The Lemmatisation of Adverbs in Northern Sotho*

D.J. Prinsloo, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (prinsloo@postino.up.ac.za)*

Abstract: To date Northern Sotho metalexicographers have focused their attention on lemmatisation problems in respect of the so-called main or primary part of speech categories, viz. nouns and verbs. See, for example, Prinsloo and De Schryver (1999) and Prinsloo and Gouws (1996). No attention has been given to the lemmatisation of *adverbs*. The latter are regarded by Ziervogel and Mokgokong (1975: 114, Introduction) as a "secondary part of speech". The treatment of adverbs in Northern Sotho dictionaries is marred by inconsistencies such as omissions from the macrostructure, insufficient and inconsistent labelling, inferior treatment in the microstructure, under-utilization of the mediostructure and outer texts, and reflects a lack of a strategy of selection of items for lemmatisation. Linguistic descriptions of adverbs in currently available grammars vary substantially and therefore confuse learners of the language and inexperienced lexicographers[1]. The aim of this article is to offer solutions to the lemmatisation problems regarding adverbs in Northern Sotho and to propose guiding entries for paper and electronic dictionaries which could serve as models for future dictionaries. The treatment of adverbs in Northern Sotho dictionaries will also be critically evaluated, especially in terms of frequency of use and target users' needs.

Keywords: LEXICOGRAPHY, LEMMATISATION, ADVERBS, INFORMATION RETRIEVAL, ELECTRONIC DICTIONARY, MACROSTRUCTURE, MICROSTRUCTURE, CROSS-REFERENCING, MEDIOSTRUCTURE, DICTIONARY, AFRICAN LANGUAGES, BACK MATTER, NORTHERN SOTHO

Opsomming: Die lemmatisering van bywoorde in Noord-Sotho. Tot dusver het Noord-Sotho metaleksikograwe hulle aandag bepaal by lemmatiseringsprobleme ten opsigte van die sogenaamde primêre woordkategorieë, naamlik naamwoorde en werkwoorde. Vergelyk byvoorbeeld, Prinsloo en De Schryver (1999) en Prinsloo en Gouws (1996). Geen aandag is gegee aan die lemmatisering van *bywoorde* nie. Laasgenoemde word deur Ziervogel en Mokgokong (1975: 72, Inleiding) as 'n "sekondêre rededeel" beskou. Die bewerking van bywoorde in Noord-Sotho woordeboeke word bederf deur inkonsekwenthede soos weglatings uit die makrostruktuur, onvoldoende en inkonsekwente etikettering, minderwaardige bewerking in die mikrostruktuur, onderbenutting van die mediostruktuur en buitetekste, en vertoon 'n gebrek aan 'n strategie vir seleksie van items vir lemmatisering. Taalkundige beskrywings van bywoorde in tans beskikbare grammatikas verskil grootliks en verwar dus aanleerders van die taal en onervare leksikograwe.[2] Die doel van hierdie artikel is om oplossings aan die hand te doen vir die lemmatiseringsprobleme rakende bywoorde in Noord-Sotho en gidsinskrywings voor te stel vir papier- en elektroniese woordeboeke

* This article is based on a paper read at the Eighth International Conference of the African Association for Lexicography, organized by the Department of Germanic and Romance Languages, University of Namibia, Windhoek, Namibia, 7–9 July 2003.

wat as modelle vir toekomstige woordeboeke kan dien. Die bewerking van bywoorde in Noord-Sotho woordeboeke sal ook krities geëvalueer word, veral ten opsigte van gebruiksfrekwensie en teikengebruikers se behoeftes.

Sleutelwoorde: LEKSIKOGRAFIE, LEMMATISERING, BYWOORDE, INLIGTINGSONTSLUITING, ELEKTRONIESE WOORDEBOEK, MAKROSTRUKTUUR, MIKROSTRUKTUUR, KRUISVERWYSING, MEDIOSTRUKTUUR, WOORDEBOEK, AFRIKATALE, AGTERWERK, NOORD-SOTHO

Introduction

According to Prinsloo and Gouws (1996: 103), the lexicographer is the mediator between theoretical linguistics and the everyday language user. In practical terms, this often means that the African-language lexicographer has to take great pains in lemmatising grammatically complex systems in a user-friendly way on the level of the target user. Typical examples are the lemmatisation of nouns, verbs, reflexives, adjectives and especially copulatives (cf. Prinsloo 2002). A dictionary should not primarily reflect the attitude of the lexicographer; it should rather be aimed at specific needs of a well-defined target user. It will be illustrated in terms of adverbs that lexicographers should strive to lemmatise adverbs in Northern Sotho in such a way that the whole spectrum of occurrences of adverbs is covered with maximum utilization of all lexicographic mechanisms at their disposal. The user-perspective, and especially the need for modern dictionaries to be user-friendly, has been prominent in lexicographic studies of the past decade (cf. Gouws and Prinsloo (1998), Hartmann and James (1998), Prinsloo and De Schryver (1999), Gouws (2000), etc.) and will be regarded as a given in this article. The South African situation moreover often demands dictionaries to be accessible to a wider user group than originally envisaged by the compiler. Lexicographers should therefore strive towards maximum poli-functionality of their dictionaries. Special attention should be given to the encoding needs of learners, in this case to the need to find enough information in dictionaries in order to actively use adverbs in speech and writing.

The aim of this article is to offer solutions to the lemmatisation problems regarding adverbs in Northern Sotho. It will also be attempted to show how macrostructural and microstructural strategies as well as the mediostructure can be maximally utilized in order to reach this objective. The different kinds of adverbs distinguished for Northern Sotho appear thousands of times in the Pretoria Sepedi Corpus. These enormous overall counts clearly indicate not only that they should be included as lemmas but also that an exhaustive treatment is required and/or justified especially for the encoding needs of inexperienced target users. Prerequisites will be to obtain an overall picture of the adverbial system and to find appropriate lemmatisation strategies for the different types of adverbs in Northern Sotho. The question is therefore what the lexicographer has to know about the adverb in Northern Sotho in order to em-

bark on successful lexicographic treatment of adverbs and how to lemmatise them in a user-friendly way. It cannot be expected from him/her, however, to solve deeply-rooted theoretical differences between linguists on the approaches to the description of adverbs.

It will also be emphasized that in order to lemmatise adverbs successfully, the lexicographer should not hesitate to go beyond 'word boundaries'[3] in the selection of lemmas. Lexical elements smaller than words, such as affixes, and lexical elements larger than words, such as adverbial phrases, should be considered for lemmatisation. Gouws (1989: 84) correctly emphasizes that the traditional focus on the word as representative of the lexicon should be shifted to lemmas representing the lexical items of the particular language.

Although general definitions of adverbs vary, they all formulate the core function of adverbs as describing or modifying a clause or action in terms of especially time, place and manner.

> An **adverb** is a word such as 'slowly', 'now', 'very', 'politically' or 'fortunately' which adds information about the action, event, or situation mentioned in a clause. (Sinclair 1995: 27)
>
> … a word used for describing a verb, an adjective, another adverb, or a whole sentence. Adverbs in English often consist of an adjective with '-ly' added, for example 'quickly', 'mainly', and 'cheerfully'. (Rundell 2002: 20)
>
> … to describe how, where, when or how often something happens … (Procter 1995: 20, textbox)
>
> Adverbs are words which qualify or describe verbs, adjectives and other adverbs in some or other way. (Van Wyk et al. 1992: 118)
>
> … adverbs describe the nature of the action in terms of *time*, *place* and *manner*. (Louwrens 1991: 26)

It could be argued that learners and prospective, inexperienced compilers find the description and treatment of adverbs in currently available dictionaries and grammars of Northern Sotho unsatisfying and even confusing.

Firstly, a wide range of terminology is used to refer to the different kinds of adverbs, viz. basic adverbs, genuine adverbs, common adverbs, secondary derivations, derived adverbs, adverbs that developed from other categories, adopted adverbs, descriptive adjuncts and pseudo-adverbs. With particular reference to adverbial phrases, the terms particles, prepositions and prefixes are used to describe the same kind of lexical elements, depending on the theoretical framework favoured by the author in question. On the one hand different terms such as *basic adverbs* and *genuine adverbs* are used to refer to the same type of adverb while on the other, a single term, for example *derived adverbs*, is used to refer to different types of adverbs by different compilers. The learner can also easily mistakenly assume adverbs derived from other categories, and adverbs developed from other categories to be the same type of adverbs. The latter, however, refers to adopted adverbs. Louwrens (1991) regards *ka*, *le*, *go*,

etc. which introduce adverbial groups, as particles, but Poulos and Louwrens (1994) call them prefixes.

Secondly, Louwrens (1991: 26) says "it is preferable not to regard particle groups as adverbs …" but in Poulos and Louwrens (1994) these groups are indeed regarded as adverbs (see main and subcategories 1 to 6 in Table 2).

The potential confusion for the learner and the prospective lexicographer can also be illustrated by means of *kudu* 'mainly'. Lombard (1985: 166) says it is a basic adverb not related to any other word category. Poulos and Louwrens (1994: 341) agree and add that it is not derived from any other word category and that it has an inherent adverbial meaning. Ziervogel and Mokgokong (1975: 114, Introduction) refer to it as a noun which is a common adverb, and in the central text indicate the part of speech of *kudu* as adverb. Kriel and Van Wyk (1989) label it as a noun of class 9 and offer no treatment of its adverbial characteristics in the entire article of the lemma *kudu*. Van Wyk et al. (1992) and Lombard (1985) recognize 3 basic types of adverbs. Louwrens (1991) distinguishes the categories time, place and manner. Lombard (1985: 168) does not make provision for adverbs of place and says that the so-called adverbs of place are not adverbs. Van Wyk et al. (1992) only say adverbs qualify "in some or other way". Louwrens (1991: 26), in contrast to Lombard (1985) and Van Wyk et al. (1992), does not categorise adverbs in terms of basic, derived and adopted.[4] Poulos and Louwrens (1994) describe adverbs in terms of their derivations and distinguish not less than 9 main categories and up to 17 subcategories. Ziervogel and Mokgokong (1975), in contrast to the other linguists, disregard the category "basic adverb". In fact they describe the nature of adverbs in a rather clumsy way. A dead reference in respect of the final category *ga-* adds to the user's predicament since vital information required to complete the paradigm cannot be retrieved at this point.

> Other parts of speech are used as adverbs, or adverbs may be formed by affixing prefixes or suffixes to other parts of speech. Nouns are often used unchanged as adverbs. … secondary derivations with the secondary formatives ka-, le-, ga-, go- may also be regarded as adverbs. … Adverbs, usually those of quality, are derived from adjective and relative stems by means of ga-. (Ziervogel and Mokgokong 1975: 114-115, Introduction)

Such inconsistencies, whether justified or not, have a negative effect on the learner's and/or user's information retrieval efforts. The issue here is not the validity of their views — criticism on linguistic grounds lies beyond the scope of this article. Furthermore, one should also accept that the adverb can be described from more than one angle and that progressive linguists have the academic right to change their minds. The concern lies with the learner who tries to master the nature and use of adverbs in Northern Sotho and with the lexicographer in his/her role as mediator who finds it difficult to obtain a comprehensive overview of the adverb in order to treat it satisfactorily on the macrostructural and microstructural levels in dictionaries.

Thirdly, a single glance at the treatment of adverbs in Northern Sotho dictionaries reveals far too many inconsistencies and errors. Kriel (1983) includes the lemma *ga(n)nyane* which means that the lemma could either be *ganyane* or *gannyane*. This lemma is placed in the wrong alphabetical position for either *ganyane* or *gannyane*. There is also another treated lemma *gannyane*, again in an incorrect alphabetical position. He gives *ga n'.nyane* as comment on form of *ga(n)nyane* but *ga nya.ne* as comment on form for *gannyane*. Kriel (1950) is inconsistent in respect of circumflexes and POS indication regarding adverbs. As an example of the latter, he labels *gatee* 'once' and *gararo* 'three times' as adverbs but not *gabedi* 'twice'. Ziervogel and Mokgokong (1975) lemmatise the question particles *afa*, *na* and *naa* but indicate the POS of *afa* as adverb. Incorrect alphabetical sorting of lemmas is a common problem in Kriel and Van Wyk (1989), e.g. for *gakale*, compare De Schryver and Lepota (2001, Note 6). Missing punctuation, for example a question mark at *gakakang*, and typing errors such as *by.* instead of *byw.* at *gakalo* are unfortunate. In the latter case the user can interpret the incorrectly spelt label as a translation equivalent, he/she may incorrectly conclude that *gakalo* means *by* 'at' instead of 'so many'.

Form and meaning of adverbs in Northern Sotho

A prerequisite to successful lemmatisation strategies for and treatment of adverbs, is a thorough understanding of the nature of adverbs in Northern Sotho. Poulos and Louwrens (1994: 328) say:

> The analysis of the adverb can be approached in different ways. One could, for example, classify adverbs according to whether they express the concepts of time, place, manner, etc. Or one could describe them in terms of their derivation, that is, in terms of the prefixes and/or suffixes that are used.

Louwrens (1991: 26) says "adverbs describe the nature of the action in terms of *time*, *place* and *manner*" and gives the following examples.

- *Adverbs of time:* Pula e nele *maabane* It rained yesterday
- *Adverbs of place:* Ba dutše *moriting* They are sitting in the shade
- *Adverbs of manner:* Masogana a ja *kudu* The young men eat a lot

Van Wyk et al. (1992: 118) distinguish three types of adverbs namely basic adverbs, derived adverbs and adverbs that have been adopted from other word categories.

- *Basic adverbs* refer to words that are not derived from other words or stems and solely function as adverbs. Examples include *ruri* 'really', *kudu* 'much, a lot', *bjale* 'now', *bjalo* 'like that', *kae?* 'where?', and *neng?* 'when?'.
- *Derived adverbs* are derivations by means of the prefix *ga-*, from nouns and adjectives, e.g. *gabotse* 'well', *gatee* 'once', *gabohloko* 'painful', *gašoro* 'cruelly', etc.

— *Adopted adverbs* are words, such as nouns, which are overwhelmingly or even exclusively used as adverbs such as *maabane* 'yesterday', *bošego* 'at night', *godimo* 'above', *Tshwane* 'Pretoria', etc.

In addition to their discussion in terms of basic adverbs, derived adverbs and adopted adverbs, Van Wyk et al. (1992: 121) also mention that particle groups, such as *ka mehla* 'always', *le gatee* 'not at all', "often function as adverbs".

Poulos and Louwrens (1994: 328) describe adverbs "according to the way in which they are formed". They distinguish 9 categories, viz. adverbs formed by using the prefixes *ka-* (instrumental) (7 subcategories), *le-* (associative) (2 subcategories), *go-* (locative), *ga-* (locative), *mo-* (locative), *kua-* (locative), the suffix *-ng* (locative), the prefix *ga-* (adverbial) and word categories which may function as adverbs (without the addition of any prefixes or suffixes) (8 subcategories).

The nature of the description (time, place and manner), the 3 basic types of adverbs (basic, derived, adopted and particle groups) and the way in which they are formed will now be interlinked in two ways in Tables 1 and 2. Table 1 interlinks the categories of time, place and manner with the three basic types of adverbs that occur in Northern Sotho, and with the way in which they are formed. Table 2 is based upon the way in which adverbs are formed, thus reflecting the viewpoint of Poulos and Louwrens (1994), and interlinked with the three basic types of adverbs as well as with the categories time, place and manner.

In this way the viewpoints of all the above-mentioned authors as well as most of their examples are catered for.[5] The purpose of the compilation of Tables 1 and 2 is threefold. Firstly, either or both of these tables can assist the lexicographer in obtaining a comprehensive overview of the adverb in Northern Sotho. Secondly, these tables can be used in the back matter of a paper dictionary, or, thirdly, in pop-up information boxes in an electronic dictionary. It is for the lexicographer to decide whether he/she prefers to base the back matter entry (entries) and pop-up box(es) on say, Table 1 or Table 2 or both, or whether to use these tables as they are or to adapt them to the level of the target user of the dictionary.

Table 1: Time, place and manner linked to basic types of adverbs and the way in which they are formed (P&L = Poulos and Louwrens 1994)

Time	**Place**	**Manner**
bošego 'at night' ADOPTED P&L-9(v)	***ka* toropong** 'in town' GROUP P&L-1(vi)	**kudu** 'very much' BASIC P&L-9(vii)
neng? 'when?' BASIC P&L-9(vii)	***Go* Madika** 'to Madika' GROUP P&L-3	**gagolo** 'mainly' DERIVED P&L-8
lehono 'today' ADOPTED P&L- 9(v)	***Ga* Madika** 'to/at Madika's place' GROUP P&L-4	**ruri** 'really' BASIC P&L-9(vii)

***ka* Labobedi** 'on Tuesday' GROUP P&L-1(iv)	***mo* tafoleng** 'on the table' GROUP P&L-5	***ka* sefatanaga/lerato** 'with or by means of a car/love' GROUP P&L-1(i), (ii) and (vii)
***ka* letšatši** 'per day' GROUP P&L-1(iv)	***kua* Amerika** 'over there (far away) in America' GROUP P&L-6	***ka ga* molato wo** 'about this problem' GROUP P&L-1(iii)
maabane 'yesterday' ADOPTED P&L-9(v)	**toropo*ng*** 'in/at the town' DERIVED P&L-7	**ka fao** 'because' GROUP P&L-1(v)
nkgapela 'shortly' ADOPTED P&L- 9(v)	**godimo** 'above' ADOPTED P&L-9(iii);	***le* Tate** '*together with* father, *to* father' GROUP P&L-2(i) and (ii)

Thus, for example, *kudu* in Table 1 is a basic adverb of manner belonging to the subcategory (vii) "basic, non-derived adverbs with an inherent adverbial meaning" within the main category 9 "word categories which may function as adverbs without the addition of any prefixes or suffixes" of Poulos and Louwrens (1994).

Table 2: The way in which adverbs are formed, linked to the basic types of adverbs and the categories time, place and manner (P&L = Poulos and Louwrens 1994)

P&L	**The way in which adverbs are formed**		
1	**Adverbs formed by using the prefix ka- (instrumental)**	(i) 'by means of' **ka sefatanaga** 'by car' GROUP: MANNER	(ii) 'with' **ka thipa** 'with a knife' GROUP: MANNER
		(iii) 'about' **ka ga molato** wo 'about this problem' GROUP: MANNER	(iv) Time **ka Labobedi** 'on Tuesday' GROUP: TIME
		(v) 'because of, on account of' **ka fao** 'because' GROUP: MANNER	(vi) Place **ka toropong** 'in town' GROUP: PLACE
		(vii) Miscellaneous **ka lerato** 'with love' GROUP: MANNER	
2	**Adverbs formed by using the prefix le- (associative)**	(i) 'together with' **le Tate** 'together with father' GROUP: MANNER	(ii) translating the English preposition 'to' **le Tate** 'to father' GROUP: MANNER
3	**Adverbs formed by using the prefix go- (locative)**	**Go Madika** 'to Madika' GROUP: PLACE	
4	**Adverbs formed by using the prefix ga- (locative)**	**Ga Madika** 'to/at Madika's place' GROUP: PLACE	
5	**Adverbs formed by using the prefix mo- (locative)**	**mo tafoleng** 'on the table' GROUP: PLACE	

6	**Adverbs formed by using the prefix kua-**	**kua Amerika** 'over there (far away) in America' GROUP: PLACE	
7	**Adverbs formed by using the suffix -ng (locative)**	**toropong** 'in/at town' DERIVED: PLACE	
8	**Adverbs formed by using the prefix ga- (adverbial)**	**gagolo** 'mainly' DERIVED: MANNER	
9	**Word categories which may function as adverbs — (without the addition of any prefixes or suffixes)**	(i) Place names **Tshwane** 'Pretoria' ADOPTED: PLACE	(ii) Other nouns indicative of place **mošate** 'to/at/from the chief's place' ADOPTED: PLACE
		(iii) Nouns of classes 16-18 **godimo** 'above' ADOPTED: PLACE	(iv) Demonstratives of classes 16-18 **mo** 'here' ADOPTED:PLACE
		(v) Time **Bošego** 'at night' ADOPTED: TIME	(vi) Certain possessive forms **la mathomo** 'for the first time' GROUP: TIME
		(vii) Inherent adverbial meaning **kudu** 'very much' BASIC: MANNER	(viii) Either conjunctions or adverbs **fela** adverb 'merely' BASIC: MANNER

Given the presentations of the different linguists of adverbs in Northern Sotho, as well as Tables 1 and 2, it is for the lexicographer to decide on the best angle of approach for lemmatisation of these adverbs. He/she can decide to approach the lemmatisation of adverbs departing from the way in which they are formed, or from the basic types of adverbs or even in terms of their function. Whatever the preferred angle might be, sound decisions regarding lemmatisation, treatment in the microstucture, utilization of the mediostructure, and treatment in the user's guide and back matter have to be taken. In this article, lemmatisation will be attempted on the basic types of adverbs.

Lemmatising basic adverbs

Basic adverbs, in terms of Van Wyk et al. (1992), Louwrens (1991), Ziervogel and Mokgokong (1975), Lombard (1985), as well as Poulos and Louwrens' (1994) Categories 9(vii) and 9(viii), are less problematic. Only a limited number of basic adverbs exist in Northern Sotho and since they are all frequently used, they should all be lemmatised. Consider all the basic adverbs listed by Van Wyk et al. (1992: 118), Louwrens (1991: 26), Ziervogel and Mokgokong (1975: 114), Poulos and Louwrens (1994) and Lombard (1985: 166) and their respective frequency counts in the 6.1 million-word Pretoria Sepedi Corpus, henceforth simply referred to as "the corpus".

Table 3: Overall frequencies of *basic adverbs* in the corpus

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
kudu	4,996	**ntshe**	2,566	**kudukudu**	285	**fela**	17,337	**bjale**	9,397
ruri	3,582	**bjalo**	15,468	**rurirurі**	206	**neng?**	1,794	**kae?**	5,222

In addition to information on frequency, corpus lines and information on collocates obtained from the corpus can be very useful to the lexicographer. Compare for example information on the most frequent collocates of *ruri* in Table 4.

Table 4: Collocates of *ruri* 'really'

	L3	L2	L1		R1	R2	R3
sa			357	**ruri**			
e		376		**ruri**			
le		139	330	**ruri**			

From this extract from the collocates table for *ruri* it is clear that the possessive concord, classes 7/8, *sa* occurs very frequently one position to the left (L1) of *ruri*, thus 357 occurrences of *sa ruri*. Likewise, the copulative stem *-le*, and in fact the entire copulative verb *e le*, are indicated as frequent collocates of *ruri* in the positions L1 and L2 respectively. *Sa ruri* and *e le ruri* are therefore prime candidates for inclusion in the microstructural treatment of *ruri*. Let there furthermore be no doubt that the corpus is a most valuable source for, among others, sense distinction, typical examples, collocations, decisions on inclusion in or omission from the dictionary. Compare De Schryver and Prinsloo (2000, 2000a and 2000b) for an exhaustive overview of corpus compilation and corpus utilization on macro- and microstructural levels.

Lemmatising derived adverbs

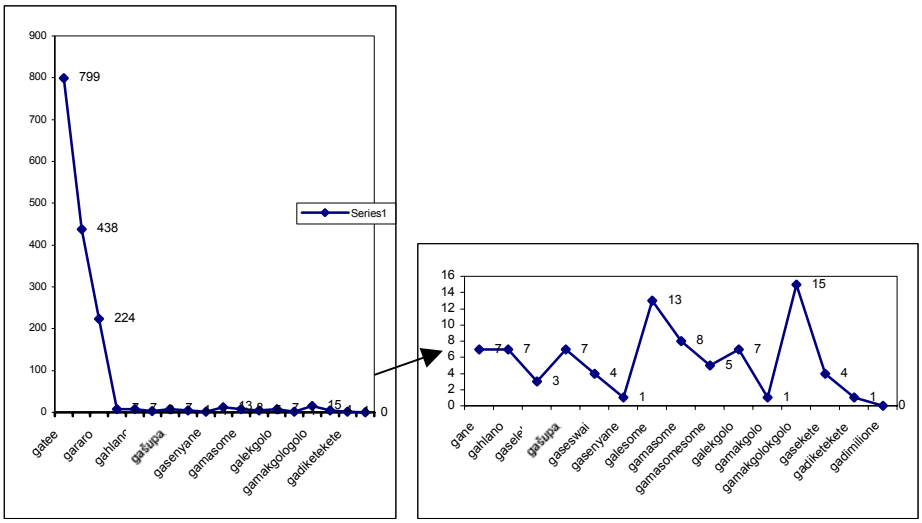
Since the number of derived adverbs is unlimited or open ended, it is not possible to lemmatise all forms separately. From a lexicographic angle, a number of issues are at stake here. Firstly, there is a need for selection in the case of such paradigms. Secondly, the lexicographer has to take decisions in terms of paradigm completion. Thirdly, the lexicographer has to consider certain affixes and particles/prepositions for inclusion in the macrostructure, i.e. as lemmas in their own right. The following analysis is a typical example of how such instances should be approached.

Consider firstly the open-ended paradigm *gatee* 'once', *gabedi* 'twice', … *galesome* 'ten times', … *galekgolo* 'hundred times', … *gadiketekete* 'thousands of times' … . The lexicographer in his/her role as mediator has to take certain decisions, e.g. in respect of inclusion into or omission from the dictionary, after having studied corpus data and available dictionaries.

Table 5: Overall frequencies of the numeral paradigm *gatee*, *gabedi*, … *gadimilione* in the corpus

		Freq.	NEnSeD (1950)	GrNSoW (1975)	Pukuntšu (1983)	Pukuntšu (1989)	Sediba (1992)	SeDiPro (2000)
gatee	once	799	yes	yes	yes	yes	yes	yes
gabedi	twice	438	yes	yes	yes	yes	no	yes
gararo	three times	224	yes	yes	yes	yes	yes	yes
gane	four times	7	yes	no	yes	yes	no	yes
gahlano	five times	7	yes	yes	yes	yes	no	yes
gaselela	six times	3	no	no	yes	yes	no	yes
gašupa	seven times	7 (40)[6]	yes	no	yes	yes	no	yes
gaseswai	eight times	4	no	no	yes	yes	no	yes
gasenyane	nine times	1	no	no	no	no	no	yes
galesome	ten times	13	no	yes	no	no	no	yes
gamasome	tens of times	8	no	no	no	no	no	no
gamasomesome	multiple/several tens of times	5	no	no	no	no	no	no
galekgolo	a hundred times	7	no	no	yes	yes	no	yes
gamakgolo	hundreds of times	1	no	no	no	no	no	no
gamakgolokgolo	multiple/several hundreds of times	15	no	no	no	no	no	no
gasekete	a thousand times	4	no	no	no	no	no	no
gadiketekete	multiple/several thousands of times	1	no	no	no	no	no	no
gadimilione	millions of times	0	no	no	no	no	no	no

Figure 1: Overall frequencies of the numeral paradigm *gatee*, *gabedi*, … *gadimilione* in the corpus



The frequency counts reveal a rather interesting pattern. From Table 5 and Figure 1 it is clear that, from a frequency angle, *once*, *twice* and *three times* are much

more frequently used than *four times* up to *nine times* with relative frequency for "rounded off" numerals such as *ten times* and *hundreds of times*. Treatment in existing dictionaries indicates that the compilers did fairly well on intuition but did miss out on frequently used items such as especially Sediba (Lombard et al. 1992) for *twice*, *four times* and *five times*, and NEnSeD (Kriel 1950), Pukuntšu (Kriel 1983), Pukuntšu (Kriel and Van Wyk 1989) as well as Sediba for *ten times*.

As far as the principle completing-a-paradigm is concerned, two strategies are suggested. Firstly, the lexicographer could complete the 1-to-10 paradigm by also entering *four times* up to *ten times* as separate lemmas, although in terms of frequency counts, this cannot wholly be justified. Secondly, the rest of the open-ended paradigm could be addressed by lemmatising the outstanding "beacons" such as *ten times*, *hundreds of times*, *thousands of times*, etc. Guidance in respect of the paradigm as a whole could be done by appropriate cross-referencing to the back matter. The back matter section would then explain the normal (rather complicated) numerical system of Northern Sotho for expressing numbers from say 1 to 10 and 11 up to 10 000 000 and/or contain references to grammar books where this system is described. Thirdly, the prefix *ga-* (used to derive these adverbs) should be entered as a separate lemma, cf. (1). Compare Gouws (1989: 84) for the importance of lemmatising elements bigger than words and also elements smaller than words.

(1) **ga-** *adv prefix* **gatee** 'once' < **tee** 'one' ***Ba mmethile gatee fela*** *They hit him only once*. **gabotse** 'well' < **botse** 'lovely' ***Sepela gabotse!*** *Go well!* **gantši** < **-ntši; gammogo** < **-mmogo; gagolo** < **-golo; gabedi** < **-bedi** ► BM 2.8[7]

This suggested entry not only caters for the numerical paradigm but also covers the other most frequent typical adverbs formed by means of this derivation strategy, cf. Poulos and Louwrens' Category 8, either as a treated sublemma in the case of *gabotse* or as untreated sublemmas such as *gantši*, *gammogo* and *gagolo*. This brings us to another open-ended paradigm, namely *all adverbs* derived by means of the adverbial prefix *ga-* (of which the numerals just discussed, are only a subsection). Once again the lexicographer has to find a strategy for inclusion or omission. Consider the most frequently used adverbs in this broader category.

Table 6: The most frequently used adverbs derived by means of the prefix *ga-*

		NEnSeD (1950)	Pukuntšu (1983)	Pukuntšu (1989)	Sediba (1992)	SeDiPro (2000)	
gabotse	4,905	yes	yes	yes	yes	yes	nicely, well, carefully
gantši	1,265	yes	yes	yes	yes	yes	often, frequently
gammogo	1,229	yes	yes	yes	yes	yes	together (with), simultaneously
gagolo	804	yes	yes	yes	no	yes	greatly; especially; mostly
gatee	799	yes	yes	yes	yes	yes	once
gannyane	466	no	yes	yes	yes	yes	little, sparingly
gabedi	438	yes	yes	yes	no	yes	twice
gabotsebotse	315	no	yes	no	no	yes	very well, clearly

gararo	224	yes	yes	yes	yes	yes	three times, thrice
gabonolo	235	no	yes	yes	no	yes	easily, with ease
gampe	112	yes	yes	yes	no	yes	badly
gabohloko	71	yes	yes	yes	no	yes	painfully
gabotsana	67	no	yes	no	no	no	somewhat beautiful/nice
gagologolo	51	no	yes	no	yes	yes	especially, chiefly

From this table it is clear that NEnSeD (Kriel 1950) missed out on very frequently used adverbs such as *gannyane*, *gabotsebotse*, *gabonolo*, Sediba (Lombard et al. 1992) on *gagolo*, etc. The high frequency counts for *gabotsebotse* and *gabotsana* furthermore urge the lexicographer to venture beyond the boundaries of the "basic word", viz. those consisting of a prefix and a stem, and also to consider forms with *reduplicated* stems and *diminutive* forms for lemmatising and not merely the basic forms ga + stem.

The next step is to look into *all other* derived adverbs especially Poulos and Louwrens' Categories 1-7 and 9(vi).

Here, each of the particles *ka*, *le*, *go*, *ga*, *mo* and *kua* as well as the suffix *-ng* should be lemmatised with *elaborate* attention in the microstructure of each article to its function as initiator of adverbial groups. For example, one should attempt to cover all of the P&L categories 1(i) to 1(vii) in the treatment of the lemma *ka*. The lexicographer should, preferable in the user's guide, take a clear stand on the use of the terms *preposition* versus *prefix* versus *particle*, and should not burden the user with grammatical labels such as *prep./pref./part.*, cf. (2).

(2) **ka** part. [intr. adv. phrases], **o sepela ka sefatanaga** she goes by car; **ka Labobedi** on Tuesday; **ka toropong** in town. ► BM 1.1-1.3

As in the case of the prefix *ga-* above, the lexicographer should not hesitate to lemmatise the locative suffix *-ng* as an article in its own right.

Poulos and Louwrens' Categories 1(iii), 1(v) and 9(vi) also require special attention. Here the lexicographer should be prepared to lemmatise multiword lemmas such as *ka ga*, *la mathomo*, 'beginning' *la bobedi* 'for the second time' and even, not mentioned by Poulos and Louwrens, *ka mo*, *ka kua*, etc. Furthermore, in the case of *la bobedi* for example, appropriate cross-references should be given to *Labobedi* 'Tuesday' and *bobedi* 'second'. Consider their frequencies in the corpus:

Table 7: Frequencies of multiword adverbs that are candidates for lemmatisation

Lemma candidate	Freq.		Lemma candidate	Freq.		Lemma candidate	Freq.
ka ga	3,709		**la bobedi**	274		**ka kua**	851
la mathomo	487		**ka mo**	4,795			

(3) **la bobedi** *adv.* for the second time, secondly; ► **Labobedi, bobedi**

(4) **la mathomo** *adv.* for the first time, firstly; ► **mathomo**

It should be reiterated that the lexicographer should also and always use the corpus as an invaluable aid to the lexicographic treatment of all types of adverbs of which the study of concordance lines like those in Table 8 generated for the adverb *galesome* 'ten times', is a typical example.

Table 8: Concordance lines for *galesome* 'ten times'

Tataweno yena o mphorile, **a fetoša moputso wa ka**	**galesome**	**fela**, Modimo o mo thibetše go ntira bošul
apa le gorogile maabane. **Malome o bethile Lesibana**	**galesome**.	Naa mošomo wa mantšu ao a ngwadilwe
ditaba tšeo ke tša Chabalala. **O ile a ingwaya hlogo**	**galesome**,	a hloka karabo, fela a tlelwa ke kgopolo y
kilwe maabane. BoMakotlo ba tlile. **Ke tlo go lebalela**	**galesome**	fela. Modimo o tlo go lebalela gamasome a

Such concordance lines are the ideal point of departure for microstructural treatment (cf. De Schryver and Prinsloo (2000b) for a detailed discussion).

Lemmatising adopted adverbs

In the case of *adopted* adverbs, the lexicographer is once again confronted by limited or even open-ended *paradigms* but also with difficult decisions regarding the functions as adverbs versus nouns, especially in terms of part-of-speech indication. Firstly a number of paradigms, this time mostly on a semantic level, have to be dealt with such as *lehono : maabane : maloba*, 'today : yesterday : the day before yesterday', *fase : godimo : morago* 'below : above : behind', *leboa* : *borwa* : *bohlabela* : *bodikela* 'north : south : east : west', etc. Frequency of use and the obligation to complete such semantic paradigms should be the norm.

Lemmatising nouns that are often or even exclusively used as adverbs, *twice*, once with POS-label *adverb* and again with POS-label *noun*, will be totally redundant. In the microstructural treatment, lexicographers often opt for indicating the POS in such cases as *noun* with no reference to a possible adverbial function. Neglecting the POS *adverb* in this way can however only be tolerated up to a point where the labelling of adverbs as nouns becomes artificial and questionable, especially in those cases where nouns are exclusively used as adverbs. The question here is whether the part of speech of nouns that are exclusively used as adverbs should be indicated as *noun*, *adverb* or *both*. What should definitely be avoided is a situation where the same adverb is labelled differently in different dictionaries, or even in different editions of the same dictionary, or where clearly "related" adverbs (i.e. belonging to the same paradigm), are labelled differently in the same dictionary. Consider the treatment of the three words listed by Lombard (1985: 167) as *adverbs that developed from class 6 nouns*, i.e. *maabane*, *maloba* and *mantšiboa*, as a case in point.

(5) (a) *Pukuntšu* (1989)
maabane, byw. … gister; *ka maabane*, in die aand …
(b) *New English–Sesotho Dictionary* (1950)
maabane, adv., yesterday …

(6) (a) *Pukuntšu* (1989)
malôba, snw. kl 6, … eergister, die ander dag …
(b) *New English–Sesotho Dictionary* (1950)
malôba, adv., the day before yesterday …

(7) (a) *Pukuntšu* (1989)
mantšiboa, snw./byw. … aand, in die aand, teen sononder, saans.
(b) *New English–Sesotho Dictionary* (1950)
mantšiboa, n., evening, in the evening, …
(c) *New English–Northern Sotho Dictionary* (1967)
mantšiboa, adv., n., evening, in the evening, …

All these dictionaries offer a single entry for each of these words. In (5) both dictionaries label *maabane* as an adverb, in (6)(a) *maloba* is labelled as a noun with no separate entry or reference whatsoever to *adverb* but in (6)(b) as an adverb. In (7)(a) a single entry is given for *mantšiboa* but with *dual* labelling of its function. In (7)(b), in contrast to (5)(b) and (6)(b), the lemma is now labelled as a *noun* but in a later edition of the same dictionary, i.e. (7)(c), also labelled as an adverb.

Different options can be considered here. The lexicographer could simply ignore the overwhelming or even exclusive function of such nouns as adverbs and consistently label them as nouns (coupled with an explanation in the user's guide and/or back matter of the dictionary) as in (6)(a) and (7)(b). Alternatively, the lexicographer could decide to label the POS in cases where nouns are exclusively used as adverbs, as in (5) and (6)(b) or even in addition to the label *noun*, as in (7)(c). A third possibility, which would represent a sound application of the metalanguage could be to order the POS-labels according to the dominant function, i.e. *n./adv.* if the nominal function is more frequent or *adv./n.* if the word is more frequently used as an adverb. This has to be clearly explained in the front matter of the dictionary. The dominant function can be determined on the basis of frequency counts in the corpus.

Electronic dictionaries

Generally speaking, many more options are available to the lexicographer in electronic dictionaries and fewer restrictions exist in terms of access, available space, mediostructure, etc. See Prinsloo (2001) and De Schryver (2003) for detailed discussions of electronic dictionaries. For example, pop-up screens alone can instantly provide the user with a wealth of information on various aspects of adverbs. This could for instance be done as shown in (8) by simply momentarily resting the cursor on the label *adverb*. Note that all this information, brought together in an instant, also narrows the gap between dictionary and grammar, which is generally believed to be "unbridgeable" (cf. Geeraerts 2000: 77).

Basic adverbs refer to words such as *ruri* 'really', *kudu* 'much, a lot', *bjale* 'now', *bjalo* 'like that', *kae?* 'where?', and *neng?* 'when?'.
Derived adverbs are derivations by means of the prefix *ga-*, e.g. *gabotse* 'well', *gatee* 'once', *gabohloko* 'painful', *gašoro* 'cruelly', etc.
Adopted adverbs are words, such as nouns, which are overwhelmingly or even exclusively used as adverbs such as *maabane* 'yesterday', *bošego* 'at night', *godimo* 'above', *Tshwane* 'Pretoria', etc.

(8) **ga-** *adv prefix* **gatee** 'once' **< tee** 'one'

Time	Place	Manner
bošego 'at night' ADOPTED	**ka toropong** 'in town' GROUP	**kudu** 'very much' BASIC
neng? 'when?' BASIC	**Go Madika** 'to Madika' GROUP	**gagolo** 'mainly' DERIVED
lehono 'today' ADOPTED	**Ga Madika** 'to/at Madika's place' GROUP	**ruri** 'really' BASIC
ka Labobedi 'on Tuesday' GROUP	**mo tafoleng** 'on the table' GROUP	**ka sefatanaga/lerato** 'with, by means of a car/love' GROUP
ka letšatši 'per day' GROUP	**kua Amerika** 'over there (far away) in America' GROUP	**ka ga molato wo** 'about this problem' GROUP
maabane 'yesterday' ADOPTED	**toropong** 'in/at town' DERIVED	**ka fao** 'because' GROUP
nkgapela 'shortly' ADOPTED	**godimo** 'above' ADOPTED	**le Tate** 'together with father, to father' GROUP

Conclusion

Compiling user-friendly dictionaries of a high lexicographic standard for African languages poses a great challenge to prospective lexicographers. They often are the mediators between complicated grammatical structures and the decoding and encoding needs of their target users. Adverbs should not be lemmatised haphazardly as they cross the compiler's way. They should be carefully researched and lemmatised in a structured way. Lexicographers should be aware of the fact that different subcategories of the same phenomenon might require different lexicographic treatments as in the case of basic adverbs versus derived adverbs versus adopted adverbs. This is even true for subcategories within a given category such as the various approaches required for different categories of adverbs derived by *ga-*. On the macrostructural level, candidates for inclusion (or omission) should carefully be considered, preferably

based on corpus data. On the microstructural level, data should be presented in such a way that the needs of both encoding and decoding users are met and the medio-structure should be maximally utilized. The ultimate aim should be to ensure an unimpeded information retrieval process in respect of

— easy access to the lemma,

— successful information retrieval in the microstructure,

— added value obtained in following up on cross-references,

— useful guidance from the user's guide in the front matter,

— a comprehensive overview of adverbs in the back matter, and

— appropriate references to external sources such as grammar books.

Notes

1. An estimated 80% of freelance lexicographers and lexicographers employed by the National Lexicography Units in South Africa have little or very limited lexicographic experience.
2. Na beraming het 80% van alle vryskut leksikograwe en leksikograwe in diens van die Nasionale Leksikografie eenhede min of beperkte ervaring van leksikografie.
3. "Word boundaries" should here be interpreted as for orthographic words.
4. The term "adopted adverb" is used in terms of Van Wyk et al. (1992) in this article and should not be interpreted in the more general sense of *adopted* 'borrowed from another language'.
5. Note that the status of ideophones as adverbs is not recognized in these classifications and requires further research. Compare also Poulos and Louwrens (1994: 351).
6. If Christian religious data in the corpus are taken into account.
7. The symbol ► is a reference marker referring the user to the reference address which in this case is Section 2.8 in the back matter (BM).

References

De Schryver, G.-M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2): 143-199.

De Schryver, G.-M. and B. Lepota. 2001. The Lexicographic Treatment of Days in Sepedi, or When Mother-Tongue Intuition Fails. *Lexikos* 11: 1-37.

De Schryver, G.-M. and D.J. Prinsloo. 2000. The Compilation of Electronic Corpora, with Special Reference to the African Languages. *Southern African Linguistics and Applied Language Studies* 18(1-4): 89-106.

De Schryver, G.-M. and D.J. Prinsloo. 2000a. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The Macrostructure. *South African Journal of African Languages* 20(4): 290-309.

De Schryver, G.-M. and D.J. Prinsloo. 2000b. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 2: The Microstructure. *South African Journal of African Languages* 20(4): 310-330.

Geeraerts, D. 2000. Adding Electronic Value: The Electronic Version of the *Grote Van Dale*. Heid, Ulrich et al. (Eds.). *Proceedings of the Ninth EURALEX International Congress. EURALEX 2000. Stuttgart, Germany, August 8th–12th, 2000*: 75-84. Stuttgart: Stuttgart University.

Gouws, R.H. 1989. *Leksikografie*. Pretoria: Academica.

Gouws, R.H. 2000. Toward the Formulation of a Metalexicographic Founded Model for National Lexicography Units in South Africa. Wiegand, H.E. (Ed.). 2000. *Wörterbücher in der Diskussion IV. Vorträge aus dem Heidelberger Lexikographischen Kolloquium*: 109-133. Tübingen: Max Niemeyer.

Gouws, R.H. and D.J. Prinsloo. 1998. Cross-referencing as a Lexicographic Device. *Lexikos* 8: 17-36.

Hartmann, R.R.K. and G. James. 1998. *Dictionary of Lexicography.* London/New York: Routledge.

Kriel, T.J. 1950. *The New English–Sesotho Dictionary.* Johannesburg: APB.

Kriel, T.J. 1967. *The New English–Northern Sotho Dictionary*. Johannesburg: Educum.

Kriel, T.J. 1983. *Pukuntšu Woordeboek*. Pretoria: J.L. van Schaik.

Kriel, T.J. and E.B. van Wyk. 1989. *Pukuntšu Woordeboek, Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho*. Pretoria: J.L. van Schaik.

Lombard, D.P. 1985. *Introduction to the Grammar of Northern Sotho*. Pretoria: J.L. van Schaik.

Lombard, D.P., R. Barnard and G.M.M. Grobler. 1992. *Sediba, Practical List of Words and Expressions in Northern Sotho, Northern Sotho–Afrikaans–English, English–Northern Sotho.* Pretoria: Via Afrika.

Louwrens, L.J. 1991. *Aspects of Northern Sotho Grammar.* Pretoria: Via Afrika.

Prinsloo, D.J. 2001. The Compilation of Electronic Dictionaries for the African Languages. *Lexikos* 11: 139-159.

Prinsloo, D.J. 2002. The Lemmatization of Copulatives in Northern Sotho. *Lexikos* 12: 21-43.

Prinsloo, D.J. and G.-M. de Schryver. 1999. The Lemmatization of Nouns in African Languages with Special Reference to Sepedi and Cilubà. *South African Journal of African Languages* 19(4): 258-275.

Prinsloo, D.J. and G.-M. de Schryver (Eds.). 2000. *SeDiPro 1.0, First Parallel Dictionary Sepêdi–English*. Pretoria: University of Pretoria.

Prinsloo, D.J. and Rufus H. Gouws. 1996. Formulating a New Dictionary Convention for the Lemmatization of Verbs in Northern Sotho. *South African Journal of African Languages* 16(3): 100-107.

Poulos, G. and L.J. Louwrens. 1994. *A Linguistic Analysis of Northern Sotho*. Pretoria: Via Afrika.

Procter, Paul. (Ed.). 1995. *Cambridge International Dictionary of English*. Cambridge: Cambridge University Press.

Rundell, M. (Ed.). 2002. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

Sinclair, J.M. (Ed.). 1995. *Collins COBUILD English Dictionary*. London: HarperCollins.

Van Wyk, E.B., P.S. Groenewald, D.J. Prinsloo, J.H.M Kock and E. Taljard. 1992. *Northern Sotho for First-Years*. Pretoria: J.L. van Schaik.

Ziervogel, D. and P.C. Mokgokong. 1975. *Groot Noord-Sotho Woordeboek*. Pretoria: J.L. van Schaik.