Compiling a Corpus-based Dictionary Grammar: An Example for Northern Sotho

Gilles-Maurice de Schryver, *Department of African Languages and Cultures, Ghent University, Ghent, Belgium; Xhosa Department, University of the Western Cape, Bellville, Republic of South Africa; and TshwaneDJe HLT, Pretoria, Republic of South Africa (gillesmaurice.deschryver@UGent.be),*
and
Elsabé Taljard, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (elsabe.taljard@up.ac.za)*

Abstract: In this article it is shown how a corpus-based dictionary grammar may be compiled — that is, a mini-grammar fully based on corpus data and specifically written for use in and integrated with a dictionary. Such an effort is, to the best of our knowledge, a world's first. We exemplify our approach for a Northern Sotho mini-grammar, to be included into a Northern Sotho–English dictionary.

Keywords: LEXICOGRAPHY, DICTIONARY, CORPUS, FREQUENCY, MIDDLE MATTER, DICTIONARY GRAMMAR, NORTHERN SOTHO (SESOTHO SA LEBOA)

Samenvatting: Een corpusgebaseerde woordenboekgrammatica samenstellen: een voorbeeld voor Noord-Sotho. In dit artikel wordt aangetoond hoe een corpusgebaseerde woordenboekgrammatica kan samengesteld worden — dit is, een minigrammatica die al z'n gegevens rechtstreeks uit een corpus haalt en die speciaal geschreven werd om in een woordenboek gebruikt te worden, en er ook volledig mee geïntegreerd is. Zo'n poging is, voor zover ons bekend, een wereldprimeur. We illustreren onze aanpak voor een minigrammatica van het Noord-Sotho, bedoeld om gebruikt te worden in een Noord-Sotho–Engels woordenboek.

Sleutelwoorden: LEXICOGRAFIE, WOORDENBOEK, CORPUS, FREQUENTIE, MIDDENWERK, WOORDENBOEKGRAMMATICA, NOORD-SOTHO

1. Using corpora beyond a dictionary's central section(s)

It is now widely accepted that the use of electronic corpora has become indispensable in modern dictionary making, and this on a variety of levels. But just on how many levels? The macrostructural and microstructural levels immediately spring to mind, and most attention in the scientific literature has indeed also gone to aspects revolving around the corpus-based selection of lemma signs on the one hand, and the corpus-based construction of articles attached to

those lemma signs on the other. Any self-respecting dictionary, however, contains much more than 'just' the central text. Good dictionaries also comprise extra matter, invariably distributed across front, middle and back matter sections. If one is serious about corpus-based lexicography, then the extra matter should also be rooted in corpus data. One can come a long way by making sure there is a one-to-one correlation between the central (corpus-based) section(s) and the extra matter (cf. below), but during practical dictionary making this quickly proves not to be sufficient.

In this article the focus will be on the creation of a corpus-based dictionary grammar, exemplified for Northern Sotho. The core principles of corpus-based lexicography will be briefly reviewed in order to set the stage, but that review is merely incidental and the reader is referred to Sinclair (1987) and Corréard (2002) for what remain to this day the best collections on the topic.

2. Corpus-based lexicography in a nutshell

In corpus-based lexicography, the main arbiter during the creation of the (initial) macrostructure is the list of frequencies attached to the lemmatised list of inclusion candidates. Clearly, there are as many lemmatisation policies as there are dictionary teams compiling dictionaries, but it remains common practice to work with and label homonyms, part-of-speech groups, senses (often structured as a hierarchy), and translation equivalents. On each of these levels, frequency considerations are used to order the items, with the more frequent ones presented before the lesser frequent items. In bidirectional explanatory dictionaries, each side draws its data from its own dedicated corpus, so there one has to do this entire exercise twice. Moving to the hard part of the microstructural level, the core process is to derive meanings from the uses as seen in the corpora queried. Each of the uses is then exemplified with corpus extracts — the very extracts that led the lexicographers to divide their articles into senses in the first place. Combinations, derivations, idiomatic expressions, and the like, are all selected by considering their respective occurrence frequencies. Misspellings, misuses (exploitations of the norms?), and anything striking is also noted and finds its way into the dictionary; for example by cross-referring erroneous forms to the correct ones, or by means of the inclusion of usage notes. In short, the very structure of each and every article, even though conforming to a strict and well-defined DTD (document type definition), is 'inspired' by what the dictionary compiler sees in and distils from the corpus. Using a corpus is not a once-off process, but rather a continuous one. This process is greatly facilitated by software where the dictionary writing system (DWS) and the corpus query package (CQP) are seamlessly integrated (De Schryver and De Pauw 2007).

As noted in Section 1, in addition to a dictionary's macrostructure(s) and microstructure(s), most dictionaries also contain extra matter material. Numerous examples exist of dictionaries where the extra matter was clearly written in

isolation, and if not to that extreme, at least not with a full integration in mind. It is classic that front matter descriptions on 'How to Use Your Dictionary' contain examples that, when followed up, lead nowhere — as the very examples used to exemplify the usage, have not been included in the dictionary. Even a cursory glance through dictionary reviews reveals this. In addition, for languages where the orthography has not yet been standardised, one needs to be extra careful to spell the material in the extra matter section(s) in the same way as was done in the central section(s) — and of course one should adhere to consistency within all of these sections as well. Here, too, recourse to a corpus during the compilation seems wise.

In a recent dictionary project, a particularly exciting aspect was the attempt to compile a mini-grammar, based, just as all other aspects of that dictionary, on corpus data. This project, as well as the core features for the write-up of a corpus-based mini-grammar, is discussed in the next section.

3. Isolating the core features for a corpus-based mini-grammar

Given the project is still under embargo, we can unfortunately not reveal the Publisher, nor can we give too many (trade) details. Future publications in the scientific literature will no doubt deal with the larger metalexicographical picture of this project. What can be revealed at this stage are the following features:

— Bidirectional, bilingual Northern Sotho–English dictionary
— Aimed at a well-defined junior target user group, but also with more advanced users in mind (given no other dictionaries are available for this market)
— Corpus-based mini-grammar is part of the middle matter
— Very few pages only for the mini-grammar, eight in all, and this for both a Northern Sotho and an English version (thus four pages each)

From this it is clear that one must attempt to (a) describe the absolute core, and (b) make sure the mini-grammar truly fits and blends into the other sections of the dictionary. The second issue is the easiest, as it means one simply has to attempt to be as consistent as possible at all times: the same abbreviations and conventions as those used in the central sections must be employed, the metalanguage must be presented in the same language as seen in the central sections, all examples given must also be in the dictionary itself, etc. When it comes to the first issue, it seems one could merely attempt to isolate, say, the ten most important grammatical features of Northern Sotho, and to present those. Indeed, and as will be seen in Section 4, some grammar aspects are frequent, while others are not, so theoretically one could just select the top ten based on frequency considerations.

Book-length, fully corpus-based grammars have been written in this way, especially for English. The best known of these is the *Longman Grammar of Spo-*

ken and Written English (Biber et al. 1999). Unlike the latter, for a corpus-based mini-grammar one does not have hundreds and hundreds of pages to describe minute corpus-based phenomena, and also unlike the latter, a corpus-based mini-grammar is not a generic stand-alone description, it is intended to be an integral part of a very specific product. In our case this means that one must keep both the product and its user in mind. Firstly, the product is a bilingual Northern Sotho–English dictionary, which means that one must also cover, even 'bridge', Northern Sotho and English. Reformulated, English grammar, if not the focus of the description, does play a role, and must be contrasted with Northern Sotho where necessary, and in some cases even take centre stage. Secondly, given the target user is not advanced, initially, it means that complicated grammatical issues must be described in as easy a language as possible, but without sacrificing too much accuracy. (After all, specialised vocabulary does have a raison d'être.) To the best of our knowledge, the compilation of a corpus-based dictionary grammar, no matter its size, has never been attempted before.[1]

Terra cognita. It is well-known that lexicographers struggle to find 'the right place' for anything that is not part of their dictionary's central section(s). Users simply do not seem to find their way to any extra matter, and in recent learners' dictionaries, compilers go as far as randomly interspersing their dictionary text with full colour plates or entire reference sections. No studies have reported on the efficacy of this approach, however. Nonetheless, in bilingual lexicography, there are three 'logical' places to present extra matter: preceding the first side of the dictionary (i.e. the front matter), between the two sides (i.e. the middle matter), and following the second side (i.e. the back matter). In our case we chose to make the mini-grammar part of the middle matter, which also includes other material that belongs to a so-called 'Study section'. The publisher is responsible for making this section stand out visually from the core dictionary sections, yet not to such an extent that it is easily skipped. The eight pages were divided as follows: three pages of text in Northern Sotho and three pages of text in English, with between these, two pages for tables and figures with double labelling so they fit both Northern Sotho and English. This seemed to be the most economical use of the allowed space.

Terra incognita. We decided to single out ten topics for the mini-grammar. Following frequency checks of both Northern Sotho as well as English topics, the decision was made to devote seven frequent topics to Northern Sotho, two frequent topics to English, and one *non*-frequent topic to Northern Sotho. The group of seven seems straightforward, and is also what intuition would have dictated. Given the bilingual imbedding of the material, the group of two, which deals with important (English) issues that are often a cause for confusion, can also be explained. The singleton group, which deals with an issue that is not even covered in the dictionary, will clearly need further explanation (cf. below, Section 4.10). The short headings for each of the ten topics are as follows:

— Articles
— Nouns
— Nominal suffixes
— Verbal suffixes, verbal prefixes, negative verbs
— Agreement system
— Adjectives
— Pronouns
— Demonstratives
— Locative particles
— Tone

The sections on 'Articles' and 'Pronouns' deal with English grammar, while the section on 'Tone' has no counterpart in the central section of the dictionary. With regard to the order of the ten topics, frequency considerations were again considered. What is not covered is presented last. Perhaps surprisingly, the issue that leads the list is basically an English issue, as it deals with the English definite and indefinite articles (and we all know that "the" leads any English-language lemmatised frequency list). The order Nouns > Verbs > Adjectives should also not surprise. In short, the ten topics were sorted from most important to lesser important, amalgamating both Northern Sotho and English frequencies. In the next section the mini-grammar itself will be presented, accompanied by liner notes for each.

4. Corpus-based mini-grammar for a Northern Sotho–English dictionary

4.0 Introduction

Northern Sotho belongs to a large family of languages, internationally known as the Bantu language family. Most of the languages spoken in the southern half of Africa belong to this group, and therefore share many grammatical features. Approximately 4.2 million people in the Republic of South Africa speak Northern Sotho as a home language. The main characteristic features of Northern Sotho are summarized in this section. Note that the tables and figures on pages *X* and *Y* accompany this summary.

The grammar is introduced by this paragraph, where, from the start, the dictionary user is alerted to the fact that various tables and figures accompany the text. In this introduction, we look both back (it is a Bantu language) and forward (spoken in South Africa, by 4.2 million people).

4.1 Articles

❶ There are ***no definite or indefinite articles*** in Northern Sotho. This means that there are no equivalents for the English words 'the', 'a' or 'an'.

Starting a mini-grammar of Northern Sotho with what is basically an English grammatical point may seem counterproductive, but was prompted by various

frequency considerations. Not only is "the" the top-frequent English lemma sign (with "a" in fifth position, and "an" in 33rd), real dictionary usage logs involving a Bantu language on the one hand and English on the other, confirm that both the definite article "the" and the indefinite article "a" are among the most frequently searched-for items (cf. De Schryver et al. 2006: 76). It is thus imperative to point out, right from the start, that these articles simply do not have corresponding translation equivalents in Northern Sotho. The lemma sign "a" of course also heads the first article of the English to Northern Sotho side of the dictionary. There, a usage note gives further information (and also focuses more on English grammar).

4.2 Nouns

❷ ***Nouns*** are grouped into classes, which are numbered according to an internationally accepted numbering system. The class to which a noun belongs can be identified by looking at the first part of the noun, which is called the noun class prefix (CP, cf. Table 1). Classes 1 to 10 are arranged **in pairs** with the unevenly numbered classes (1, 3, 5, 7, 9) containing singular forms, and the evenly numbered ones (2, 4, 6, 8, 10) the corresponding plural forms. In your dictionary, nouns are entered under their singular form, and plurals need to be looked up under their corresponding singular form. Also, the class of the headword is shown in bold, together with the corresponding one in non-bold.

Not all nouns have both singular and plural forms — some only ever occur in the singular form, whereas others only have a plural form. In your dictionary, the absence of either singular or plural form is indicated by a dash (–) in the slot where class membership is indicated, for example *setu* 'silence' shows 7/–. Nouns in class 14 normally do not have a plural form and those that do, use the plural prefix of class 6. Nouns in classes 16, 17, 18 and two unnumbered classes (sometimes called the *N*-class and the *ga*-class or class 24) refer to spatial orientation and are called **locative classes**. The locative classes, together with class 15 (the infinitive class), do not distinguish plural forms.

Nouns are by far the most frequent part of speech in the Northern Sotho to English side of the dictionary, as they account for 51% of all entries. Northern Sotho nouns are also the most important words in any sentence, in the sense that the way a sentence *looks* (cf. Section 4.5) is entirely dependent on the class prefix of the noun. Given that the composition of nouns (class prefix + noun stem) is entirely different from that of English nouns, and given that especially the way plurals are formed is very different (in Northern Sotho the prefix changes, compared to the addition of *-s* or *-es* at the end in English), a separate section was needed. (In contrast, basic Northern Sotho verbs work very much like verbs in English, so a separate section on the basic Northern Sotho verb was not needed.)

Also note that guidance is given (here, and throughout the mini-grammar) on how the grammatical characteristics impact on the way the dictionary was compiled, and thus on how certain words should be looked up.

The mentioned 'Table 1' as well as all other tables and figures from the mini-grammar have been reproduced at the end of this article.

4.3 Nominal suffixes

❸ A number of ***nominal suffixes*** can be added to a noun to change its meaning. One such suffix is the **locative marker *-ng***. Adding *-ng* to a noun adds the meaning 'in/at/to/from' to the meaning of the noun. For example: *toropo* 'town', but *toropong* 'in town'. In your dictionary, all frequent cases of nouns with locative markers are treated as derivations under the main noun. Another important nominal suffix is the **diminutive suffix *-ana***, used to express 'small, little, short, etc.', such as in *mokotla* 'bag' versus *mokotlana* 'small bag'. Because of sound changes caused by this suffix, for example *kgarebe* 'young girl' versus *kgarebjana* 'little girl', such forms have been entered as headwords in your dictionary.

Based on frequency considerations, two types of words were lemmatised *with* their nominal suffixes in the dictionary. The first, the locative marker *-ng*, does not cause any (major) sound changes, so nouns with this suffix have simply been lemmatised as sub-lemmas. The second, the diminutive *-ana*, can cause more important sound changes, and with the target user group in mind, such nouns have been lemmatised. The function of this third topic in the mini-grammar is thus dual: firstly to point out that suffixes can be attached to nouns, and secondly to explain where to find nouns with suffixes in the dictionary.

Implicit in this exposition is of course that what are prepositions in English, can be simply nominal suffixes in Northern Sotho. This is but one of many mismatches in parts of speech between the treated language pair.

4.4 Verbal suffixes, verbal prefixes, negative verbs

❹ ***Suffixes*** can also be added to ***verbs***, such as the **relative marker *-go* (or *-ng*)**, the **plural marker *-ng***, or what are known as the **verbal extensions**. The most frequently used single verbal extensions are listed in Table 3. The actual form taken on by these suffixes may vary, following certain phonological rules. Common combinations of verbal extensions are shown in Table 4. Figure 1 shows how frequent the most important verbal extensions are. In your dictionary, the presence of verbal extensions is always indicated, with a cross-reference to the verb stem (when present).

Verbal prefixes such as subject concords (SC) and object concords (OC) are usually separated from the verb stem, but three kinds of prefixes are fixed to the verb stem, namely the **reflexive prefix *i-*** 'self', the **OC of the first person singular *n-/m-*** 'me', and the **OC of class 1 *m-*** 'her/him'. For the latter, this happens only when the verb which follows starts with a *b-*. Examples: *ruta* 'teach' > *ithuta* 'learn; teach oneself', *thuša* 'help' > *nthuša* 'help me', *botša* 'tell' > *mmotša* 'tell her/him'. In your dictionary, all frequent verbs with attached prefixes have been entered.

The ***negative forms*** of verbs are formed by means of the **negative morphemes *ga, sa* and *se***, which appear as verbal prefixes. In some cases, the use of these morphemes cause the verbal ending to change from ***-a*** to ***-e***. For example: *Banna ba aga sekolo.* 'The men are building a school.' > *Banna ga ba age sekolo.* 'The men are not building a school.' In your dictionary, negative forms have been entered as combinations under verbs that end in *-e*, and are preceded by '**ga/sa/se (…)**'.

This fourth topic of the mini-grammar deals with verbs, good for 37% of all entries in the dictionary. The focus is not on the verb stems (i.e. the basic verbs),

as these are straightforward, but on the various suffixes, prefixes and negative forms that surround verb stems.

The suffixes *-go* and *-ng* are non-problematic, which is why a mere mention is adequate. However, no less than 57% of all the verbs entered into the dictionary contain one or more verbal extensions, the orthographic form of which is governed by a multitude of complex rules, which is why Tables 3 and 4, as well as Figure 1, devote a considerable amount of space to this issue. The focus, once again, is on the frequent possibilities only. This is especially clear from Figure 1.

The second paragraph deals with those verbs where prefixes have been written conjunctively, which occurs for reflexive verbs (76 cases in the dictionary), object concords of the first person singular (31 cases), and object concords of class 1 for *b*-initial verbs (13 cases). These forms have been lemmatised as such, as the various morphophonological sound changes make it very hard to isolate the stems of these verbs.

In the third paragraph the so-called *ga/sa/se*-convention for Northern Sotho is briefly described. This convention was introduced by Prinsloo and Gouws (1996), and is a useful (approximate) tool to summarise numerous (negative) tenses into just one compact dictionary article. Negative morphemes can of course be used in combination with any type of verb, including those with pre- and suffixes. In all, 14% of all verbs entered into the dictionary are so-called 'negative verbs'.

4.5 Agreement system

❺ Northern Sotho has a complex linking system (also called an ***agreement system***) in which nouns are linked by means of concords to verbs, adjectives, pronouns and other nouns. Subjects for example, are linked to verbs by means of subject concords (SC): *Baithuti ba bala Seisimane.* 'The students read English.' When the subject is deleted, these concords function as pronouns: *Ba bala Seisimane.* 'They read English.' In the phrase *mmotoro wa gagwe* 'her car' (literally 'the car of her'), the possessive concord (PC) *wa* is the link between the possession *mmotoro* and the possessor *gagwe*. The form of the concord depends on the possessor noun, in this case *mmotoro.* If a noun from a different class is used, the concord will change, for example: *dipuku tša gagwe* 'her books'. See Tables 1 and 2 for all concords.

The noun class system together with the linked concordial agreement system forms the heart of the Northern Sotho grammar. Once nouns and noun classes (Sections 4.2 and 4.3) as well as verbs (Section 4.4) have been introduced, it can be presented. Due to the complexity, one has no other option than to do this in tabular form, accompanied by selected examples. What is rather unique in this corpus-based presentation, however, is that all frequent forms have been singled out and are typographically different from the lesser or non-frequent forms. Indeed, in Tables 1 and 2 all the frequent concords are printed in bold, and these also correspond with those — and only those — items that have been lemmatised in the dictionary.

4.6 Adjectives

❻ In English, we say 'red dress', thus the adjective 'red' precedes the noun which it describes. In ***Northern Sotho***, ***adjectives*** follow the noun: *roko ye khubedu*, where *roko* 'dress' is the noun, and *ye khubedu* 'red' the adjective. Northern Sotho adjectives are made up of an adjective stem (of which there are no more than 30), in most cases preceded by a corresponding class prefix (CP) fixed to this adjective stem, and also a preceding demonstrative (DEM). In both sides of your dictionary, all the frequent adjectives have been entered in full, with their corresponding demonstrative also being shown. See for example the entry for 'black' in your dictionary.

As a result of the limited number of adjective stems, other grammatical constructions are used to describe nouns in Northern Sotho. These constructions often correspond to ***English adjectives***. They are the following:

- Possessive construction, in your dictionary '**[PC +]**', which consists of a possessive concord (PC) followed by a noun: *meetse a borutho* 'warm water' (literally 'water of warmth').
- Nominal relative construction, in your dictionary '**[DEM +]**', which consists of a demonstrative (DEM), followed by a relative noun. Relative nouns often belong to class 14: *bophelo bjo bonolo* 'easy life'.
- Verbal relative construction, in your dictionary '**[DEM + SC +]**', which consists of a demonstrative (DEM) plus a subject concord (SC), followed by a verb that has the relative suffix ***-go***: *mamapo a a elago* 'liquid honey'.

In contrast to Northern Sotho, the English language has hundreds of adjectives (there are around 660 in your dictionary). On the English–Northern Sotho side, adjectives that are not frequent in Northern Sotho are therefore abbreviated. For instance, under 'enormous': **[DEM +] CP**gologolo; **[DEM +]** kgolokgolo. The form with 'CP' (the class prefix) is valid for all classes except 8 to 10, while the other form is valid for classes 8 to 10 only.

In the Bantu languages, there are only about 30 adjective stems. These stems take a class prefix, but even then the total number of Northern Sotho adjectives in the dictionary only amounts to 131 (2.6% of all lemma signs). In contrast to existing dictionaries, where the user must 'construct' the adjective him-/herself, or where a haphazard list of options is given, an explicit approach was followed in the two sides of the dictionary. An example will make this clear. In Kriel et al. (1989[4]) one finds for 'black':

> **swart**, -so (adj.-st.); … [followed by a selection of combinations]

While Prinsloo and Sathekge (1996) have:

> **black** ntsho, moso, baso, leso

Compare this to the treatment of 'black' in the dictionary under discussion:

> **black** **** adjective, noun*
> **I** *adjective* [**blacker**, **blackest**] ⇨*cl. 1* **yo moso** Her father was a tall, black man. • *Tatagwe e be e le monna yo motelele yo moso.* ⇨*cl. 2* **ba baso** This can help black women, so that they have something to plant with. • *Se se ka thuša basadi ba baso gore ba be le se sengwe seo ba bjalago ka sona.* ⇨*cl. 3* **wo moso** My black hair is beautiful. • *Moriri wa ka wo moso o botse.* ⇨*cl. 4* **ye meso** She was wearing a dress with black and white stripes. • *O be a apere roko ya methaladi ye meso le ye*

mešweu. ⇨*cl. 5* **le leso** A fire was made, it was lit and the smoke drifted upwards; it was like a black cloud. • *Gwa gotšwa mollo, ya tshungwa, muši wa kuelela godimo e le leru le leso fela.* ⇨*cl. 6* **a maso** The sky was covered in black clouds that were boiling like the waves of the sea. • *Legodimo le apere maru a maso a a bilogago bjalo ka maphoto a lewatle.* ⇨*cl. 7* **se seso** That black car is moving. • *Sefatanaga sela se seso se a sepela.* ⇨*cl. 8, cl. 10* **tše ntsho** She was wearing black shoes (on her feet). • *Kua maotong o be a apere dieta tše ntsho.* ⇨*cl. 9* **ye ntsho** Please write in black ink. • *Hle ngwalang ka enke ye ntsho.*

II *noun* [*pl.* **blacks**] ⇨ **boso**

◊ **black with white back** ⇨*cl. 9* **ye kgwadi** ⇨*cl. 10* **tše kgwadi** *(said of a bull)*

As can be seen, not only are full forms given (unlike in Kriel et al.), but the translation equivalents immediately contain the demonstrative as well (unlike in Prinsloo and Sathekge). This approach was followed for all 'real' Northern Sotho adjectives.

In order to describe nouns in the absence of enough adjective stems, various grammatical constructions are employed in Northern Sotho, as can be seen from the mini-grammar, viz. possessive concord + noun, demonstrative + relative noun, demonstrative + subject concord + Vgo.

Lastly, in the English to Northern Sotho part of the dictionary, frequent English adjectives that have no frequent counterpart in Northern Sotho, are shown in an abbreviated way, by means of the adjective construction that also underlies the treatment of the article for 'black' just shown, viz., for 'black': [DEM +] CPso; [DEM +] ntsho.

The treatment of adjectives is a clear case where one actually tries to sit on two chairs simultaneously, catering for both the English grammar and the Northern Sotho grammar, and mapping one part of speech onto another one across languages. See for a more in-depth discussion De Schryver (2006).

4.7 Pronouns

❷ ***No distinction is made between 'she', 'he' and 'it'*** in Northern Sotho. Absolute pronouns (and other words that are used as pronouns) are neutral with regard to gender; therefore the absolute pronoun *yena* can mean either 'she' or 'he'; the possessive pronoun [PC +] *gagwe* can mean either 'her(s)' or 'his', and when used as a possessive pronoun, [PC +] *yona* can mean either 'her(s)', 'his' or 'its', for example: *mahlo a yona* 'her/his/its eyes'.

This seventh point in the mini-grammar is comparable to the first, in that it is only mentioned as a result of the pull of the English language. If one had been dealing with, say, a Northern Sotho–Zulu dictionary, there would not have been a need for this point. The above description is self-contained, rests to add that for the translations of the examples in the Northern Sotho to English side of the dictionary, an attempt was undertaken to distribute the genders according to overall corpus statistics. The occurrence of "she" versus "he" is 40/60%, and likewise for "her(s)" versus "his".

4.8 Demonstratives

❽ Three basic positions are distinguished for the ***demonstratives*** of Northern Sotho. Demonstratives are used to indicate the relative distance between the speaker, the person who the speaker is speaking to (the addressee), and the object or person to which the demonstrative refers. These demonstratives can refer to an object that is (a) relatively close to both speaker and addressee (position I), (b) closer to the addressee where addressee and speaker are relatively far apart (position II), and (c) far away from speaker and addressee, who are quite close to one another (position III). See Table 1 for the various possibilities, including variant positions (Ia, Ib, etc.). Note that these positions are also found for the ***demonstrative copulatives***.

Although there are only 47 demonstratives that were frequent enough to be entered into the dictionary (0.9% of all the entries), two thirds of these belong to the top two frequency bands (which means they belong to the top 1 000 words of the Northern Sotho language).

A sound treatment of the demonstratives, as well as of the related demonstrative copulatives, is actually especially difficult on the microstructural level. See De Schryver et al. (2004) for an in-depth, article-length discussion of the lexicographic treatment of the demonstrative copulative in Northern Sotho. As shown there, a sound treatment consists of (a) extra guidance on the microstructural level (by means of context or usage notes), and of (b) multiple levels of cross-referencing, including cross-references between an overview table in the extra matter and the dictionary articles themselves. Both of these have been implemented.

4.9 Locative particles

❾ Northern Sotho makes use of ***locative particles*** to give a detailed description of the place in which an action or process takes place. Five locative particles are distinguished, namely ***ka***, ***kua***, ***mo***, ***ga*** **and** ***go***. When the particles *ka*, *kua* and *mo* are used, they are followed by a noun with a locative meaning, which is often a noun with the locative suffix *-ng*. The locative particles can also be combined, thus forming ***locative bigrams***, consisting of two locative particles each, and ***locative trigrams***, which are combinations of three locative particles. Frequent bigrams are *ka kua* and *ka mo*. Both of these bigrams can combine with *ga* and *go*, leading to the formation of the trigrams *ka kua ga*, *ka kua go*, *ka mo ga* and *ka mo go*. See Figure 2 for the relative distribution of these particles, and Figure 3 for a mnemonic to remember which combinations occur.

On average, every one hundredth word in plain Northern Sotho text and speech is an individual locative particle (cf. De Schryver and Taljard 2006: 141-142). Clearly, one cannot write or speak without them. Even some of the bigrams and trigrams are relatively frequent. As with the demonstratives and the demonstrative copulatives, the main difficulty from a lexicographic point of view is to correctly treat the semantics of the locative particles. As such, the main function of this ninth point in the mini-grammar is to serve as a pointer to the particles themselves — unigrams, bigrams and trigrams — where more information about them may be found.[2]

4.10 Tone

⓾ Northern Sotho is a ***tone*** language, distinguishing two basic tones, namely high (H) and low (L). Every word has its tone pattern, which might change according to the phonological (or sound) environment in which it appears. The tone pattern for the word *mosegare* 'midday' is LHLH, or *mòségàré*. Tone can be used to distinguish between words which are spelled the same, but have different meanings: *anega* (LLL) means 'tell; narrate', whereas *anega* (HLL) means 'hang (something wet)'. It is not the tradition to indicate tone in Northern Sotho texts, and it is not shown in your dictionary either.

As dictionary compilers, we have frequently been asked why we do not indicate tone in our Northern Sotho reference works. The above paragraph summarises the issue. Tolle lege.

Acknowledgements

Gilles-Maurice de Schryver would like to thank *Ghent University* for its continued support of his field trips to South Africa. Many thanks, too, to the Publisher who was willing to embark on this innovative project.

Endnotes

1. For what it is worth, at the time of writing, a *Google* phrase-search for "corpus-based dictionary grammar" (with the quotes) returns zero hits. In comparison, a search for a word as exotic as "automagically" returns over 2.3 million hits (with a nod to D. Joffe).
2. Doing so, the user will surely be rewarded. At *ka mo ga*, for instance, what must be the most stunning corpus example may be found, viz:

   ka mo ga *locative particles* ⇨ **on the side of** Megokgo e tletše mahlo ka lebaka la moya, mongwe o etla ka mo ga hlogo mongwe ka thoko yela gomme e gahlana ka morago ga hlogo. • *Tears filled his eyes because of the wind, one going round this side of the head, the other one on that side, meeting at the back of the head.*

References

Biber, D. et al. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. Grenoble: EURALEX.

De Schryver, G.-M. 2006. Compiling Modern Bilingual Dictionaries for Bantu Languages: Case Studies for Northern Sotho and Zulu. Corino, E. et al. (Eds.). 2006. *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6–9 settembre 2006/Proceedings XII Euralex International Congress, Torino, Italia, September 6th–9th, 2006*: 515-525. Alessandria: Edizioni dell'Orso.

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

De Schryver, G.-M. and E. Taljard. 2006. Locative Trigrams in Northern Sotho, Preceded by Analyses of Formative Bigrams. *Linguistics, An Interdisciplinary Journal of the Language Sciences* 44(1): 135-193.

De Schryver, G.-M. et al. 2004. The Lexicographic Treatment of the Demonstrative Copulative in Sesotho sa Leboa — An Exercise in Multiple Cross-referencing. *Lexikos* 14: 35-66.

De Schryver, G.-M. et al. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

Google. 2007. Google Search Engine [online]. Available http://www.google.com/.

Kriel, T.J., E.B. van Wyk and S.A. Makopo. 1989[4]. *Pukuntšu woordeboek, Noord-Sotho–Afrikaans*, *Afrikaans–Noord-Sotho*. Pretoria: J.L. van Schaik.

Prinsloo, D.J. and R.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.

Prinsloo, D.J. and B.P. Sathekge. 1996. *New Sepedi Dictionary, English–Sepedi (Northern Sotho)*, *Sepedi (Northern Sotho)–English*. Pietermaritzburg: Shuter & Shooter.

Sinclair, J.M. (Ed.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London/Glasgow: Collins ELT.

Table 1: Noun class system and concordial agreement in Northern Sotho

Class (cl.)	CP	Example	SC	OC	PC	DEM pos. I	DEM pos. II	DEM pos. III	DEM pos. Ia	DEM pos. Ib	DEM pos. IIb	DC pos. I	DC pos. II	DC pos. IIIa	PRO Abs	PRO Poss	PRO Poss, communal	PRO Quant
1 1a	mo- Ø-	*mo*sadi 'woman' malome 'uncle'	**o / a**	**mo**	**wa**	**yo**	**yoo**	**yola**	yono	yokhwi	**yowe**	**šo**	**šoo**	**šole**	**yena**	**gagwe**	**wešo**	yohle
2 2b	ba- bo-	*ba*sadi 'women' *bo*malome 'uncles &c'	**ba / ba**	**ba**	**ba**	**ba**	**bao**	**bale**	bano	**bakhwi**	bawe	**šeba**	šebao	šebale	**bona**	**bona**	**bešo / gabo / gabobona**	**bohle**
3	mo-	*mo*nwana 'finger'	**o / wa**	**o**	**wa**	**wo**	**woo**	**wola**	wono	**wokhwi**	wowe	**šo**	**šoo**	**šole**	**wona**	**wona**	**wešo**	**wohle**
4	me-	*me*nwana 'fingers'	**e / ya**	**e**	**ya**	**ye**	**yeo**	**yela**	**yeno**	**yekhwi**	yewe	**še**	**šeo**	šele	**yona**	**yona**	**yešo**	**yohle**
5	le-	*le*bone 'light'	**le / la**	**le**	**la**	**le**	**leo**	**lela**	**leno**	**lekhwi**	lewe	**šele**	šeleo	šelele	**lona**	**lona**	**lešo**	**lohle**
6	ma-	*ma*bone 'lights'	**a / a**	**a**	**a**	**a**	**ao**	**ale**	ano	**akhwi**	awe	**šea**	šeao	šeale	**ona**	**ona**	**ešo**	**ohle**
7	se-	*se*lepe 'axe'	**se / sa**	**se**	**sa**	**se**	**seo**	**sela**	seno	**sekhwi**	sewe	**sese**	seseo	sesele	**sona**	**sona**	**sešo**	**sohle**
8	di-	*di*lepe 'axes'	**di / tša**	**di**	**tša**	**tše**	**tšeo**	**tšela**	tšeno	**tšekhwi**	tšewe	**šedi**	**šedio**	šedile	**tšona**	**tšona**	**tšešo**	**tšohle**
9	N- Ø-	*m*pša 'dog' hlogo 'head'	**e / ya**	**e**	**ya**	**ye**	**yeo**	**yela**	**yeno**	**yekhwi**	yewe	**še**	**šeo**	**šele**	**yona**	**yona**	**yešo**	**yohle**
10	diN- di-	*dim*pša 'dogs' *di*hlogo 'heads'	**di / tša**	**di**	**tša**	**tše**	**tšeo**	**tšela**	tšeno	**tšekhwi**	tšewe	**šedi**	**šedio**	šedile	**tšona**	**tšona**	**tšešo**	**tšohle**
14	bo-	*bo*dulo 'residence'	**bo / bja**	**bo**	**bja**	**bjo**	**bjoo**	**bjola**	bjono	bjokhwi	bjowe	šebo	šeboo	šebole	**bjona**	**bjona**	**bješo**	**bjohle**
(6)	ma-	*ma*dulo 'residences'	**a / a**	**a**	**a**	**a**	**ao**	**ale**	ano	**akhwi**	awe	**šea**	šeao	šeale	**ona**	**ona**	**ešo**	**ohle**

Class (cl.)	CP	Example	SC	OC	PC	DEM pos. I	DEM pos. II	DEM pos. III	DEM pos. Ia	DEM pos. Ib	DEM pos. IIb	DC pos. I	DC pos. II	DC pos. IIIa	PRO Abs	PRO Poss	PRO Poss, communal	PRO Quant
15	go-	*go* ruta 'to teach'	**go / gwa**	**go**	**ga**	mo	moo	mola	mono	mokhwi	mowe	šego	šegoo	šegole	gona	gona		**gohle / gohlegohle**
16	fa-	*fa*se 'below'	**go / gwa**	**go**	**ga**	**fa**	**fao**	**fale**	fano	**fakhwi**	fawe	**šefa**	šefao	šefale	**gona**	**gona**		**gohle / gohlegohle**
17	go-	*go*dimo 'above'	**go / gwa**	**go**	**ga**	mo	moo	mola	mono	mokhwi	mowe	šego	šegoo	šegole	**gona**	**gona**		**gohle / gohlegohle**
18	mo-	*mo*rago 'behind'	**go / gwa**	**go**	**ga**	**mo**	**moo**	**mola**	**mono**	**mokhwi**	**mowe**	šemo	šemoo	šemole	**gona**	**gona**		**gohle / gohlegohle**
N-	N- Ø-	*n*tle 'outside' pele 'in front'	**go / gwa**	**go**	**ga**	mo	moo	mola	mono	mokhwi	mowe	šemo	šemoo	šemole	**gona**	**gona**		**gohle / gohlegohle**
(24) *ga-*	ga-	*ga*re 'middle'	**go / gwa**	**go**	**ga**	mo	moo	mola	mono	mokhwi	mowe	šemo	šemoo	šemole	**gona**	**gona**		**gohle / gohlegohle**

Table 2: Corresponding core information for first and second persons

	SC	OC	PRO Abs	PRO Poss	PRO Poss, communal
1p sg	**ke / ka**	**N-**	**nna / nnaena**	**ka**	
1p pl	**re / ra**	**re**	**rena**	**rena**	**gešo / gaborena**
2p sg	**o / wa**	**go**	**wena**	**gago**	
2p pl	**le / la**	**le**	**lena**	**lena**	**geno**

Notes for Tables 1 and 2:

- All the words printed in bold in Tables 1 and 2 belong to the top-frequency words in Northern Sotho, and are thus included in your dictionary. In order to complete the table, the other forms (in non-bold) have been added, as you may also read or hear them, even though their frequency is low.
- For the N- in Table 2, see the discussion of the verbal prefixes on page *X*, point ❹, on the object concord of the first person singular *n-/m-* 'me'.

Abbreviations used in Table 1, as well as throughout your dictionary:

CP	class prefix	Ø-	zero prefix
DC	demonstrative copulative	1p sg	first person singular
DEM	demonstrative	1p pl	first person plural
OC	object concord	2p sg	second person singular
PC	possessive concord	2p pl	second person plural
PRO	pronoun (absolute, possessive, communal possessive, quantitative)	cl.	class
		N-	class prefix *n-* or *m-*
SC	subject concord	pos.	position

Table 3: Single verbal extensions (most frequent ones only, in order of frequency)

Verbal extension	Name	Meaning
-ile	perfect	Indicates that an action was carried out in the past, or that someone/something is in a specific state. English past tenses and past participles are often translated by verbs with the extension *-ile* in Northern Sotho.
-(i)wa	passive	Adds the meaning of 'be/being' to that of the verb.
-ela	applicative	Adds the meaning of 'to', 'for', 'on behalf of', 'in/to/from the direction of' to that of the verb.
-iša	causative	Adds the meaning of 'cause to', 'help', 'make' to that of the verb.
-ega	neuter-passive	Adds the meaning of 'be/become' to that of the verb.
-ana	reciprocal	Adds the meaning of 'each other / one another' to that of the verb.
-ol, -oll	transitive-reversive	Indicates that the action or state expressed by the basic verb has been reversed. In this way, antonymous verb pairs are formed.

Table 4: Multiple verbal extensions (most frequent ones only, in order of frequency)

Combination of verbal extensions	Example
-ile + -(i)wa	*biditšwe* 'was/were called' < *bitša* 'call'
-ela + -(i)wa	*agelwa* 'be built for' < *aga* 'build'
-iša + -(i)wa	*dirišwa* 'be used' < *dira* 'do'
-ela + -ile	*thabetše* 'be happy about' < *thaba* 'be happy'
-ela + -iša	*fetetša* 'infect; pass on to' < *feta* 'pass'
-ela + -ela	*tsenelela* 'penetrate' < *tsena* 'enter'
-iša + -ela	*tiišetša* 'strengthen' < *tia* 'be strong'
-iša + -ile	*feditše* 'finished' < *fela* 'finish'
-ega + -ile	*diregile* 'happened' < *dira* 'do'

Figure 1: Graphical representation of the frequency of the different verbal extensions (in %)

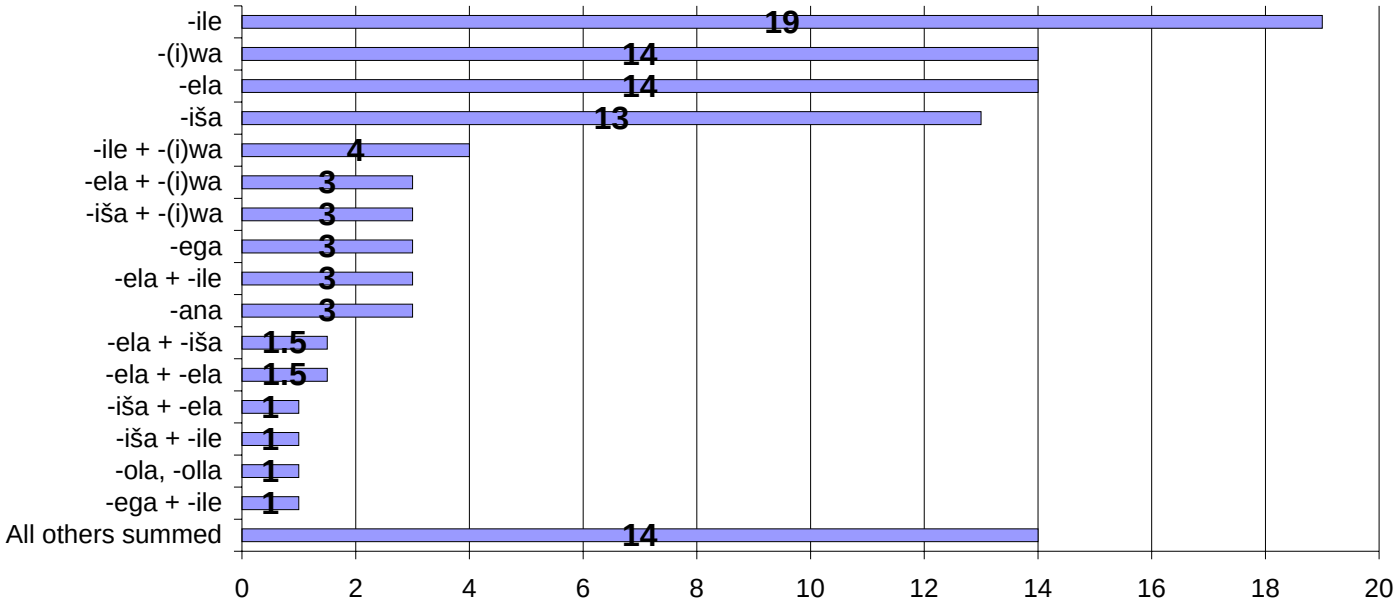

Figure 2: Graphical representation of the frequency of the locative particles

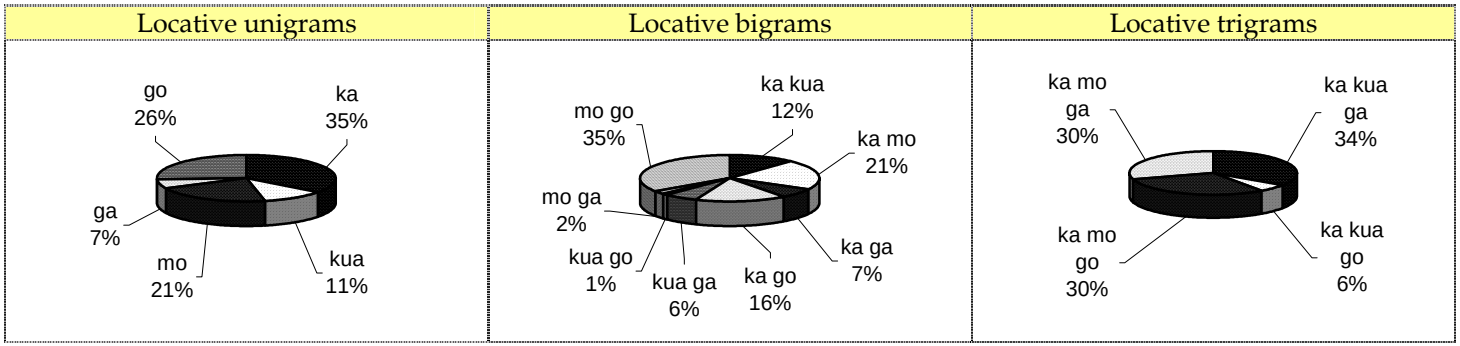

Figure 3: Mnemonic for the possible combinations of locative particles