retrieved by using a basic word-processing tool such as *MSWord* or *WordPerfect*, simply by making use of the *Search* function. This makes the manual scanning of all textual material unnecessary, since the terminologist can automatically pinpoint all occurrences of the particular term in the text where definitional information is provided on it. Syntactic patterns that signal definitional information, on the other hand, will be of much value when hopefully in the near future, Northern Sotho will have a tagged corpus, enabling the terminologist to automatically search for these patterns.

4.3.3 Lexical and syntactic markers of definitional information in Northern Sotho

The lexical items as well as the syntactic patterns found as markers of definitional information in the special purpose corpus on Northern Sotho linguistics are given in (4):

(4) Table 1 (See attached spreadsheet)

The procedure followed to arrive at these results can be briefly explained: the 292 KWIC lines in which definitional information was found, were scrutinised in order to identify any textual markers which signalled the presence of definitional information on any of the 45 terms. These markers of definitional information function on a very non-theoretical level. They simply indicate that some form of definitional information on a specific term is to be found in the textual vicinity flagged by these markers. The definitional information can either precede or follow the term and its marker, and may even be spread across sentence boundaries. After the markers had been identified, these were then indicated on a spreadsheet by means of a cross next to the term on which it provided information. From the information provided under B (leftmost column), for example, it is clear that for 22 of the 45 terms, definitional information is textually marked by the presence of the identifying copulative particle **ke**, which follows the term (**T**) in question. Markers that appeared only once were not taken into account. Analysis of the marking strategies brought to light that two distinct kinds of strategies could be identified, i.e. general strategies and strategies which are suspected to be subject-specific. The strategies in columns B, C, D, E, G, H, J and K, as well as those in the unshaded columns under F, represent general strategies. This means that regardless of the subject field, these strategies would in all probability function as markers of definitional information in Northern Sotho texts. The strategies listed in the shaded columns under F and the one in column I are of a slightly different nature. Their function as marking strategies are probably directly linked to the lexical items that are found in these patterns and the relationship of these items with the particular subject field. Verbs such as **ganetšwa** 'is/are negated', **šuthišwa** 'is/are moved', **tlogelwa** 'is/are deleted', etc. form such an inherent part of the subject field of Northern Sotho linguistics, that it is highly improbable that these strategies would be useful in identifying definitional information on terms appearing in a corpus based on, for example, chemistry or music texts.

TABLE 1

A	B				C			D	E	F								G		H		I			J		K			
Strategies →	T + Identifying cop				Identifying cop + T			lereo le (la) + T 'this term (of)'	lereo le (la) + T + COP 'this term (of) is'	T (+dem) + s.c.(+ka) + PASS V								PASS V + T		PASS V + ke + T		T + s.c. + V			inchoative V + T		Orthographic strategies			
Lexical markers →	ke 'is/are'	e le(go) 'which is/are'	e ba 'become(s)'	e ka ba(go) '(which) can be/become'	ke 'is/are'	e le(go) 'which is/are'	e tla/ka ba 'can/will be'			hlolwa 'is/are caused'	šomišwa 'is/are used'	bitšwa 'is/are called'	ganetšwa 'is/are negated'	šuthišwa 'is/are moved'	tlaleletšwa 'is/are qualified'	tlogelwa 'is/are deleted'	tswalanywa 'is/are connected'	bitšwa 'is/are called'	bopša 'is/are formed'	hlolwa 'is/are caused'	laolwa 'is/are determined'	hlaloša 'mean(s)'	tšweletša 'produce(s), result(s) in'	-na 'has, have'	hlola 'cause(s)'	bopa 'form(s)'	(T)	T(*)	T:	
Terms																														
deiktiki	x																													
dikafoko		x √																	x						x √	x x √				
direwa		x																												
ditlaleletšo	x																													
kamano										x x x √								x x		x										
kgatelelo																									x x x x √					
kgokagano	x x					x √				x √															x x √					
lebopikganetši					x x																									
lediri	x				x	x x √					x				x x x x x x															
lefoko	x x x x √			x																			x		x x √					
legoro																					x	x								
lehlathi	x √	x					x											x √												
leina					x x x √		x														x									
leinataodi																		x x √			x									
lekgokasediri	x √									x								x x x √												
lekgokedi					x							x x x								x x x √										
lešala											x	x x																		
lethuši						x x √																								
madiri																										x				
mafoko		x																							x x √	x x √				
mafokofokwana					x √														x x √							x x		x		
mafokotheo	x √																x													
mantšu	x	x √				x √													x √					x x		x x x x √				
mašala	x √	x									x																			
medirišo						x																								
mmoledišwa	x x √																											x x		
modirišo	x x x x x x																													
mmodirišogore	x					x x		x x x x		x	x											x x							x	
modirošokanegelo								x					x										x							
modirišopego	x x √				x x																			x x						
modirišopegotlhaodi	x							x					x																	
modirišotaelo	x √											x																		
phetlekokgokagano	x																											x		
poledišano																												x		
sediri		x			x x			x x						x		x	x												x	
sedirwa	x			x										x		x													x	
seemotikologo		x		x x				x x x x x x													x x			x						
serewa	x x x x √		x √		x																						x			
tatelanontšu																											x			
tiro	x x √	x x x x x √		x x x x x √					x x x x						x															
tlhalošo	x																													
tlhalošotheo	x x x x √	x √																												
togaganyo						x		x	x	x x															x x			x		
tšhupetšogotee								x												x										
tswalano					x															x x √								x		
	40	14	1	9	14	9	2	16	5	8	4	6	2	2	7	2	2	8	4	7	5	3	2	5	13	11	2	7	3	
Total no. of occurrences	General strategies			64			26	16	5			18							12		12					24			12	189
	General and specific strategies			64			26	16	5			18					15		12		12			10		24			12	214
No. of new terms on which definitional info is provided				27			8	2	0			2					0		1		1			0		2			2	45

This implies that the strategies **T (+dem) + s.c. (+ka) + PASS V** and **T + s.c. + V** would probably function as markers of definitional information in any text, but containing in the VERB (**V**) slot, verbs which are relevant to the particular subject field. This would imply that these verbs would have to be identified anew for every subject field within which the terminological activity takes place.

4.3.4 Analysis of results

With regard to the statistical analysis of the table, two sets of figures are relevant. Firstly, at the bottom of each column, the total number of occurrences for each strategy is given, thus the strategy **Term + Identifying copula** in column B appears in a total of 64 out of a possible 292 KWIC lines. The total number of KWIC lines in which the general strategies appear, is 189. This implies that by using the patterns appearing in these general strategies as a search node in WST's *Concord* tool, roughly 65% of all concordance lines in which definitional information appears, can be retrieved automatically from the corpus. To illustrate this: by using the pattern **Term + ke** as a search node, a total of 40 concordance lines, providing definitional information on 22 different terms, can be culled from the corpus. Compare example (5) in this regard, where an excerpt of 10 concordance lines is given by way of illustration. Translations of the relevant concordance lines have been added in order to illustrate the definitional information to be found in these lines.

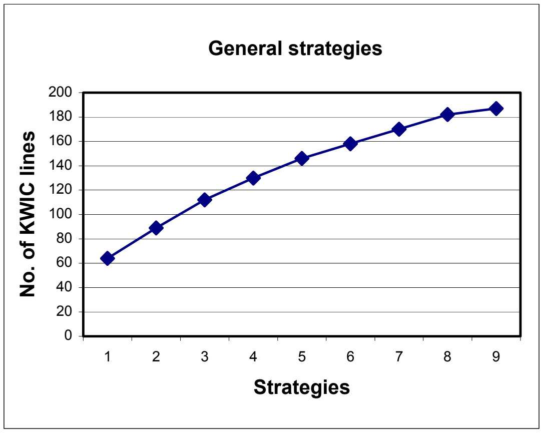
(5) Definitional information marked by the pattern **Term + ke**

<table>
<tr><th></th><th>Term</th><th>ke</th><th></th></tr>
<tr><td>Mantšu a</td><td><b>deiktiki</b></td><td><b>ke</b></td><td>mantšu a a šupago motho yoo goba selo seo se bonalago mo seemotikologong sa kgokagano.</td></tr>
<tr><td colspan="4"><b><i>Deictic</i></b> words <b><i>are</i></b> words that indicate a person or an object which is visible within the discourse context.</td></tr>
<tr><td>mehuta ye e fapanego ya</td><td><b>ditlaleletšo</b></td><td><b>ke</b></td><td>gore mantšu ao a šomišwago mo dikafokong go tlaleletša leina goba lediri</td></tr>
<tr><td colspan="4">… different kinds of <b><i>complements is</i></b> that those words that are used in phrases to complement a noun or a verb …</td></tr>
<tr><td>Malebišatheo a</td><td><b>kgokagano</b></td><td><b>ke</b></td><td>go fihliša molaetša wo o tšwago go mmoledi (goba mongwadi) go mmoledišwa (goba mmadi).</td></tr>
<tr><td colspan="4">The basic function of <b><i>discourse is</i></b> to carry a message coming from the speaker (or writer) to the addressee (or reader).</td></tr>
<tr><td>Re ka re</td><td><b>lediri</b></td><td><b>ke</b></td><td>kokwane ye e thekgilego mantšu a mangwe mo lefokong</td></tr>
<tr><td colspan="4">We can say that a <b><i>verb is</i></b> the base which supports the other words in the sentence …</td></tr>
<tr><td>Elelwa gore</td><td><b>mafokotheo</b></td><td><b>ke</b></td><td>mafoko a a nago le tiro e tee fela</td></tr>
<tr><td colspan="4">Remember that <b><i>basic sentences are</i></b> sentences which have only one predicate …</td></tr>
</table>

Mašala	**Mašala**	**ke**	mantšu a a šomago bjalo ka maina, gomme a ka bewa sebakeng sa maina mo lefokong:
Pronouns are words that function like nouns, and they can be placed in the place of a noun in a sentence:			
kgethollwago mo lefokong.	**Sedirwa**	**ke**	motho yoo goba selo seo se angwago ke modiro wo o dirwago ke sediri.
… distinguished in a sentence. The ***object is*** the person or object which is influenced by the action carried out by the subject.			
Ke ka fao ge re itše ra re	**tiro**	**ke**	motheo wa lefoko, ka ge tlhalošo ya yona e le kokwane ye e thekgilego ditlhalošo tša mantšu a mangwe mo lefokong.
Therefore we say that the ***predicate is*** the foundation of the sentence, since its meaning is the base which supports the meanings of the other words in the sentence.			
yona ga e bonale. Re ka re	**tlhalošo**	**ke**	senaganwa goba boikakanyetšo bja mmoledi le mmoledišwa.
… is not visible. We can say that ***meaning is*** the idea or the intention of the speaker and the addressee.			
le na le tlhalošo ya motheo.	**Tlhalošotheo**	**ke**	ke tlhalošo ye e rwelwego ke lentšu ge le eme le le nnoši.
… has a basic meaning. The ***basic meaning is*** the meaning which is carried by the word when it stands alone.			

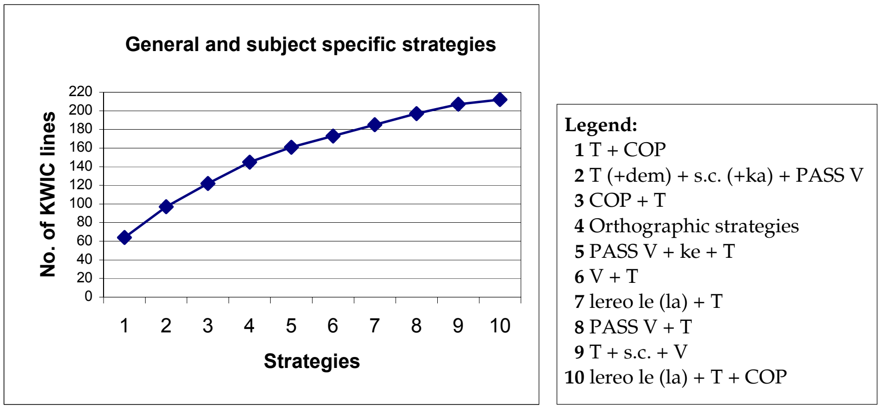
When the occurrence of the subject-specific patterns is added to those of the general strategies, the retrieval rate goes up from 189 concordance lines out of a possible 292, to 212, thus from 65% to 73%. These figures are represented in (6) and (7) respectively:

(6) Retrieval rate of concordance lines using general strategies only



Legend:
1 T + COP
2 COP + **T**
3 Orthographic strategies
4 T (+dem) + s.c.(+ka) + PASS V
5 PASS V + ke + **T**
6 V + **T**
7 lereo le (la) + **T**
8 PASS V + **T**
9 lereo le (la) + **T** + COP

(7) Retrieval rate of concordance lines using both general and subject-specific strategies



Although an increase of 8% seems almost negligible, it was found that the definitional information identified by means of these subject-specific strategies, provided the terminologist with highly relevant information, which would contribute much to the formulation of a good terminological definition of the particular term. Granted, from a statistical point of view, 73% retrieval does not seem to amount to much, but it has to be pointed out that there is a large amount of repetition to be found in the concordance lines, possibly due to the instructional nature of the text. Compare (8) below, in which five different KWIC lines provide definitional information on the term **mantšu** 'linguistic words', but the definitional information provided in these lines, is exactly the same, all of them basically indicating that sounds are combined to form words.

(8) Definitional information on the term **mantšu** 'linguistic words'

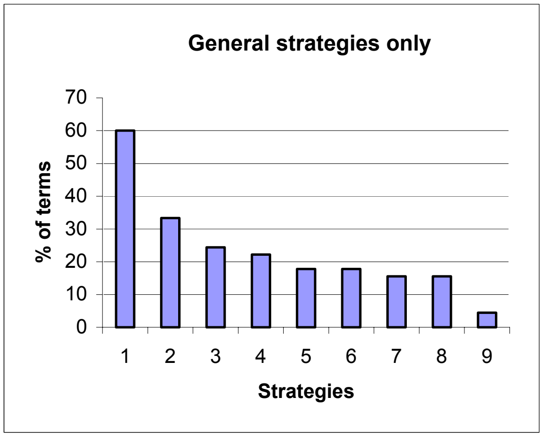
<table>
<tr><th></th><th>Term</th><th></th></tr>
<tr><td>gomme e re lemoša ge motheo wa</td><td><b>mantšu</b></td><td>e le medumo. Gape re lemoga ge mantšu a</td></tr>
<tr><td colspan="3">… and it indicates to us that <b><i>the basis of words are sounds</i></b>. Furthermore we understand if words …</td></tr>
<tr><td>medumo e tswalana go bopa</td><td><b>mantšu</b></td><td>mantšu a tswalana go bopa mafoko</td></tr>
<tr><td colspan="3">… <b><i>sounds are combined to form words</i></b>, words are combined to form sentences …</td></tr>
<tr><td>tswalanya medumo ye e itšego go bopa</td><td><b>mantšu</b></td><td>Ge a re mma, o tswalanya medumo</td></tr>
<tr><td colspan="3">… <b><i>combine(s) certain sounds to form words</i></b>. If he/she says mma, he/she combines sounds …</td></tr>
<tr><td>medumo e šomišwago ka gona go bopa</td><td><b>mantšu</b></td><td>Nepo ye bohlokwa ya fonolotši ke gona go</td></tr>
<tr><td colspan="3">… <b><i>sounds which are combined to form words</i></b>. The important aim of phonology is to …</td></tr>
</table>

Ge medumo e tswalanywa, go bopša	**mantšu**	Lekala la thutapolelo le le nyakišišago

When ***sounds are combined, words are formed***. The section of grammar which investigates …

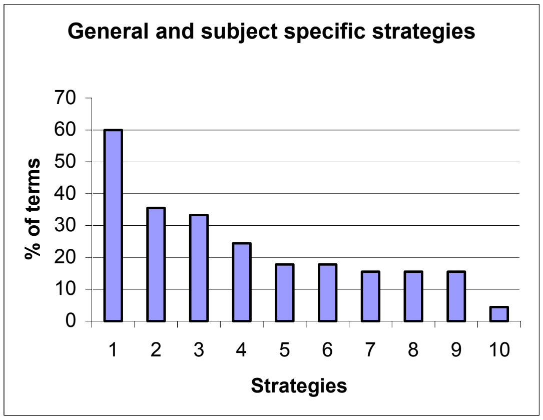
The second set of figures that is relevant to the table, concerns the relationship between a given strategy and the number of terms for which it provides definitional information. In (9), the retrieval rate of definitional information for each of the general strategies is given, whereas (10) reflects the retrieval rate for general and subject-specific strategies.

(9) Retrieval rate: general strategies (%)



Legend:
1 **T** + COP
2 COP + **T**
3 Orthographic strategies
4 **T** (+dem) + s.c. (+ka) + PASS V
5 PASS V + ke + **T**
6 V + **T**
7 lereo le (la) + **T**
8 PASS V + **T**
9 lereo le (la) + **T** + COP

(10) Retrieval rate: general and subject-specific strategies (%)



Legend:
1 T + COP
2 T (+dem) + s.c. (+ka) + PASS V
3 COP + T
4 Orthographic strategies
5 PASS V + ke + T
6 V + T
7 lereo le (la) + T
8 PASS V + T
9 T + s.c. + V
10 lereo le (la) + T + COP

The strategy **Term + Identifying copula** provides information on 27 of the 45 terms; thus by employing this strategy as a search node in a KWIC line, one can retrieve information on 60% of the total number of terms. When the pattern **Identifying copulative + Term** is utilised, definitional information on a further 8 terms can be retrieved, pushing the retrieval rate up to 78%. Since these are both general strategies, one can expect that the use of these patterns as search nodes in WST's *Concord* tool would result in a high retrieval rate of definitional information, regardless of the subject field within which the terminologist may be working. This hypothesis will of course have to be tested by means of further research, making use of special field corpora based on a variety of special subject fields.

5. Retrieval of information on conceptual relationships

It has already been pointed out that an analysis of a special field text not only provides general definitional information, but that it may also reveal the conceptual relationships in which terms appear. Information on these relationships is particularly important when the purpose of the retrieval of definitional information is writing of a terminological definition. An analysis of the definitional information which was retrieved from the corpus brought to light that conceptual information on quite a number of terms had in the process of definitional retrieval, also come to the fore. Although Sager (1990: 29) indicates that the "simplistic view of the past that concepts are adequately represented by three types of relationships (generic, partitive, other) has been generally abandoned", he nevertheless indicates that these relationships are still frequently used in terminology, together with a fourth category, labelled 'complex relationships'. Therefore, the conceptual relationships that were investigated for the purpose of this study are:

— Generic relationships, which deal mainly with the concepts 'superordinate' and 'subordinate' and which establish a hierarchical order of concepts;

— Partitive relationships, also called part-whole relationships, which indicate the connection between concepts consisting of more than one part and their constituent parts;

— Polyvalent relationships, which include polyhierarchical relationships, which occur when a concept is placed within more than one hierarchy within a given subject field; and

— Complex relationships, which include quite a large range of relationships such as causal relations, instrumental relations, production relations, functional relations, etc.

An analysis of the 292 KWIC lines that provided definitional information on the 45 test terms, revealed that the definitional information uncovered for these terms also include information on the conceptual relationships between the particular term and other related terms. For 24 of the 45 terms, i.e. 53%, conceptual information could be retrieved from the text. The terms as well as the strategies that led to the recovery of the conceptual relationships have been marked by means of the symbol √ on the table in (4). A summary of the strategies and the number of terms for which conceptual information is provided by each strategy is given in (10):

(10) Retrieval of conceptual information

Strategy	**No. of terms**
T + COP	13
COP + **T**	6
T (+dem) + s.c. (+ka) + PASS V	2
PASS V + **T**	5
PASS V + ke + **T**	2
Inchoative V + **T**	6

From the above, it is clear that the strategy **T** + COP which is responsible for the highest retrieval rate of definitional information, is also linked to the retrieval of conceptual information for the highest number of terms. Also noticeable is the fact that none of the subject-specific strategies seems to function as a marker of conceptual information. Some examples of conceptual information that were retrieved from the corpus appear in (11)–(14):

(11) Generic relationship

… go kgethollwa mehuta ye mebedi ye bohlokwa ya **kgokagano**, e lego (i) *kgokagano ka molomo*; le (ii) *kgokagano ka go ngwala*.

'… two important kinds of ***discourse*** are distinguished, which are (i) *spoken discourse*; and (ii) *written discourse*.

(**kgokagano** 'discourse' = superordinate; *kgokagano ka molomo* 'spoken discourse' and *kgokagano ka go ngwala* 'written discourse' = subordinates)

Ka fao *lekgokedi* le le bitšwa **lekgokasediri**.

'Therefore this *agreement morpheme* is called a ***subject concord***.'

(**lekgokedi** 'agreement morpheme' = superordinate; **lekgokasediri** 'subject concord' = subordinate)

(12) Partitive relationship

Dikarolo tše pedi tše bohlokwa tšeo di hlolago **lefoko** ke *sekafokoina* le *sekafokodiri*.

'The two important sections that form a ***sentence*** are the *noun phrase* and the *verb phrase*.'

(**lefoko** 'sentence' = whole; *sekafokoina* 'noun phrase' and *sekafokodiri* 'verb phrase' = parts)

Ngwana o thoma go tswalanya *medumo* ye e itšego go bopa **mantšu**.
'A child starts to combine certain *sounds* to form ***words***.'
(**mantšu** 'words' = whole; *medumo* 'sounds' = part)

(13) Polyvalent relationship

… lentšu la mathomo mo lefokong e tla ba **lehlathi**.
'… the first (linguistic) word in the sentence will be the ***adverb***.'

Tlaleletšatiro ya mohuta wo e bitšwa **lehlathi**.
'This kind of verbal adjunct is called an ***adverb***.'
(The term **lehlathi** is situated within two hierarchies: on the one hand it is a (linguistic) word, which — within the structuralist framework — implies that adverbs constitute a word class in Northern Sotho. Secondly, it is indicated that adverbs are also verbal adjuncts.)

(14) Complex relationships

Causal relationship
Togaganyo ka tlhalošo e hlolwa ke **tswalano** ye e bonalago gare ga tlhalošo ya mantšu a a fapanego.
'*Semantic cohesion* is caused/established by the ***relationship*** which exists between the meanings of different words.'
(**tswalano** 'relationship' = cause; *togaganyo (ka tlhalošo)* '(semantic) cohesion' = effect)

Instrumental relationship
Malebiša a … phetlekokgokagano … ke go nyakišiša le go hlatholla ka fao *polelo* e dirišwago ka gona go hlola **kgokagano** gare ga batho.
'The purpose of … discourse analysis … is to investigate and to explain how language is used to establish communication/discourse between people.'
(*polelo* 'language' = instrument; **kgokagano** 'communication' = process)

The same kind of investigation that was carried out in order to identify textual markers of definitional information in corpora would have to be carried out with regard to the identification of markers of conceptual information. It does seem highly unlikely though that strategies for the marking of conceptual information would differ radically from those used to mark definitional information, since conceptual information is nothing but a specific kind of definitional information.

6. Using definitional information retrieved semi-automatically for the writing of a formal definition — a case study

Definitional information retrieved from a corpus can be used either as a rough and ready tool by technical translators to assist them in decoding the text which is to be translated, or it can be utilised for the writing of a formal terminological definition. The actual content of the definition will naturally be determined by the usual factors that should be taken into account when any definition is compiled, i.e. target user, available space (in the case of paper dictionaries), etc. What follows is a brief illustration of how definitional information retrieved from the text can be used in the formulation of a terminological definition. The term selected for illustrative purposes is **lehlathi** 'adverb'.

In (15), the definitional information on **lehlathi** as culled from the text is cited.

(15) Definitional information on the term **lehlathi**

Ka ge boemotheo bja **lehlathi** e le mafelelong a lefoko, le swanetše go šutišwa gore le eme mathomong ge go nyakega gore le tšwelele ka tsela ya kgatelelo.

'Since the basic position of an ***adverb*** is at the end of a sentence, it must be moved to the beginning if it is to be emphasized.'

Lehlathi ke karolofoko yeo e hlathago ka moo modiro o phethagalago ka gona.

'An ***adverb*** is a sentence part which distinguishes/explains the way in which an action is carried out.'

Tlaleletšatiro ya mohuta wo e bitšwa **lehlathi** ka gobane lehlathi le hlatholla modiro wo o bolelwago ke tiro.

'A verbal adjunct of this kind is called an ***adverb*** since an adverb qualifies the action expressed by the predicate.'

Ge a nyaka go gatelela **lehlathi** (ke gore maabane), lentšu la mathomo mo mafokong e tla ba lehlathi.

'If he/she wants to emphasise the ***adverb*** (which is yesterday), the first word in the sentence will be the adverb.'

Based on the information provided in (15), the following terminological definition can be formulated:

(16) Formal definition of **lehlathi**

lehlathi: tlaleletšatiro yeo e hlathollago modiro wo o bolelwago ke tiro. Boemotheo bja lehlathi ke mafelelong a lefoko, eupša ge go nyakega gore le tšwelele ka tsela ya kgatelelo, le swanetše go šutišwa gore le eme mathomong. Mohlala: *maabane*.

'adverb: a verbal adjunct which qualifies the action expressed by the predicate. Its basic syntactic position is at the end of the sentence, but it can be moved to the beginning of the sentence for purposes of emphasis. Example: *maabane* 'yesterday'.'

It must again be stated very clearly that the process of retrieving definitional information semi-automatically does not make the involvement of the subject-field expert redundant. There is no guarantee that this process will for example, succeed in retrieving the most salient features of a particular concept from the text, simply because they are not marked by any of the identified text markers, or because they are not mentioned in the text. A case in point is the definitional information retrieved for the term **leina** 'noun'. The following definitional information was retrieved for this term:

(17) Definitional information on the term **leina**

Sediri mo lefokong e ka ba **leina** go swana le Moithuti o a bala.
'The subject in a sentence may be a ***noun*** as in *The student reads/studies*.'

Mo sekafokoineng, lentšu le bohlokwa ke **leina** mola lentšu le bohlokwa mo sekafokodiring e le lediri.
'In a noun phrase, the important word is the ***noun***, whereas the important word in the verb phrase is the verb.'

baithuti ke **leina** la Legoro 2
'*students* is a ***noun*** of class 2'

dipuku ke **leina** la Legoro 10
'*books* is a ***noun*** of class 10'

Leina le ka hlaolwa ke ditlaleletšo tše di fapanego.
'A ***noun*** can be determined by different adjuncts.'

… re lemoga ge lešala gagwe le laolwa ke **leina** le Makena.
… we understand that the pronoun *gagwe* is determined by the ***noun*** Makena.'

It is only a subject expert who will be able to point out that the most salient feature of nouns in Northern Sotho does not appear amongst the information retrieved, i.e. that nouns are morphologically characterised by a class prefix which is affixed to a stem. It is also not possible to deduce the correct superordinate, i.e. **lentšu** 'linguistic word' from the retrieved information. Although not foolproof, the process of semi-automatic retrieval of definitional information does however reduce the time that has to be spent on consultation with special field experts, which in turn, might make them more willing to participate in terminology projects.

7. Conclusion

The main aim of this article was to investigate whether authors provide definitional information for high frequency terms in technical texts, and if so, whether it would be possible to retrieve this information in a semi-automatic way. To this end, a case study was done for Northern Sotho, using a special field corpus on linguistics. It was found that definitional information on high frequency terms is indeed provided in these texts: definitional information could be identified for 45 of the 50 test terms, i.e. 90%. It was pointed out that semi-automatic retrieval of this information could significantly reduce the dependency of the terminologist on the input of the special field expert. Consequently, a number of lexical and syntactic patterns that function as textual markers of definitional information were identified. When these patterns are entered as search nodes in WST's *Concord* tool, or for that matter in the *Search Box* of an ordinary *MSWord* document, the need to physically read through the textual material in search of definitional information is eliminated. The patterns **T + Identifying copula** and **Identifying copulative + T** distinguished themselves as highly effective markers of definitional information. By making use of these patterns as search nodes, definitional information could be retrieved for a surprising 78% of the test terms.

A close reading of the concordance lines containing definitional information revealed that important information regarding conceptual relationships existing between terms could also be retrieved. This information is particularly relevant in cases where the aim of information retrieval is the writing of a terminological definition.

Finally, it was stated that special field experts would always remain the final arbiters in evaluating the correctness and applicability of definitional information retrieved from texts. The value of the semi-automatic retrieval of definitional information is that the demands on the input of these experts are significantly reduced, since the software can isolate information, which could potentially be used in terminological definitions, leaving the expert the task of evaluating, and if necessary, supplementing the information proffered by the software.

Currently for Northern Sotho, terminologists do not have access to tagged and marked corpora, they do not have morphological parsers enabling them to retrieve all kinds of information from corpora with the proverbial pressing of a button, they do not have sophisticated software programmes, but they do have basic tools. This study hopefully serves to illustrate that by making innovative use of the basic tools that are available, one is indeed able to do terminological work of high quality.

Endnotes

1. In this article, (South African) Bantu languages is used in accordance with its international application, i.e. as a purely linguistic term referrring to a specific family of languages.
2. For more information on *WordSmith Tools*, see the home page of Mike Scott, the creator of the software: <http://www.lexically.net> (or else: <http://www. liv.ac.uk/~ms2928>).
3. For a detailed discussion of this procedure, see Taljard and De Schryver (2002: 51-56).

Bibliography

Ahmad, Khurshid and Margaret Rogers. 2001. Corpus Linguistics and Terminology Extraction. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001: 725-760.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Pearson, Jennifer. 1998. *Terms in Context.* Amsterdam: John Benjamins.

Sager, Juan C. 1990. *A Practical Course in Terminology Processing.* Amsterdam: John Benjamins.

Sager, Juan C. 2001. Terminology Compilation: Consequences and Aspects of Automation. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001: 761-771.

Shreve, Gregory M. 2001. Terminological Aspects of Text Production. Wright, Sue Ellen and Gerhard Budin (Eds.). 2001: 772-787.

Taljard, Elsabé and Gilles-Maurice de Schryver. 2002. Semi-Automatic Term Extraction for the African Languages, with Special Reference to Northern Sotho. *Lexikos* 12: 44-74.

Wright, Sue Ellen and Gerhard Budin (Eds.). 2001. *Handbook of Terminology Management. Volume 2. Application-Oriented Terminology Management*. Amsterdam: John Benjamins.