Chapter 1

Introduction

1.1 Language introduction

Figure 1.1: Geographical extension of the Northern Sotho dialects in South Africa

Sesotho sa Leboa 'Northern Sotho' is one of the three written languages of the Sotho group consisting of Northern Sotho, Tswana ('Western Sotho') and Southern Sotho (All three are comprised in group S.30 in the classification of Guthrie (1971)). What is termed *Sesotho sa Leboa*, however, is in fact a standardised written form of about 30 dialects of the North-Eastern area of today's South Africa and the very south of Botswana (cf. Figure 1.1[1]), some of which differ significantly from others. *Sepedi*, the Pedi language forms its basis,

[1]Figure 1.1 is a cropped form of a map taken from `http://africanlanguages.com/northern_sotho`

according to Ziervogel (1988, p. 1) "with Kôpa elements incorporated".

The Sotho group belongs to the greater group of the South-Eastern Bantu languages (cf. Figure 1.2[2]). The peoples speaking these languages originally made no use of abstract writing systems and it was European missionaries who were the first to record them (mainly in order to enable a translation of the Bible). However, such tasks were enormous undertakings, as, according to (Louwrens, 1991, p. 1 et seq.), "they very soon realised, however, that the words in these languages differ substantially from those found in the European languages." Hence a variety of orthographic systems were developed, which can be distinguished broadly by their word division. The standardisation of this issue is mainly based on the work of Doke, (e.g. Doke (1921)), who set the principle that any approach to word division should be based on pronunciation. According to Louwrens (1991, p. 3), Doke indicated that each orthographic word should have one stress (on the penultimate syllable), and his definition formed the basis of today's conjunctive writing systems of the Nguni languages. Here, one linguistic word could contain a number of (bound and free[3]) morphemes, and is written as one orthographic word as in e.g. the Zulu *Ngingakusiza?* 'Can I help you?'.

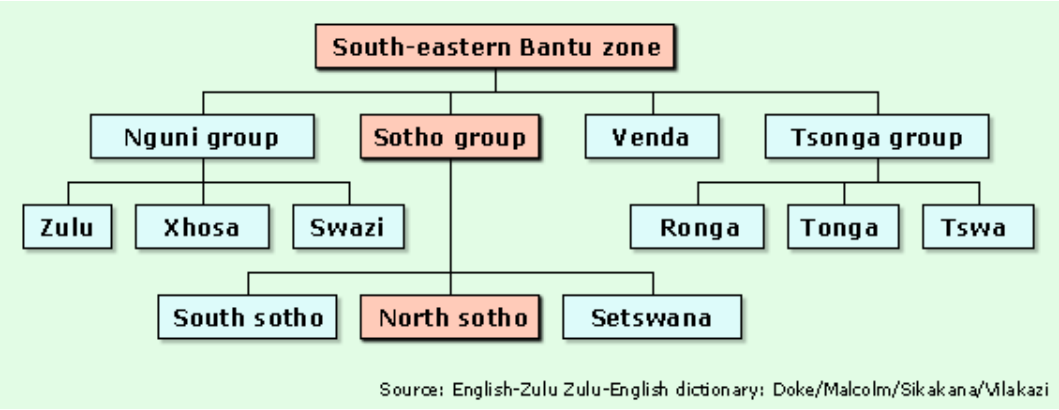

Figure 1.2: Northern Sotho as a part of the South-Eastern Bantu languages

It was Doke, too, who classified the definition of part of speech categories for the South-

[2](ibid.)

[3]In this study, the terms 'free' morphemes are used for words that may appear independently, while 'bound' morphemes are seen as part of an independent word. A noun, for example, is a free morpheme, while a subject concord is bound, because it can only appear as a part of a verb. See also Kosch (2006, p. 5 et seq.)

Eastern Bantu languages based on sound morphosyntactic principles (cf. Louwrens (1991, p. 4)). However, his works concentrated on the Nguni languages, while concerning the Sotho languages, Van Wyk (cf. e.g. Van Wyk (1958)) prepared their standardisation. He based the identification of linguistic words on two principles, namely "isolatability" and "mobility" (cf. (Lombard, 1985, p. 11 et seq.)). These principles are similar to the constituent tests developed by American structuralists (based on Bloomfield (1933)), where the identification of a constituent is mainly based on the possibility of moving and permutation operations.

By application of these principles, Van Wyk defined the linguistic words of the Sotho languages differently from Doke. The Northern Sotho expression of 'Can I help you?', *Nka go thuša?* is however considered one linguistic word, though it consists of three orthographic units, *Nka*, a fused form of *ke ka* 'I may/can', *go* 'you', and *thuša* 'help'. However, the first two of these units are bound morphemes[4].

Still today, there are a number of Northern Sotho phenomena which have not yet been standardised or for which there is no classification. However, its orthography is well developed and the parts of speech mostly defined. Hence the language is considered to be ready to serve for a study viewing it from a computational perspective, for example in a comprehensive morphosyntactic description based on the written form of the words.

1.2 Aims

So far, there have been a number of publications on Northern Sotho, be it on its grammar in the form of prescriptive study books for language learners (inter alia Lombard (1985)), or on some of its morphosyntactic aspects published in several descriptive articles in linguistic journals/conference proceedings (e.g. Anderson and Kotzé (2006)). Others have worked on statistical analysis on explicit grammatical phenomena like, e.g. De Schryver and Taljard (2006), on locative trigrams. However, to the author's knowledge, so far no attempt has been made to describe a comprehensive grammar fragment of the language from a computational perspective, which is the general aim of this study.

Information contained in a message from one human to another is distributed over several

[4]For a more detailed comparison of the Sotho/Nguni writing systems and the similarities between them on morpheme-level, cf. Taljard and Bosch (2006).

levels of communication, non-verbal and verbal, of which only one is the written form, i.e. flowing text. To limit the analysis to written text will consequently lead to a loss of information. Moreover, because a computational text analysis focuses on limited portions of this flowing text, namely sentences, the outcome of such an exercise is reduced even more. Hence an electronic grammar that analyses and generates sentences, can only deliver a fraction of the information generated by the originator of the message. As it is a computer analysing human language, it is therefore necessary to forego world knowledge, tone, etc. As a whole, computational processing of text generally entails such loss of information.

Jurafsky and Martin (2000, p. 285) describe syntax as the "skeleton" of speech and language processing (as words are their foundation). We consider this term to be also valid for morphology, as it describes how words can be made up of morphemes. One could therefore begin with the (orthographic) 'material' of the language, i.e. bound and free morphemes, followed by the description of the language's skeleton, i.e. how these morphemes constitute the grammatical units of the language. Such an approach is very similar to that of mathematics, where the material consists of the numerals of which there are a number of kinds (integers, algebraic numbers, fractions, etc.) and the formulas that can be described as the rules of how the numerals can combine. For example, a 'grammar' of mathematics would describe the different ways in which the numerals can be added together; here, integers can be added by the operation '+'. Fractions however cannot be added this easily, they need specific rules describing the methodology necessary. Other rules would be of a rather restrictive kind in order to e.g. forbid any division with the cypher zero as a denominator. As computers are basically nothing more than mathematical calculators, this approach mirrors the basic layout of every computer program distinguishing 'data' and 'rule'.

Language teachers use a variety of different levels of communication when they explain language items to a learner. Firstly, the medium of instruction is – at least in the first lessons – another language. Secondly, the descriptions often entail instructions on how intentions of a speaker are to be formulated, i.e. what rules the speaker must apply to encode his/her message correctly. The ability of the speaker to also decode utterances subsequently develops from there. A computer can be compared to a learner that can store information on all these units of the language and how they can be combined in seconds. However, it must learn the language usually from scratch, without world knowledge and without anything to compare it to. Consequently, the second aim of this study entails a translation of what

the textbooks describe (as they are the only comprehensive sources) into a reduced level of description of the Northern Sotho language following the 'data versus rules' principle.

There are several approaches for encoding a grammar. Traditionally, generative grammars keep units and (morphosyntactic) rules apart, their rules are based on the fairly strict unit order of the major European languages. The smallest unit that can be utilised by a rule is the word as it is contained in the lexicon. Such units may combine to (or constitute[5]) constituents, i.e. phrases (named after their head, e.g. VP for a phrase headed by a verb), and phrases to sentences. There are indeed a number of approaches to describe languages, for example dependency grammars which view language more as a network of dependencies, where the units are described by the relations between them. Dependency grammars (when they are implemented according to theory) are usually better suited for free constituent order languages. Northern Sotho has a rather strict word order, therefore, in this study, the focus is on the generative approach, not only describing rules for building linguistic constellations of the parts of speech, but also, more generally, attempting to find generalisations of their distributional patterns leading towards a future, more general linguistic modelling of Northern Sotho grammar. For its implementation, we make use of a grammar system capturing both, dependency and constituent structures: the Lexical-Functional Grammar formalism (LFG, cf. Chapter 5).

An electronic grammar can be used for many different purposes, like, e.g. Computer Aided Language Learning (CALL), assisting learners of a language in producing correct sentences. Such grammars might even be error-tolerant, i.e. the analysis of inaccurate sentences is also possible. An error-tolerant grammar informs the learner what error (s)he has produced (e.g. Faaß (2005)). Other electronic grammars are utilised in the development of grammar checkers. A number of machine translation (MT) approaches contain an electronic grammar not only to parse, i.e. to analyse source language sentences, but also to generate the appropriate target sentences. The fourth aim of this study is to show one possible way of utilising such a grammar for the purposes of MT.

In summary, we plan to firstly identify all the core language units in Northern Sotho, secondly their formal relationships, thirdly how an implementation of a fragment of Northern

[5]Note that one unit can constitute a constituent, a phrase, and a sentence as well, e.g. *Bolela!* 'Speak!', an imperative sentence (S) consisting of one unit, the verb stem (V) which constitutes a verbal phrase (VP).

Sotho morphosyntax could be approached, and finally, how an automated translation into English could be designed.

1.3 Methods

Our first aim is to interpret Northern Sotho grammar descriptions in a way that these can be reformulated into computational terms. As our study defines how flowing text could be processed, it is necessary to describe its orthographic units, hence chapter 2 is dedicated to the units the text consists of, categorised in terms of word classes (part of speech). These units present (orthographic) words, i.e. the 'foundation' as described by Jurafsky and Martin (2000, p. 285) and hence the material that the morphosyntactic rules described in chapter 3 may utilise.

The current literature interprets the Northern Sotho constellations from different angles, we opt for choosing one of the most comprehensive descriptions to be our starting point, Lombard's 'Introduction to the Grammar of Northern Sotho' of 1985. Views from other available publications, inter alia Poulos and Louwrens 'A Linguistic Analysis of Northern Sotho' of 1994 are then added for comparison of the given categorisations and definitions. Lombard's grammar was no arbitrary selection, this book is considered by linguists as being not only comprehensive in terms of covering the major phenomena of the language in understandable terms, it also provides information on the morphosyntax of the language in a way easy to reformulate into computational terms. Secondly, Lombard follows the traditional views of Northern Sotho linguistics in that it describes the language's phenomena according to a modal system found in several of the other Southern Bantu languages, too. By following such an approach the methodology of this study qualifies to be transferred to similar languages, like. e.g. Tswana.

All literature utilised indeed begins with a linguistic category, e.g. an indicative present tense sentence, and then describes the morphosyntactic rules to be applied for this category. Chapter 3 will follow this principle. For the sake of completeness, however, chapter 4 begins with a general introduction to parsers and continues with descriptions of the many possible verbal phrases from the perspective of a feature–category relation.

After defining a comprehensive grammar fragment of Northern Sotho, our next task is to define ways in which to implement units and rules. Chapter 5 describes solutions for the

specific challenges of Northern Sotho in the context of an implementation.

Lastly, as the grammar that is described might serve in future as a part of a rule-based machine translation (MT) software, a brief introduction to rule-based MT is provided in chapter 6 together with some ideas on how to solve specific problems concerning an automated translation from Northern Sotho into English.

1.4 A general introduction to grammar

1.4.1 Introduction

As far as we are aware, this is the first attempt to describe a comprehensive fragment of an electronic grammar of Northern Sotho based on literature describing the language and on samples of flowing text, i.e. a corpus. In order to examine the possibilities of its implementation, one first has to explore the nature of the phenomena of the language. Secondly, an existing grammar system should be examined for its capability to handle those language specific phenomena.[6] However, in order to be able to describe the phenomena appropriately, grammatical systems in general and their requirements should be examined as a first step.

The term 'grammar' is assumed by generative linguists to describe the formation of structures containing the units of a specific language. This task includes the following four fields of linguistic research:

- The formation of sounds (phonology);
- The formation of words (morphology);
- The rules that are to be applied when words combine (syntax);
- The meaning of the described components and their composition (semantics).

In his introduction, Katamba (1993) describes how these issues were traditionally seen as four hierarchical levels of linguistic analysis, i.e. levels of representation, each of which could be ideally processed separately. However, information from a higher level can indeed

[6]More details on this issue will be provided in chapter 5, where Lexical Functional Grammar (LFG) will be discussed from the perspective of implementing some Northern Sotho morphosyntactic constellations.

influence a lower level analysis. In terms of applications resulting in modular software systems, these levels of analysis were often implemented in a cascading style, i.e. the output of one level was the input for the next, from sounds to words to sentences to meaning. Such an approach is also called derivational, as one level derives information from another.

Phonological rules describe the system of sounds of one specific language. Northern Sotho as a Bantu language is regarded as a tonal language (e.g. Zerbian (2006)[7]). This opinion is proven inter alia by the minimal pair of *bona/bóna* (the accent marking a high tone, the orthographic word form for both is *bona*). The word *bona* represents the pronoun 'they', while bóna is a verb meaning '[to] see'. A number of researchers therefore use diacritics to resolve such ambiguities when writing in and about this language. However, this study concerns text written according to the official orthography of the language which does not use diacritics on vowels[8]. This consequently means that a number of lexical ambiguities has to be dealt with, based on the orthographic rules of Northern Sotho.

Traditionally, electronic grammars (cf. paragraph 1.4.4) only include morphology and syntax; moreover, the term 'syntax' is sometimes even used synonymously with the term 'grammar', while morphology is seen as another, separate issue of Natural Language Processing (NLP). In this study, the terms 'morphosyntax' or 'morphosyntactic' will appear if we include morphological aspects on the syntactic level, the term grammar will always entail both, i.e. morphosyntactic analysis.

1.4.2 What is a grammar?

In a more narrow sense, a grammar usually describes the use of words being part of a specific language – versus a lexicon listing these words. A lexicon contains a set of words while the grammar describes what to do with them to create sentences. Note that not all sentences which are grammatical are also comprehensible (cf. example (2)).

However, before we can describe the use of *words*, it is necessary to define what is meant by this term, as there are linguistic and orthographic words, content and function or non-content words, etc. That is, before defining morphosyntactic rules of a language, one must describe the 'material', i.e. the units that will become the elements of these rules, as e.g.

[7]Available at `http://www.zas.gwz-berlin.de/index.html?publications_zaspil(Jan2009)`.

[8]The letters a–p,r–u,y, and one special character, namely *š*, occur.

Grefenstette and Tapanainen (1994) state, namely that "any linguistic treatment of freely occurring text must provide an answer to what is considered as a token".

1.4.2.1 Word versus token

Automatic processing of text generally begins with tokenization[9], i.e. the identification of linguistic tokens (a graphical token could be defined as any character sequence surrounded by spaces). A Tokenizer often transforms a flowing text to a one-token-per-line format to ease the further computational processing. Sentence borders are usually identified and marked, too. Linguistic tokens, like, e.g. alphanumeric references, dates, acronyms, or abbreviations, may contain periods. Automatic detection of a linguistic token border is fairly tricky, especially for tokens containing hyphens or quotes (Grefenstette (1994, e.g. p. 118) mentions enclitic forms like 'he's'). However, rule-based tokenizers have been developed that work with a high accuracy, e.g. Grefenstette (1994). Tokenization can be processed by means of statistical procedures, too (cf. e.g. Schmid (1994)). In this study, we categorise tokens roughly as graphical tokens, i.e. as sequences of alphabetic or numeric characters surrounded by spaces or punctuation with the exceptions of dates (e.g. '2009.01.20') and alphanumeric references (e.g. '1.', 'a.' or '(a)').

Linguistic words are seen as the smallest units dealt with in syntactic research while morphological research examines word-internal structure. However, it is not a trivial task to identify linguistic words automatically, as they are often not identical to linguistic tokens: several of them may be contained in one (n:1, cf. 'he's') or they may contain more than one token (1:n). These words occur in a number of languages, like, e.g. the Latin *ad hoc* appearing in English text, or the French negation *ne pas*. In those languages, however, such a phenomenon can be seen as an exceptional case while the language this study deals with, Northern Sotho, is a disjunctively written language, i.e. linguistic words consisting of several tokens are to be considered as the rule, like, e.g., (1) containing a disjunctively written verb, *ke tlo apea* '`subj-1st-sg fut` cook', within a clause[10].

(1) *Nna*$_{\text{PROEMPPERS\_1sg}}$ *ke*$_{\text{CSPERS\_1sg}}$ *tlo*$_{\text{MORPH\_fut}}$ *apea*$_{\text{V\_tr}}$ *dijo*$_{\text{N10}}$
I$_{\text{emphasis}}$ `subj-1st-sg` `fut` cook food
'I (personally) will cook the food'

[9] In this study, other issues like the format of the file where the text is stored, are not accounted for.

[10] Contrary to Anderson and Kotzé (2006), who describe Northern Sotho linguistic words as tokens (their finite state tokenizer identifies linguistic words like *ke tlo apea* as one token), in this study, *ke tlo apea* is considered as three tokens that form one linguistic word.

If a parser (an operational electronic grammar) shall process on a token basis, its developers would have to be conscious of the fact that morphological and syntactic analyses are based on the same kind of units.

1.4.2.2 Semantics versus syntax

Basically, it could be claimed that semantics is the field of linguistics where the previously purely structural form-function analysis of a parser reaches another level. It becomes an analysis of the meaning of a certain structure, i.e. this level of linguistic analysis should lead to a representation of the content message in the uttered or written sentence. However, particularly when the processing begins with the written text, this is no trivial task, as e.g. Palmer (1986) states in his introduction book on semantics. He says that "in language it is extremely difficult, perhaps even impossible, to specify precisely what the message is". He furthermore argues that the problem is to "describe language in terms of language". While language can explain other communication systems like traffic signs, the meaning of a sentence can depend, amongst other considerations, on who is uttering it.

Describing the syntactic/semantic approach of Human Language Technology (HLT), Ramsay (1989) therefore does not claim that the message can be detected completely by analysis, as the context in which an utterance is made is usually not known to the analysis system. He formulates carefully in saying that only the "part of the message carried by the text" may be analysed.

A well developed way to achieve this aim is to define rules, similar to grammar rules that will show how the basic meaning of a single unit in the sentence is modified and/or extended when it combines with the other units. Categorial grammars are a well-known example of such semantic construction algorithms.

A point to add in this matter is that unlike the possible modular implementation of morphology, syntax and phonology, semantics always interfere with all of the other issues. Interference of semantics with grammar is demonstrated when looking e.g. at how the meaning of a verb influences its transitivity. Levin (1992, p. 53 et seq.) states: "A verb denotes an action, state, or process involving one or more participants, the arguments of the verb [...] This set of properties follows directly from the meaning of the verb and plays an essential part in determining how it is used in a sentence."

Nevertheless, the interdependency of semantics with morphology and syntax does not interfere with the grammaticality of a given clause. Chomsky (1957, p. 15) showed with the (famous[11]) example (2) that a sentence can be grammatical without actually making sense:

(2) Colorless green ideas sleep furiously.

One question for a developer of an electronic grammar is hence how deep one should go when implementing these interdependencies between the levels of analysis. Bender et al. (1999) suggest for example a lexical marker on each verb defining whether the verb can be used in a reflexive construction. Such a marker could then be taken into account by the rules forming reflexives, thereby inhibiting the generation of semantically illegal structures like *'I killed myself'.

However, a 'non-reflexive-use' marker has to be set somewhere in the lexicon and this task might take some effort to be put into practice, considering the fact that some verbs, like e.g. 'kill' can indeed be used as a reflexive in certain aspects and tenses ('I am (in the process of) killing myself', 'I might kill myself', 'I kill myself' and 'I will kill myself' are all acceptable). Moreover, in figurative speech a sentence like 'I killed myself trying to fulfil the expectations' is fully acceptable. Such pragmatic aspects of language might lead to the conclusion to not include such semantic restrictions at all.

Another question to be decided upon is on the arguments a verb might select, not only from a morphosyntactic point of view (an issue which will be dealt with in chapter 3), but also by examining the arguments' semantics. It is technically possible to prevent a system from permitting the noun 'idea' to be selected by the verb 'sleep' if there are appropriate characteristics described in the lexicon.

Such decisions influence not only the quality of the resulting grammar, especially in terms of it possibly accepting meaningless sentences, but also practical issues like the manpower needed for its implementation or the processing time necessary to analyse a sentence.

[11]The meaninglessness of this sentence has been challenged a number of times,
cf. `http://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously` for a discussion

1.4.3 The computational perspective

The aim of this study is to develop an electronic grammar for Northern Sotho. The grammar is based upon existing linguistic descriptions of Northern Sotho. In practice, this task entails summarising traditional views of especially Lombard (1985), but also Van Wyk et al. (1992); Ziervogel (1988); Louwrens (1991) and Poulos and Louwrens (1994), from a computational perspective.

Additionally, parts of the *University of Pretoria Sepedi Corpus* (PSC, cf. De Schryver and Prinsloo (2000)) will be utilised, that were collected as a 'gold standard corpus' of Sepedi, containing legal sentences of this language. The tokens were annotated with their parts of speech semi-automatically. All these sentences have been examined by language practitioners for their correctness, they have been analysed manually in order to foresee the results that are expected as an outcome of a parser. Some of these sentences are mentioned in arguments towards an extension of the definitions provided by the above mentioned literature.

When developing a computational grammar it is often necessary to review some of the views previously described by linguists or to put them into a different perspective, because computers have no world knowledge and therefore fully rely on the data and a proper description thereof. For example, when describing noun classes (cf. chapter 2.2.1), we note that all noun classes are marked by certain prefixes. Usually these class prefixes are attached to a nominal root. However, in the case of the noun class 15, the 'infinitive' class, the appropriate prefix is written separately. This issue is noticed and noted in linguistic descriptions, but apart from that it is not of much relevance for the linguist. However, from a computational angle it is quite significant because a computer working on tokenized text will find two tokens instead of one and hence will need additional rules enabling it to process class 15 nouns. Furthermore, the class 15 prefix – unlike any other noun class prefix – can only be attached to verbal phrases (which can be recursive, a fact that is generally neglected in the available literature). Lombard (1985, p. 49) states that "infinitives are nouns and infinitive verbs at the same time". From a computational perspective, however, this infinitive on the level of morphological analysis is to be defined as a verbal phrase with an initial prefix of class 15, and only on the level of syntactic analysis, such a phrase can then be defined as possibly occurring in a nominal function.

Since the 1990s, when the above mentioned literature was published, some developments in the linguistic views on the language of Northern Sotho have also occurred. A tagset has been developed by Taljard et al. (2008) which is not in full agreement with what has been described earlier by other linguists. For example, what previously (e.g. by Louwrens (1991, p. 91 et seq.)) had been seen as a demonstrative "pronoun" or "nominal qualifier" is now described as a demonstrative concord, and some units viewed as particles are now categorised as morphemes. This study provides an attempt to adapt the descriptions found in the above mentioned literature to these newer views, that now rely on a classification of items according to their grammatical function(s).

However, besides the specific issues concerning the development of an electronic grammar of Northern Sotho, there are also general issues to be examined of which the following paragraphs offer a brief overview.

1.4.4 A brief introduction to electronic grammars

In this section, some key issues and terms referring to electronic grammars will be introduced to provide a background for the following chapters 2 and 3, which will define the units of Northern Sotho and the constellations of these units that form legal constituents of the language.

There are a number of approaches to implement electronic grammars, however, this paragraph will focus on an introduction of Context–Free Grammar (CFG)[12] in order to provide a basis. We rely heavily on Jurafsky and Martin (2000) and Sag et al. (2003). Northern Sotho examples will however be used for demonstration purposes.

Context–Free Grammar (CFG) The most famous model used when it comes to developing a parser is the Context–Free Grammar, CFG, sometimes also called Context–Free Phrase Structure Grammar. According to Jurafsky and Martin (2000, p. 327), "it dates back to the psychologist Wilhelm Wundt (1900), but was not formalised until Chomsky (September 1956) and, independently, Backus (1959)".

A phrase structure grammar consists of a set of static rules usually applied to a set of tokens

[12]In chapter 5, Lexical Functional Grammar (LFG) will be introduced.

considered to be words[13] of the language, the lexicon. The phrase structure rules show the form A $\rightarrow \vartheta$, where A is a non-lexical category and ϑ is either a part of speech or another phrase. The arrow (roughly) stands for 'consists of'. Phrases are usually seen as 'equal' in the hierarchy of language, e.g. a verbal phrase (VP) can contain a nominal phrase (NP) and an NP can contain a VP, however, one phrase, usually called 'S' (sentence) stands out as the primary phrase, i.e. nothing can contain 'S' and 'S' must contain all other phrases and their substructures, in other words, whatever the sentence in question includes. This makes 'S' the highest possible node in the parse tree resulting of the parse and guarantees the inclusion of all elements belonging to the sentence. An implementation of these rules can either process top-down (starting from the 'S'-rule to the selection of the correct lexicon entries), or bottom-up (starting by analysing the lexicon entries and finding rules up to the 'S'-rule). The lexicon entries however are each annotated with (at least) their part of speech.

In order to demonstrate an attempt at describing (simplified) Northern Sotho sentences in terms of CFG, we will set up a basic set of rules and a small lexicon for Northern Sotho in CFG as in (3). Each entry in the lexicon will be an element of one set of part of speech, however, for the sake of simplification it is at present necessary in (3) to generalise from the many different categories Northern Sotho has to offer. For example, demonstrative concords and emphatic pronouns will both be categorised as determiners ('D', both are used in Northern Sotho purely for emphasis and can therefore appear with the noun they refer to, but might replace them, too). Concerning the category 'P' note that Northern Sotho is described by linguists as not containing any prepositions, however, some particles appear in a prepositional role. An example is the instrumental particle *ka* 'with', for which the traditional 'P' will be used as part of speech category. A category 'C' is assigned to subject concords, which are part of the verb and responsible for the agreement with its subject. For the moment, noun classes used by Northern Sotho (cf. paragraph 2.2.1) are not considered, as we only distinguish between singular and plural. Lastly, 'A' is used to represent adverbs, locative nouns may also function as adverbs, like $godimo_{\mathrm{ADV}}$ 'high, above'. Our lexicon therefore consists of the following units, sorted by their categories:

- nouns ('N'): *monna* 'man', *banna* 'men', *apola* 'apple', and *maoto* 'foot';
- determiner ('D'): *wena* 'you', *yo* 'this', *ba* 'these';

[13] In the case of Northern Sotho these are orthographic, but often not linguistic words.

- verb stems (as all other parts of the verb are usually written separately) ('V'): *reka* 'buy', *fofa* 'fly', *boa* 'return', *boile* 'returned';
- adverbs ('A'): *godimo* 'high, above';
- particle ('P'): *ka* 'with';
- subject concords ('C'): *o* (sg), *ba* (pl);

The first grammar to be defined should cater for simple sentences, like *monna o reka apola* '(a) man buys an apple', or *wena o boile ka maoto* 'you returned on foot'. It is shown in 3 (a).

3 (a) A first morphosyntactic CFG for Northern Sotho

Rules
S → (NP) VP
NP → N (D), NP → D
VP → C V (NP) (A) (PP)
PP → P NP

Grammar 3 (a) can generate a number of utterances, like, e.g. *monna o reka apola* '(the) man buys (an) apple', or *monna yo o reka apola* 'this man buys (an) apple', or *monna o boile ka maoto* '(the) man returned on foot', or *wena o boa ka maoto* 'you return on foot', etc. Other constellations utilising adverbs, like *(banna) ba fofa godimo* '(the men) they fly high' are licensed, too. The analyses result in parse trees like the two examples shown in Figure 1.3.

However, the grammar also generates illegal forms, because it does not take features of tokens, like, e.g. their valency, into account. Consider the verb *fofa* '(to) fly' that does not require any syntactic object to follow, as it is intransitive: **monna fofa apola* '*(the) man flies (the) apple' would be licensed by this grammar. Such a grammar is called 'over-generating', as it licenses sentences which are not grammatical.

In order to prevent the grammar from such over-generation, different VP rules and finer-grained parts of speech, like IV for intransitive verb and TV for transitive verb, must be introduced, shown in 3 (b).

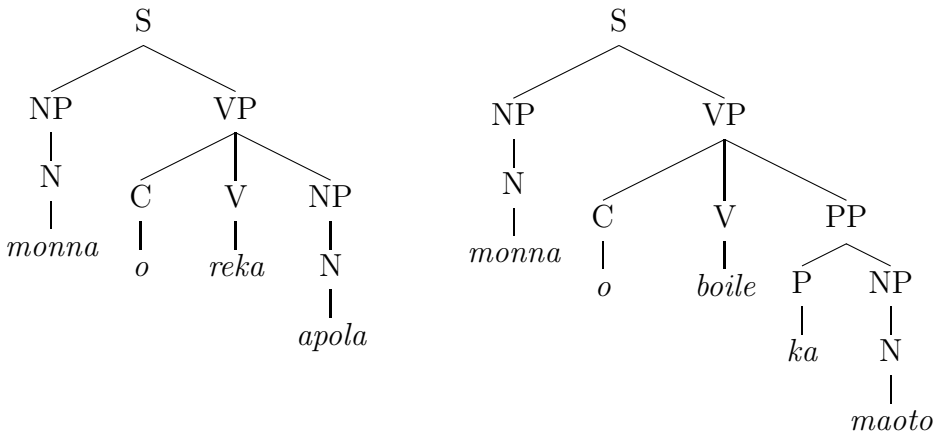

Figure 1.3: Example parse trees in CFG

3 (b) Extended CFG, taking verbal transitivity into account

1.Rules	2.Lexicon
S → (NP) VP	N: *monna*, *banna*, *apola*, *maoto*
NP → N (D), NP → D	D: *wena*, *yo*, *ba*, A:*godimo*
VP → C IV (A) (PP)	IV: *fofa*, *boa*, *boile*
VP → C TV NP (A) (PP)	TV: *reka*
PP → P NP	C: *o*, *ba*, P: *ka*

Though the grammar in 3 (b) does not over-generate in terms of transitivity, it still does not take double transitives into account, i.e. verbs that subcategorise two objects, like, e.g. *efa*, 'give'. Another unsolved problem is the necessary agreement of the subject concords or of the pronouns with the nouns they each refer to, hence, **monna wena ba boile ka apola* *'man you they returned with (an) apple' would be licensed.

Hence, the set of part of speech must be refined even more, e.g. as follows:

- nouns
  - singular ('SN'): *monna* 'man', *apola* 'apple', and *maoto* 'foot';
  - plural ('PN'): *banna* 'men';
- determiner
  - singular ('SD'): *wena* 'you', *yo* 'this';

  - plural ('PD'): *ba* 'these';

- verb stems
  - intransitive ('IV'): *fofa* 'fly', *boa* 'return', *boile* 'returned';
  - transitive ('TV'): *reka* 'buy';
  - double transitive ('DV'): *efa* 'give';
- adverbs ('A'): *godimo* 'high, above';
- particle ('P'): *ka* 'with';
- subject concords
  - singular ('SC'): *o*;
  - plural ('PC'): *ba*;

The grammar rules have to be re-defined accordingly, catering for the newly defined parts of speech, as in grammar 3 (c).

3 (c) The third version of a morphosyntactic CFG for Northern Sotho

Rules
S → (NP) VP
NP → SN (SD), NP → SD
NP → PN (PD), NP → PD
VP → SC IV (A) (PP)
VP → PC IV (A) (PP)
VP → SC TV NP (A) (PP)
VP → PC TV NP (A) (PP)
VP → SC DV NP NP (A) (PP)
VP → PC DV NP NP (A) (PP)
PP → P NP

So far, very simple sentences have been analysed, however, when considering all possible combinations and their various constraints, one can easily appreciate that hundreds of rules would be necessary to reduce the many ungrammatical structures otherwise permitted by the grammar even when dealing with the very simplest constellations of Northern Sotho. The problem lies mainly in the fact that there is only one piece of usable information about an entry in the lexicon, i.e. the part of speech that the rules may utilise. The solution

therefore lies in the definition of more information, to be defined as sets of attribute-value pairs (parametrisation) and in making use of **unification**[14].

In the following sections, legal Northern Sotho constellations will be described on the basis of such pairs.

1.5 Layout of the study

This study consists of seven parts:

- Chapter 1
  - The study's layout and a brief introduction to (electronic) grammars;
  - a definition of the aims of this study;
  - a brief outline of the methodology.
- Chapter 2
  - A description of the word classes of Northern Sotho;
- Chapter 3
  - A description of a grammar fragment of Northern Sotho;
- Chapter 4
  - A description of features of verbal phrases from a computational perspective, i.e. an examination of distributional patterns of the parts of speech contained therein;
- Chapter 5
  - A description of an implementation of parts of the grammar fragment;
- Chapter 6
  - A brief overview of Machine Translation;

[14]These meta data and the uniqueness principle, applied by a number of today's parsers, will be described more thoroughly in paragraphs 4.2 and 5.1.2.2.

  - a description of possible ways to translate Northern Sotho morphosyntactic phenomena into English making use of the grammar approach described.

- Chapter 7

  - Summary and conclusions.

Chapter 2

The word classes of Northern Sotho

2.1 Introduction

The question why word class categorisation is a necessary instrument for linguistic research has many answers; however, we will only state two of them. First, the same word class is usually found in the same textual environment of a language. For example, the nouns of English can be semantically modified by adjectives but not by verbs. Secondly, morphological rules are only valid for specific parts of speech. The German 'weak', i.e. regular verbs, for example, add the suffix *–te* to their roots when forming the past (*kauf–te* 'bought'), while the 'strong', i.e. irregular verbs experience changes in their roots (*ging* vs. **geh–te* 'went'). Therefore a distinction between 'weak' and 'strong' verbs is necessary whenever the lexicon containing the verbs will be used by a morphology system generating the appropriate past tense forms.

The aim of such a categorisation is described by Taljard (1995, p. 11) as "to establish major word categories which will aid the linguist in arriving at the simplest and logically most consistent description of the grammar of the language in question".

To assign word classes or parts of speech (POS) has a long tradition in European Linguistics. According to Jurafsky and Martin (2000, p. 287) the first known classification had already been written as long ago as 100 B.C. (possibly by Dionysius Thrax of Alexandria). This set contained eight elements: noun, verb, pronoun, preposition, adverb, conjunction, participle and article. Since then, a number of word classes have been assigned, like e.g. the class containing adjectives, usually in order to define such sets for languages other than Greek - though these very basic classes assigned over 2000 years ago are still in use for a great number of languages today.

We heavily rely on Jurafsky and Martin (2000, p. 288) in the following paragraph, when reasoning why especially computer applications usually require the assignment of POS to the words of a language before doing any further processing. However, we will only list a few of the arguments mentioned there.

In Text-to-Speech (TTS) applications, information on the POS of a word is often decisive when it comes to deciding upon putting stress on the correct syllable, as in 'CONtent' (noun) vs. 'conTENT' (adjective) or 'TRANSport' (noun) versus 'transPORT' (verb). However, this approach does not cover all cases. There are words like the German umFAHRen 'drive around' vs. UMfahren 'drive over', which both are verbs. Stemming and lemmatising are further applications where it is essential for a computational system to be informed about the class of a word. Following the generative grammar approach, morphological rules are hence defined not on word, but on POS-level[1]. To abstract words on a POS-level also allows further generalisations on a language, e.g. in defining grammatical rules that entail POS rather than words. The POS therefore usually supply the lexical contents of the syntagmatic rules describing a language's grammar. The grammar rule NP → DET N (a noun phrase consists of a determiner followed by a noun) is an example where such generalisation is put into practice, it makes use of a lexicon where all determiners of the language are listed as DET and all nouns as N.

In the following paragraphs, the POS of Northern Sotho will be introduced as they are labelled in the system's lexicon used for this project, following the tagset definitions by Taljard et al. (2008) with some minor updates. This tagset describes all orthographic units written separately, no difference is made between bound and free morphemes.

It is not the aim of this chapter to give a systematic overview of all POS-features. The discussion will be limited to a brief introduction of the issue and should rather be seen as an explanation of key concepts and of terms that will be used in the remaining study. For a more detailed understanding, the literature mentioned in the following paragraphs should be consulted.

In this study, part of speech labels are written as a subscript on the righthandside of the

[1]The theory of distributive morphology (cf. `http://www.ling.upenn.edu/~rnoyer/dm/`), see also paragraph 3.1 on page 69, rejects this approach. However, this study is theoretically based on generative grammar, such as it describes a lexicon component.

word, whenever deemed necessary. The tagset defined by Taljard et al. (2008) utilised for this study defines two levels of annotation of which the second is separated by an underscore, e.g. N01_aug where 'N01' (noun of noun class 1) is on the first level of annotation and 'aug' (augmentative derivation) on the second. In this study, we also make use of a variable 'category', ($_{\text{categ}}$) whenever we mention a major word class category which contains a number of POS, e.g. N_{categ} containing N01, N01a, N02, N02b, N03, N04, ... , N10, N14, N15, NLOC, or CS_{categ} containing CS01, CS02, CS03, ..., CS10, CS14, CS15, CSPERS, CSINDEF, CSNEUT, etc.

2.2 The noun (N_{categ})

2.2.1 The noun class system

We introduce the noun classes by referring to Ziervogel (1988, p. 1):

> "Each person or thing, concrete or abstract, is placed in a particular category or group in Northern Sotho. In grammatical terms we speak of nouns placed into classes. Take for instance the following words which indicate persons or things, i.e. nouns:
>
> *motho* (person) plural *batho* (persons, people);
> *motse* (village) plural *metse* (villages);
> *selepe* (axe) plural *dilepe* (axes)."

The use of grammatical gender, i.e. a noun class system, is a distinctive feature of the Bantu Language family as classified by Guthrie (1967). According to Lombard (1985, p. 4), there are "close to a thousand Bantu languages and dialects". As these show many "linguistic similarities", it can be assumed that they all developed from one proto-language form. In all these languages each grammatical gender is represented by a morpheme and each morpheme becomes overt as a set of allomorphs prefixed to noun and nominal stems. Mutaka (2000, p. 151) lists 24 noun classes at large, however, usually Bantu languages do not use all of these classes. Noun classes are not named, but simply numbered, with a distinction in the grammatical number, i.e. the singular and plural form of one noun is found in two different classes. Word classes that have to agree with nouns when referring

to them show their agreement by using a class specific prefix.

Northern Sotho uses 17 classes: 1 to 10 and 14 to 18, plus two 'unnumbered' locative classes, the so-called *N–*[2] and *ga–*locative classes.

Some noun stems occur in more than the two classes representing their singular and their plural form. Consider the root *–tho* (which occurs in *motho* and *batho* above), this stem form[3] uses the prefixes *se–* and *di–* as well, thus forming *setho* 'ghost' (plural: *ditho*).

There are stems that appear in as many as 10 classes, like *–dimo* which occurs in the class specific forms $ledimo_{\text{N05}}$ '(thunder-)storm' (plural: $madimo_{\text{N06}}$), $Modimo/modimo_{\text{N01}}$ 'God/ghost or spirit of a deceased' (plural: $badimo_{\text{N02}}$), $modimo_{\text{N03}}$ 'evil spirit' (plural $medimo_{\text{N04}}$), $sedimo_{\text{N07}}$ 'sacrifice' (plural $didimo_{\text{N08}}$), $bodimo_{\text{N14}}$ 'cannibalism' (no plural) and $godimo_{\text{NLOC}}$ 'high, above, in the air' (no plural). There seem to be semantic relations between at least some of these forms. Such an observation leads to the idea of grouping nouns into their respective classes by using semantic features. Endemann (1911, p. 22), for example, describes the nominal prefixes *m–*, *mo–* , *me–* , *ma–* as specifying something static; of these, according to Endemann (1911, ibid.), *mo–* stands for something static being singular, a condition, a circumstance, something locally standing, something that exists in itself. Ziervogel (1988, p. 2) classifies such groups inter alia as person class, class for terms of relationship, or including natural phenomena, abstracts, etc. Poulos and Louwrens (1994, p. 13) describe e.g. one class as referring "in most cases to the various entities and elements that characterise our world of nature". Others, more formal attempts have been made to a semantic classification as in Givón (1971), however, such a classification system cannot be used by a computer system at the current point in time because an electronic lexicon containing the meaning of each root (and a system which would possibly determine the meaning shifts occurring when a specific root combines with different affixes) has not yet been implemented successfully. Furthermore, for each such rule, there are a number of exceptions. Lombard (1985, p. 42), for example, describes class 6 as containing

[2] *N–* is used by Northern Sotho linguists do describe nasals, like, e.g. *m–* or *n–*. Prefixing *N–* e.g. to verbal stems may lead to plosivation.

[3] We explicitly refer to the surface form of the examples when stating that some are 'identical', as we describe their orthographic forms only. The system to be developed in this study does not include any non-textual analyses. Note that tonal differences do not show in the official orthography of Northern Sotho, hence some of the stated 'identical' forms might instead be homographs belonging to different discourse entities.

nouns that among others "indicate times and seasons" (e.g. *marega*$_{\text{N06}}$ 'winter', *maabane*$_{\text{N06}}$ 'yesterday'). However, *lehlabula*$_{\text{N05}}$ 'summer' and *gosasa*$_{\text{N09}}$ 'tomorrow' do not belong to this class.

A computer system is therefore dependent on the surface forms when determining the class of a noun, i.e. the noun class prefixes that will be described in the following paragraphs. We will however suggest some general semantic categories in the following tables as well in order to refer back to the literature.

2.2.2 Noun class prefixes - an overview

The tables in this section reflect an attempt to summarise the viewpoints with respect to noun formation of Endemann (1911), Lombard (1985), Poulos and Louwrens (1994) Van Wyk et al. (1992), Ziervogel and Mokgokong (1975) and Ziervogel (1988). In addition to the 'base' forms of nouns, these tables also contain nominal derivations of nouns; i.e. the diminutive, augmentative and locative forms (locativised nouns). The diminutive is formed by adding the suffix *–ana* or *–yana* to a noun. An augmentative noun is formed by adding the suffix *–gadi*, which entails an augmentative meaning and/or a feminine one. Lastly, the locative is formed by suffixing *–ng* to a noun, thereby adding an aspect of locality that can be local, like in *toropong* 'in/at the town' (a derivation of *toropo* 'town'), or temporal, like in *bekeng* 'in/during the week' (derived from *beke* 'week'). A more abstract locality can be found in the example contained in Table 2.1, *mererong* 'in the plans/intentions'.

Note that some of the noun prefixes mentioned in the following paragraphs are not linguistically defined prefixes but results of fusing processes. However, as only the surface forms are taken into account in the frame of this study, these results of fusing processes are treated the same way as morphemic prefixes.

2.2.2.1 Noun classes 1 to 4

Table 2.1 shows a summary of all prefixes used by classes 1 to 4, sorted alongside the class numbers they refer to. The symbol Ø– stands for the zero-prefix (as described by Poulos and Louwrens (1994, p. 11)). The third column shows the word class label assigned in the system's lexicon as they appear in this work.

Table 2.1: Overview of the prefixes of the noun classes 1 to 4 and their referring annotations

Class no.	***Class prefixes***	***Annotation***	***Morphosyntactic and other properties, examples***
01	*mo–*, *mm–* *ngw–*	N01	singular, personal nouns only, *morutiši* 'teacher', *mmuši* 'governor', *ngwana* 'child'
		N01_dim	*ngwanana* 'little child' (diminutive derivation)
		N01_aug	*morutišigadi* 'female teacher' (augmentative derivation)
		N01_loc	*molwetsing* 'at the sick person' (locative derivation)
01a	*mm–*, Ø-	N01a	singular; nouns expressing kinship. *mme–* 'my/our mother', *tate* 'father', *kgaitšedi* 'sister'
		NPP	names of places: *Tshwane* etc.
01a	Ø-	N01a_name	singular, proper names, always beginning with an upper case letter, *Dikeledi, Thabo, Mphahlele, Sekhukhune*
02	*ba–*	N02	plural of class 01 *barutiši* 'teachers', *babuši* 'governors', *bana* 'children'
		N02_dim	*bašemanyana* 'little boys' (dimunitive derivation)
		N02_aug	*barutišigadi* 'female teachers' (augmentative derivation)
		N02_loc	*bathong* 'amongst the people' (locative derivation)
02b	*bo–*	N02b	generally, N02b is the plural of N01a and also of other classes, it may be translated as 'X and company', cf. Van Wyk (1987) *botate* 'fathers'/'father and company' *bomme* 'mothers'/'mothers and company' *botau* 'Mr. Lion and company' (*tau* 'lion' is a class 9 noun)
02b	*bo–*	N02b_name	singular, respect form of class N01a_name which is still written with an initial upper case letter *boMphahlele, boSekhukhune*
03	*mo–*, *mm–*, *mph–*, *mpsh–*, *ngw–*	N03	singular, impersonal; *mmotoro* 'car', *mmala* 'colour', *mphago* 'road food', *mpshiri* 'copper (bangle)', *ngwaga* 'year'
		N03_dim	*mokgwanyana* 'a little habit'
		N03_aug	*modimogadi* 'goddess'
		N03_loc	*motseng* 'in the town/village'
04	*me–*, *nyw-* *mengw–*	N04	plural of class 3 *mebotoro* 'cars', *mebala* 'colours', *nywaga* 'years', *mengwaga* 'years' (alternative form)
		N04_dim	*meropana* 'tambourines'
		N04_aug	*medimogadi* 'goddesses'
		N04_loc	*mererong* 'in the plans/intentions'

2.2.2.2 Noun classes 5, 14, and 6

As Table 2.2 demonstrates, the classes 5 and 14 both use class 6 as their plural class. Additionally, other nouns occur in this class. Van Wyk et al. (1992, p. 11), amongst others, lists liquids like *meetse* 'water' or *maswi* 'milk'. On the other hand, there are also nouns in class 6 that do not necessarily indicate a plural, e.g. *maabane* 'yesterday'.

Table 2.2: Overview of the prefixes of the noun classes 5, 14 and 6 and their referring annotations

Class no.	***Class prefixes***	***Annotation***	***Morphosyntactic and other properties, examples***
05	*le–*, Ø–	N05	singular, *leoto* 'foot', *lapa* 'yard', *leino* 'tooth'
		N05_dim	*lebakanyana* 'short period of time'
		N05_aug	*ledimogadi* 'biggest thunderstorm, tornado'
		N05_loc	*lebakeng* 'during that time'
14	*bo–*, *bu–* *bj–*	N14	singular (often abstract nouns) *bodiidi* 'poverty', *bodutu* 'loneliness, boredom', *bupi* 'meal', *bjala* 'beer'
		N14_dim	*bolotšana* 'wickedness, fraud', *bošemanyana* 'boyishness'
		N14_aug	*borutišigadi* 'great teaching', *bohlologadi* 'very bad (immoral) thing'
		N14_loc	*bolwetšing* 'in sickness'
06	*ma–*, *m–*	N06	plural class of classes 5 and 14 *magotlo* 'mice', *mahodu* 'thieves', *maoto* 'feet' *meetse* 'water', *meno* 'teeth'
		N06_dim	*mafotwana* 'young of birds'
		N06_aug	*madimogadi* 'tornados'
		N06_loc	*maotong* 'on the feet/legs'

2.2.2.3 Noun classes 7 and 8

Classes 7 and 8 are two of the few classes that use only one class prefix. This is the class of nouns that is the easiest to determine, because *se-* is exclusively used as the prefix of class 7[4]. Van Wyk et al. (1992, p. 11) list inter alia names of languages and cultures in this class, which constitute proper nouns. These should be written upper case in all positions of the sentence, e.g. *Seisemane* 'English'. Table 2.3 shows examples.

Table 2.3: Overview of the prefixes of the noun classes 7 and 8 and their respective annotations

Class no.	***Class prefixes***	***Annotation***	***Morphosyntactic and other properties, examples***
07	*se-*	N07	singular *selepe* 'axe', *semumu* 'lazy/mute person', *Seisemane* 'English' *Sepedi* 'Language and/or Culture of the Bapedi'
		N07_dim	*sešwana* 'small cake of dry dung'
		N07_aug	*sefefegadi* 'something huge that flies (e.g. a jumbo jet or a 'very big bird'), *setšhabagadi* 'big tribe'
		N07_loc	*sekgoweng* 'in white/urban areas'
08	*di-*	N08	plural class of class 7 *dilepe* 'axes', *dimumu* 'lazy/mute persons'
		N08_dim	*dikgalabjana* 'little/worthless old men'
		N08_aug	*difefegadi* 'big things/animals that fly'
		N08_loc	*diatleng* 'in the hands'

2.2.2.4 Noun classes 9 and 10

Class 9 makes use of the zero prefix Ø– only if the nominal root of a noun consists of two or more syllables. Whenever the root consists of just one syllable, the nasal prefix *N*–is added, *m*– in the case of a root beginning with *p*–, otherwise *n*–. In terms of its contents, a number of loan words can be found in this class, like, e.g. *namoneiti* 'lemonade'.

Class 10 nouns can easily be mistaken for class 8 nouns because they use the same prefix, *di*–. Identification of the correct class is however usually possible by examining the singular

[4]Note that there are indeed words of Northern Sotho that begin with this prefix, but are not class 7 nouns, e.g. $seba_{\text{V}}$ 'whisper', $sebe_{\text{ADJ}}$ 'bad', $semangmang_{\text{N01a}}$ 'so-and-so' (as a person), cf. De Schryver (2007).

Table 2.4: Overview of the prefixes of the noun classes 9 and 10 and their referring annotations

Class no.	*Class prefixes*	*Annotation*	*Morphysyntactic and other properties, examples*
09	Ø–, *m–* *n–*	N09	singular *hlapi* 'fish', *mpšhe* 'ostrich', *ntlo* 'hut', *nko* 'nose'
		N09_dim	*kgalabohlajane* 'little knowledge'
		N09_aug	*namagadi* 'female (animal)'
		N09_loc	*nkong* 'on the nose'
10	*di–*	N10	plural class of class 9 *dikgoši* 'chiefs', *dinko* 'noses', *ditau* 'lions'
		N10_dim	*ditemana* 'paragraphs'
		N10_aug	*dinamagadi* 'female (animals)'
		N10_loc	*ditabeng* 'in/concerning this matters'

form of a given noun. If the singular form uses the prefix *se–*, the noun will belong to class 8, in any other case, it will belong to class 10. Table 2.4 summarises the classes 9 and 10.

2.2.2.5 Noun class 15 - The infinitive

Class 15 is described in the literature as containing the "infinitives", which could be interpreted as the non-finite verbs of the language. However, though this class displays features very similar to the constellations of the English infinitive particle 'to', its contents are still defined as nouns by Lombard (1985, p. 49) or Van Wyk et al. (1992, p. 13). The class is formed by the prefix *go* which precedes a verbal stem. Note that the prefix is not written conjunctively to the stem like in the other noun classes, but stands as a separate token, as in *go sepela* 'to walk, walking'[5] or in *go kitima* 'to run, running'. Poulos and Louwrens (1994, p. 42) mention that "the prefix has the form *go–*, and the part that follows the prefix is a stem". Lombard (1985, p. 49) solely uses some intransitive verb stems to demonstrate the infinitive.

This way of interpreting *go*, i.e. as a noun class prefix that must exclusively be followed by a verb stem, as in 4(a) (Figure 2.1 demonstrates a morphosyntactic analysis according to the

[5]There is no gerund defined for Northern Sotho, class 15 nouns however may appear as such.

definitions given by the authors mentioned above), causes a problem when transitive verbs appear with the infinitive marker *go*. The object of these verb stems could be represented by a pronominal object concord (cf. paragraphs 2.4.3 and 3.2.2), that would occur between *go* and the verb stem, as in 4 (b).

4(a)	*ba*	*rata*	*go*	*bala*	*dipuku*
	`subj-3rd-cl2`	like	to	read	books
	'they like to read books '				

(b)	*ba*	*rata*	*go*	*di*	*bala*
	`subj-3rd-cl2`	like	to	`obj-3rd-cl10`	read
	like	to	them	read	
	'they like to read them'				

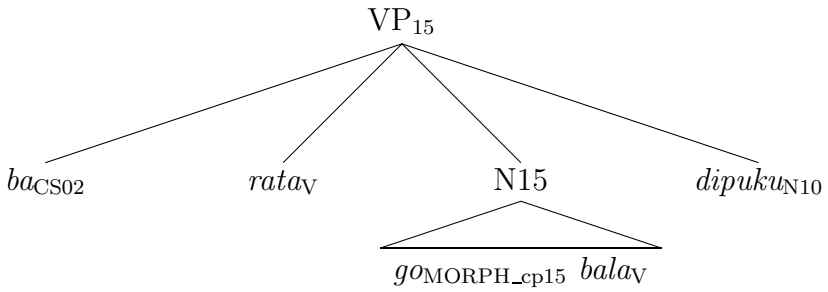

Figure 2.1: First analysis: *ba rata go bala dipuku* 'they like to read books'

Consequently, 4 (b) cannot be analysed in the same way to 4 (a), because the class 15 noun in this case would either be discontinuous or it would contain another nominal. We should like to avoid the definition of nouns of class 15 as discontinuous phrases because there is an easier solution: we suggest introducing several levels of analysis. Firstly, the agreement marker *ba* could be described on the same level as the infinitive marker *go* (see also the introduction on verbal phrases 3.2.1 on page 71). Additionally, we consider it necessary to insert a further level of analysis which follows the Head Principle[6], to name the overall

[6]Shapiro (1997) defines the Head Principle as follows: "Every phrasal category contains a head; the head and its phrasal counterparts share the same properties". As the verbal phrase embedded in the infinitive is headed by the verb stem (V), this phrase should be called VP. The properties of this VP (e.g. number or tense information) are further shared with the phrase that it is embedded in, hence the overall structure is also to be named VP (see also Chomsky's projection principle, described e.g. in Chomsky (1986)). In the case of a grammatical function, like, e.g., subject, being assigned to the constellation, it may be analysed similarly to the English gerund which in this case is treated like a nominalized verb, cf. (Bresnan, 2001, p.295f).

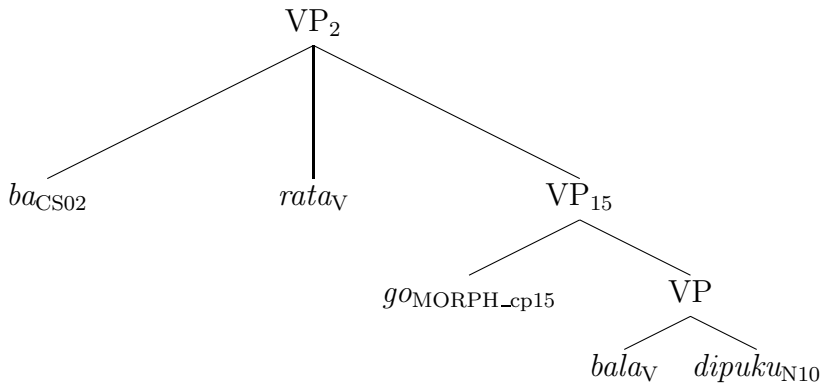

Figure 2.2: Second analysis: *ba rata go bala dipuku* 'they like to read books'

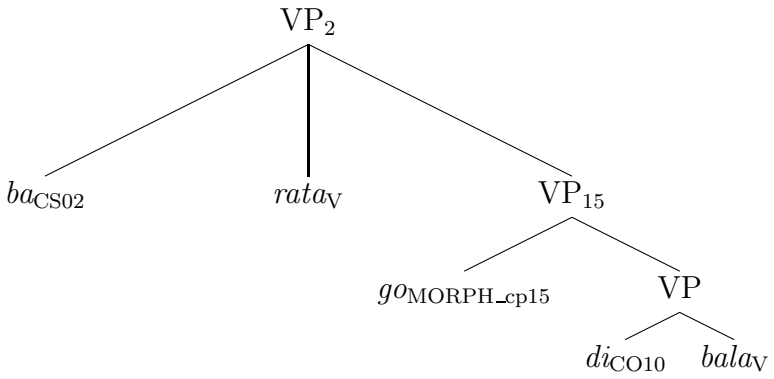

Figure 2.3: *ba rata go di bala* 'they like to read them'

structure a verbal phrase, as demonstrated in Figure 2.2[7].

Following this strategy, 4(a) and (b) can be analysed isomorphic, cf. Figure 2.3.

Our second example in (5) demonstrates the use of a qualifying adverb[8] that is – like the object concord – to be analysed on the same level of the verb stem. The infinitive particle is represented on the next higher level (cf. Figure 2.4), as before.

(5) *go botša monna fela*
to tell man only
'to tell (a) man only'

[7]Note that in section 3.2, we will describe the verbal phrases in more detail. The analyses shown here are rather approximate and only appear for the sake of demonstration.

[8]Poulos and Louwrens also mention the possibility of the infinitive being qualified by other part of speech as we will show e.g. in example 5, however, no structural analysis is given there.

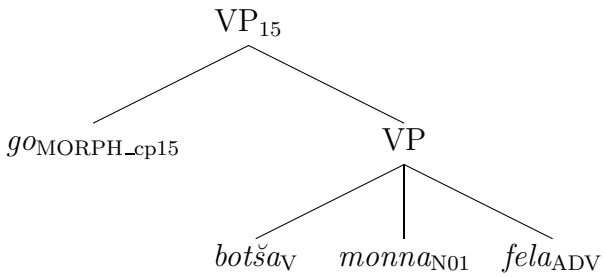

Figure 2.4: *go botša monna fela* 'to tell (the) man only'

The next example, (6), shows that negation morphemes may also occur between the two components described by the respective literature. Negation morphemes appear between the infinitive class prefix and the verb stem. Morphysyntactically, these could be treated like adverbs, i.e. we analyse them on the same level as the verb stem, cf. Figure 2.5.

(6) *go* *se* *dirwe*
to **neg** be-done
'not to be done'

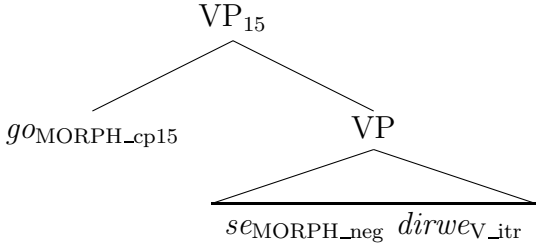

Figure 2.5: *go se dirwe* 'not to be done'

Lastly, recursive verbal phrases can also follow this class prefix like in *go ya go nyala* 'to go to marry'. In this clause, meaning 'going to marry', *go* together with the transitive verb *ya* is followed by a nested infinitive constellation, as shown in (7), illustrated in Figure 2.6.

(7) *go* *ya* *go* *nyala*
to go to marry
'going to marry'

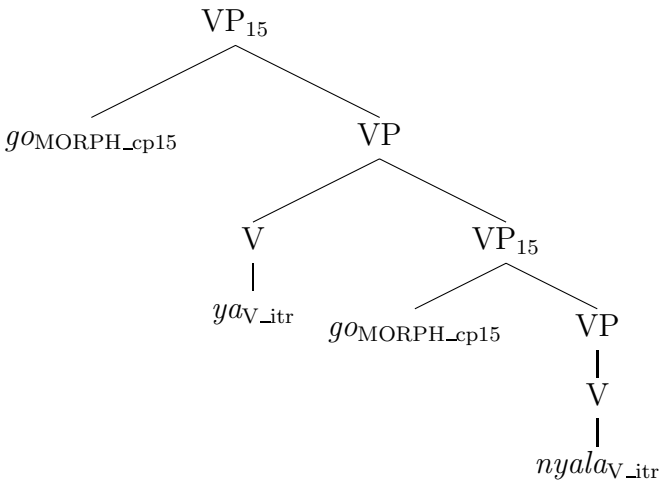

Figure 2.6: *go ya go nyala* 'going to marry'

In this study, class 15 'nouns' are therefore not considered nouns in the sense of the word class. They usually appear in embedded verbal clauses, If these infinitive verbal clauses should appear in a nominal function, they could be analysed in a similar way to the English gerund, i.e. as nominalized verbs. For details and examples, see 3.2.4 on page 92.

2.2.2.6 The locative classes 16 – 18, *N*– and *ga*–classes

In Northern Sotho, the noun classes 16 – 18, together with the *N*–, and *ga*-classes differ from all other classes by being non-productive and small, i.e. they are closed classes containing only a few nouns. All these classes behave identically from a syntactical point of view, i.e. they occur in the same environment(s) and with the same function(s). Different labels for these classes are therefore not considered necessary, all are labelled locative nouns ("NLOC"), cf. Taljard et al. (2008). Examples of such nouns are *fase* 'down, below', *morago* 'behind', or *godimo* 'high, above, in the air'.

Ziervogel (1988, p. 25) describes these nouns as "being used as adverbs"; Poulos and Louwrens (1994, p. 45) furthermore mention "adverbial [...] significance". Therefore, in the grammar described in chapter 3, the NLOC nouns (amongst other structures) will also be classified as adverbials.

2.2.2.7 Notes on the (semi-)automated identification of noun classes

A parser usually needs a lexicon containing a number of (orthographic) Northern Sotho words labelled with their part(s) of speech because the parsing rules defined for morphosyntactic analysis only describe part of speech order, rather than word order. Such a lexicon can be filled fairly easily with information on the elements of the non-productive, i.e. closed classes and some of the words of the productive classes as well. However, in any new text that is to be analysed by such software, words of the productive classes (nouns, verbs and adverbs in Northern Sotho) may appear that are not contained in the system's lexicon. As the productive classes cannot be summarised manually, an automated, or at least semi-automated methodology is necessary to identify them and label them correctly, before any morphosyntactic rule can be applied.

The tables in 2.2.1 show that the noun class system contains a number of ambiguous prefixes. Even the zero-prefix Ø– occurs in several classes (1a, 5 and 9). Automated noun- and noun class identification is therefore a far from trivial task. The prefixes of the classes 1 and 3, or 8 and 10, for example, are identical and thus any automatic determination of the correct class membership of a noun beginning with such prefixes is impossible without also taking other, i.e. lexical or contextual information into account.

To solve the problem of automatically identifying nouns and of determining their correct class, the Northern Sotho noun guesser was developed, as described in Heid et al. (2009). This guesser identifies nouns and the noun class(es) they could belong to. The noun class of a candidate is determined by looking for matching singular/plural forms and certain class-specific keywords that usually co-occur in the *University of Pretoria Sepedi Corpus* (PSC), (cf. De Schryver and Prinsloo (2000)).

However, there are a number of exceptions where such matching strategies are not successful, like *meetse* 'water', a class 6 noun that is without a singular form, or when trying to identify rare words scarcely found in corpora. Such exceptions must therefore be listed in the lexicon utilised by the analysing processes.

2.3 The pronoun

The general definition of a pronoun being a substitute for a noun or a nominal phrase is widely accepted (cf. e.g. Bußmann (2002, p. 541)) and has been assumed to be valid for

the Bantu Languages in general and Zulu in particular, cf. Doke (1954), as referenced by Wilkes (1976).

Wilkes (1976) however, observed that the words that were called pronouns by Doke are rather similar to determiners that may co-occur with nouns and that in the noun's absence acquire its status and function, as demonstrated in Wilkes (1976, p. 66). We adapt his (Zulu-) example (9) in (8), where the demonstratives (dem) *lezi-* and *leso-* appear in such determiner and also pronominal roles.

8(a) *lezi* *(-)* *zingane*
`dem` children
'these children'

leso *(-)* *sihlala*
`dem` shrub/bush
'that shrub/bush'

(b) *lezi*
`dem`
'these ones'

leso
`dem`
'that one'

Wilkes (1976, p. 77) consequently advocates a process of deletion instead of substitution:

> "Wanneer 'absolute voornaamwoorde' sonder hul antesedente optree, tree hulle inderdaad as voornaamwoorde op en is hierdie optrede ook in ooreenstemming met die opmerking wat vroeër gemaak is, naamlik dat pronominalisering in Zulu en waarskynlik die meeste ander Bantoetale, 'n pro-cum-delesi proses is."[9]

This statement is valid for Northern Sotho as well, as Van Wyk et al. (1992, p. 60) describe it (for emphatic pronouns): "they may be used with or without nouns and pronouns".

Louwrens (1991, p. 154) summarises the pronominal function in Northern Sotho as follows:

> "Strictly speaking, any linguistic element which agrees with a noun can acquire a pronominal function when that noun is deleted [...] these words do not stand in place of the deleted nouns."

[9] "When 'absolute pronouns' occur without their antecedents, they indeed occur as pronouns and such occurrence is in accordance with the remark made earlier, namely that pronominalisation in Zulu and probably in most other Bantu languages, is a pro-cum-deletion process."

The pronominal function is therefore a secondary one, fulfilled by any word or other grammatical constituent which agrees concordially with a noun. The primary function of the 'pronouns' of Northern Sotho is indeed a qualifying or determining one which is carried out when these forms appear in apposition to the nouns with which they agree (cf. Kgosana (2005, p. 18)). The secondary, pronominal function is fulfilled when these nouns are regarded as given information and thus deleted from the discourse.

Three types of pronouns are distinguished for each of the noun categories. Of these, possessive and absolute/emphatic pronouns occur in the first and second person as well, cf. $nna_{\text{PROEMPPERS\_1sg}}$ 'I' or $gago_{\text{PROPOSSPERS\_2sg}}$ 'your(s)$_{\text{sg}}$'. The third category contains the quantitative pronouns. All of these will briefly be discussed in the following paragraphs.

Note that Northern Sotho linguists usually describe demonstrative pronouns as well. However, these are categorised by Taljard et al. (2008) as having more of a concordial character, therefore the demonstrative will be attended to in paragraph 2.4.5 (on demonstrative concords).

2.3.1 The emphatic (or absolute) pronoun (PROEMP$_{\text{categ}}$)

The examples in (9) demonstrate that the absolute pronoun fulfils two pragmatic functions, "a particularisation of a nominal referent" (8a), "on the one hand, and the contrasting thereof, on the other" (8b), as Louwrens (1991, p. 103) states. Both functions are emphatic in nature, therefore Taljard et al. (2008) describe this pronoun as emphatic ("PROEMP"). Information on the class is then added to the name on the first of the two levels of annotation. The information on person is added on the second level, as shown in Table 2.5.

9(a) *dimpša*$_{\text{N10}}$ *tše*$_{\text{PROEMP10}}$
dogs `emp-3rd-cl10`
'these dogs'

(b) *tše*$_{\text{PROEMP10}}$ *dimpša*$_{\text{N10}}$
`emp-3rd-cl10` dogs
'these (specific) dogs (not the others)'

In Table 2.5, the pairs of classes 4/9, 8/10 and 15/LOC are homographous, i.e. if the noun to which they refer is missing, it is not usually possible to identify the correct class without taking the context into account. A hybrid disambiguation process that utilises rule-based

procedures combined with the utilization of a statistical tagger has been proposed to solve this problem, compare e.g. Prinsloo and Heid (2005) in this respect.

2.3.2 The possessive pronoun (PROPOSS$_{\text{categ}}$)

Van Wyk et al. (1992, p. 64) state that any noun describing a possessor can be replaced by a possessive pronoun. Poulos and Louwrens (1994, p. 90 et seq.) categorise this POS together with others as 'qualificative'. However, it is the only (anaphoric) pronoun that never occurs in apposition to the noun and that therefore demonstrates an exception to the rule stated in the introduction of this paragraph: This pronoun substitutes a noun to which it refers anaphorically, as in (10).

It would appear that the only possessive pronouns available in Northern Sotho refer to the first and second person, the classes 01 and 02, with only a few others existing where, according to Poulos and Louwrens (1994, p. 90), "the possession is owned by a family or a community as a whole". To cover the other classes (02 - 10, 14, and LOC), the emphatic pronouns are used. For the purpose of this study, emphatic pronouns used in texts as possessive pronouns, are labelled appropriately as possessive pronouns, cf. Table 2.6, (compare Table 2.5 with Table 2.6 in this respect).

(10) *mogopolo*$_{\text{N03}}$ *wa*$_{\text{CPOSS03}}$ *gago*$_{\text{PROPOSSPERS\_2sg}}$
idea of `poss-2nd-sg`
'your idea'

2.3.3 The quantitative pronoun (PROQUANT$_{\text{categ}}$)

We refer to Poulos and Louwrens (1994, p. 78 et seq.) when categorising quantitatives as pronouns. Only one pronoun stem exists, *-ohle* 'the whole of, all', it appears in different surface forms depending on the class that it occurs in. As shown in Table 2.7, these forms are the result of a concordial element being prefixed to the stem, similar to the emphatic and possessive pronouns.

Table 2.5: The emphatic pronouns

Noun class	**Annotation**	**Emphatic pronouns**
(pers)	PROEMPPERS_1sg	*nna*, *nnaena*
	PROEMPPERS_2sg	*wena*, *wenaena*
	PROEMPPERS_1pl	*rena*, *renaena*
	PROEMPPERS_2pl	*lena*
01	PROEMP01	*yena*, *yenaena*
02	PROEMP02	*bona*, *bobona*
03	PROEMP03	*wona*, *wonaona*
04	PROEMP04	*yona*
05	PROEMP05	*lona*
06	PROEMP06	*ona*
07	PROEMP07	*sona*
08	PROEMP08	*tše*, *tšona*
09	PROEMP09	*yona*
10	PROEMP10	*tše*, *tšona*
14	PROEMP14	*bjona*
15	PROEMP15	*gona*
LOC	PROEMPLOC	*gona*, *gonaena*

Table 2.6: The possessive pronouns

Noun class	**Annotation**	**Possessive pronouns**
(pers)	PROPOSSPERS_1sg	*ka*
	PROPOSSPERS_2sg	*gago*, *nago*
	PROPOSSPERS_1pl	*rena*, *gešo* 'our families'
	PROPOSSPERS_2pl	*lena*, *geno* 'your families'
01	PROPOSS01	*gagwe*
02	PROPOSS02	*bona*, *gabo* 'their families'
03	PROPOSS03	*wona*, *wonaona*
04	PROPOSS04	*yona*
05	PROPOSS05	*lona*
06	PROPOSS06	*ona*
07	PROPOSS07	*sona*
08	PROPOSS08	*tšona*
09	PROPOSS09	*yona*
10	PROPOSS10	*tšona*
14	PROPOSS14	*bjona*
15	PROPOSS15	*gona*
LOC	PROPOSSLOC	*gona*

Table 2.7: The quantitative pronouns

Noun class	**Annotation**	**Quantitative pronouns**
01	PROQUANT01	*yohle*
02	PROQUANT02	*bohle*
03	PROQUANT03	*wohle*
04	PROQUANT04	*yohle*
05	PROQUANT05	*lohle*
06	PROQUANT06	*ohle*
07	PROQUANT07	*sohle*
08	PROQUANT08	*tšohle*
09	PROQUANT09	*yohle*
10	PROQUANT10	*tšohle*
04	PROQUANT14	*bjohle*
15	PROQUANT15	*gohle*
LOC	PROQUANTLOC	*gohle*

2.4 The concords

2.4.1 Introduction

The term 'concord' usually is associated with agreement, sometimes even used synonymously (cf. Corbett (2001)). In Northern Sotho, the term concord refers 'to a structural element [...] which formally marks the relationship between a noun and other words in a sentence' (as defined by Louwrens (1994, p. 30)).

Orthographically, concords appear with few exceptions as standalone words in Northern Sotho. For the purpose of this study, subject concords, object concords, possessive concords and demonstrative concords are distinguished. They all generally agree with the noun class of the word they refer to. A concord can be categorised as a morpheme, however, when occurring as part of a verb, it appears with an explicit function different from the other morphemes of Northern Sotho, namely to guarantee agreement with a nominal that the verb refers to. Secondly, like the pronouns, they can acquire a pronominal function whenever this nominal is omitted (cf. (Louwrens, 1991, p. 154)).

2.4.2 The subject concord ($\mathbf{CS}_{\text{categ}}$)

The subject concord is the part of the verb that links it to its subject, usually a noun. As nouns appear in noun classes, there are concord forms available for each noun class. Additionally, the neutral form *e* and the indefinite form *go* can occur. The neutral concord *e* is used when the (usually anaphoric) relationship to the referent cannot be established, like in (11).

(11)

Aowa$_{\text{INT\_neg}}$,	*le*$_{\text{PART\_con}}$	*yo*$_{\text{CDEM01}}$	*e*$_{\text{CSNEUT}}$	*sego*$_{\text{VCOP\_neg-rel}}$	*morutiši*$_{\text{N01}}$
No,	`con`	`dem-3rd-cl01`	`subj-neut`	which-is-not	teacher

'No, and that one which is not (a) teacher'

The indefinite subject concord *go* on the other hand, does not refer to a specific nominal. Instead, it marks cases where such a referent does not exist, i.e. the indefinite case, like in example (12) taken from (Lombard, 1985, p. 102). Therefore, this subject concord never co-occurs with a subject noun.

(12)

go$_{\text{CSINDEF}}$	*a*$_{\text{MORPH\_pres}}$	*fiša*$_{\text{V\_itr}}$
`subj-indef`	`pres`	is hot

'it is hot'

Different sets of subject concords exist for the same class, as demonstrated in Table 2.8. As will be described in more detail in paragraph 3.2, verbs occur in different moods. Every mood appears in its own morphosyntactics and makes use of one of the sets. For example, a present tense (imperfect) positive mood with a class 1 subject will only make use of the class 1 subject concord *o*, while other moods will select *a* only, as Poulos and Louwrens (1994, p. 170) state.

Poulos and Louwrens (1994, p. 168) and Lombard (1985, p. 152) all mention two sets of subject concords, however the first of these sets uses either *a* or *o* as the respective class 1 members of their set 1. In other words, certain moods select *a*, while others select *o*, these concords are hence not interchangeable. Instead of summarising them in one set, we will define two sets, set 1 and 2 which only differ in the class 1 subject concord. This methodology will allow us to define appropriate morphosyntactic rules describing these moods (cf. paragraph 3.2, page 71 et seq.).

Set 2 of Poulos and Louwrens (1994, p. 168) describing the consecutive set (also described in (Lombard, 1985, p. 153)) is therefore named set 3 in this work.

As mentioned above, homography is a common property of the closed word classes of Northern Sotho. The class pairs 1 and 3, 4 and 9, 8 and 10, and also 15 and LOC of most sets use the same subject concords and in some cases, this ambiguity of subject concords cannot be resolved on the word level. Parsing (as described in paragraph 4.2 on page 192) might in some cases be necessary, see the discussion in (Faaß et al., 2009).

Table 2.8: The three sets of subject concords

	subject concords			
	set 1	**set 2**	**set 3**	**fused forms**
categ	$1CS_{categ}$	$2CS_{categ}$	$3CS_{categ}$	
...PERS_1sg	*ke*	*ke*	*ka*	$ke_{CSPERS} + ka_{MORPH\_pot} \rightarrow nka$
...PERS_2sg	*o*	*o*	*wa*	
...PERS_1pl	*re*	*re*	*ra*	
...PERS_2pl	*le*	*le*	*la*	
...01 (incl.01a)	*o*	*a*	*a*	
...02 (incl.02b)	*ba*	*ba*	*ba*	
...03	*o*	*o*	*wa*	
...04	*e*	*e*	*ya*	
...05	*le*	*le*	*la*	
...06	*a*	*a*	*a*	
...07	*se*	*se*	*sa*	
...08	*di*	*di*	*tša*	
...09	*e*	*e*	*ya*	
...10	*di*	*di*	*tša*	
...14	*bo*	*bo*	*bja*	
...15	*go*	*go*	*gwa*	
...LOC	*go*	*go*	*gwa*	
...NEUT	*e*	*e*	*ya*	
...INDEF	*go*	*go*	*gwa*	

2.4.3 The object concord (CO_{categ})

If an object (of a verb) is not found in its designated post-verbal position, i.e. if it is either known and therefore not mentioned (omitted) or moved to another position in the sentence (for example when being topicalised), an object concord is to be inserted, cf. (Van Wyk et al., 1992, p. 25). The function of object concords is therefore pronominal in the traditional sense of the word as they substitute an omitted or moved noun. There are however cases where an object noun is present, though the object concord is present as well, like in 13 (a). Note that in such cases it will still be the object concord that represents the functional object of the verb, while the noun is seen as adjunctive to the clause, i.e. it may be left out, as in 13 (b).

13 (a) *ke*$_{\text{CSPERS}_{1sg}}$ *mo*$_{\text{CO01}}$ *thušitše*$_{\text{V}}$ *mosadi*$_{\text{N01}}$ *yo*$_{\text{CDEM01}}$
`subj-1st-sg` `obj-3rd-cl1` helped woman `dem-3rd-cl01`
'I helped her, this woman'

(b) *ke*$_{\text{CSPERS}_{1sg}}$ *mo*$_{\text{CO01}}$ *thušitše*$_{\text{V}}$
`subj-1st-sg` `obj-3rd-cl1` helped
'I helped him/her'

The pronominal use of object concords could actually suggest their categorisation as pronouns. However, unlike pronouns, concords are bound morphemes and as such part of the verb (Van Wyk et al., 1992, p- 25). Some (proclitic) object concords even fuse with the verb stem and are thereby causing changes in the morpho-phonology and hence the orthography of its root, like in *mpona* 'see him/her', a fused form of *mo* 'him/her' + *bona* 'see'. Note that most of the object concords are homographous with their subject concord counterparts. An overview of the object concords is shown in Table 2.9.

2.4.4 The possessive concord ($CPOSS_{categ}$)

Lombard (1985, p. 172 et seq.) refers to "possessive particles" when describing these elements linking a possession with its possessor. The possessive concord is a bound morpheme and like most other concords, it is written separately. It refers anaphorically to the possession and hence appears in the same noun class. In some cases, the possessive refers not to a possession, but describes a more general relation between the participants, like in *mosadi*$_{\text{N01}}$ *wa*$_{\text{CPOSS01}}$ *Piti*$_{\text{N01a}}$ '(the) woman/wife of Peter', or in *leoto*$_{\text{N05}}$ *la*$_{\text{CPOSS05}}$ *tafola*$_{\text{N09}}$ '(a) leg of (a) table'. If the word describing the possession should be ommitted, this concord acquires its grammatical function, as in *ba*$_{\text{CPOSS02}}$ *Piti*$_{\text{N01a}}$ *ba*$_{\text{1CS02}}$ *adima*$_{\text{V}}$ *tšhelete*$_{N09}$ '(the) ones

Table 2.9: The object concords

class	annotation	object concord	comments
(pers)	COPERS_1sg	*N–*	occurs as fused form only
	COPERS_2sg	*go*	
	COPERS_1pl	*re*	
	COPERS_2pl	*le*	
01 (01a)	CO01	*mo*	fused forms possible
02 (02b)	CO02	*ba*	
03	CO03	*o*	
04	CO04	*e*	
05	CO05	*le*	
06	CO06	*a*	
07	CO07	*se*	
08	CO08	*di*	
09	CO09	*e*	
10	CO10	*di*	
14	CO14	*bo*	
15	CO15	*go*	
LOC	COLOC	*go*	

of Peter (i.e. Peter's children) borrow money'. Table 2.10 lists all possessive concords of Northern Sotho.

2.4.5 The demonstrative concord ($\text{CDEM}_{\text{categ}}$)

Modern English demonstratives or deictic determiners describe two distances, zero and non-zero, appearing as 'this'(sg.)/'these'(pl.) and 'that'(sg.)/'those'(pl.). Northern Sotho on the other hand marks four distances, which Van Wyk et al. (1992, p. 37) explains by using the translations: 'this/these (here)', 'this/these (next to)', 'that/those (there)' and 'that/those (yonder)'. If the noun that such a demonstrative refers to is omitted, the demonstrative acquires a pronominal function, e.g. in bao_{CDEM02} ba_{CS02} $re_{\text{V\_tr}}$ $\text{gore}_{\text{CONJ}}$ ...) 'those (people) say that ...'.

One should note that these deictic concords are not only used in order to point the listener to a physical location but also to a point in time relative to the current time. Van Wyk et al. (1992, p. 37) list the example *ditiro tšela* 'those deeds (which happened long ago)'.

Table 2.10: The possessive concords

class	annotation	possessive concord
01 (01a)	CPOSS01	*wa*
02 (02b)	CPOSS02	*ba*
03	CPOSS03	*wa*
04	CPOSS04	*ya*
05	CPOSS05	*la*
06	CPOSS06	*a*
07	CPOSS07	*sa*
08	CPOSS08	*tša*
09	CPOSS09	*ya*
10	CPOSS10	*tša*
14	CPOSS14	*bja*
15	CPOSS09	*ga*
LOC	CPOSS09	*ga*

Taljard et al. (2008) use the label Concord DEMonstrative ("CDEM") to signal a less pronominal, but more of a concordial function of the demonstrative. Table 2.11 summarises this pronoun/concord and is taken partially from Lombard (1985, p. 37–38), partially from our system's lexicon.

2.4.6 The demonstrative copulative (CDEMCOP$_{\text{categ}}$)

The character of this demonstrative is two-fold. Lombard (1985, p. 163 et seq.) mainly describes its predicative capacity, while Poulos and Louwrens (1994, p. 87, 90) refer to it generally as a deictic expression, a "predicative form of a demonstrative". Taljard et al. (2008) describe it as a demonstrative concord with an added copulative function ("CDEM-COP", Concord DEMonstrative COPulative), as in 14 (a).

Ziervogel (1988, p. 83) explains that its English equivalents are 'here he/she/it is; here they are', and that it describes three positions relative to the speaker, similar to the ordinary demonstrative. All forms of the demonstrative copulative begin with *še-*, followed by a pronominal root. They can also be used without a complement, as in 14 (b).

14(a) *šeba*$_{\text{CDEMCOP\_02}}$ *bašemane*$_{\text{N02}}$
here-is-`obj-3rd-cl2` boys
'here are the boys'

Table 2.11: The demonstrative concords (pronouns)

class	**annotation**	**demonstrative concord**			
		'here'	'here (next to)'	'there'	'yonder'
01 (01a)	CDEM01	*yo*	*yono*, *yokhwi*	*yoo*, *youwe*, *yowe*	*yola*
02 (02b)	CDEM02	*ba*	*bano*, *bakhwi*	*bao*, *bauwe*, *bawe*	*bale*
03	CDEM03	*wo*	*wono*, *wokhwi*	*woo*, *wouwe*, *wowe*	*wola*
04	CDEM04	*ye*	*yeno*, *yekhwi*	*yeo*, *yeuwe*, *yewe*	*yela*
05	CDEM05	*le*	*leno*, *lekhwi*	*leo*, *leuwe*, *lewe*	*lela*
06	CDEM06	*a*	*ano*, *akhwi*	*ao*, *auwe*, *awe*	*ale*
07	CDEM07	*se*	*seno*, *sekhwi*	*seo*, *seuwe*, *sewe*	*sela*
08	CDEM08	*tše*	*tšeno*, *tšekhwi*	*tšeo*, *tšeuwe*, *tšewe*	*tšela*
09	CDEM09	*ye*	*yeno*, *yekhwi*	*yeo*, *yeuwe*, *yewe*	*yela*
10	CDEM10	*tše*	*tšeno*, *tšekhwi*	*tšeo*, *tšeuwe*, *tšewe*	*tšela*
14	CDEM14	*bjo*	*bjono*, *bjokhwi*	*bjoo*, *bjouwe*, *bjowe*	*bjola*
15	CDEM15	*mo*	*mono*	*moo*	*mola*
LOC	CDEMLOC	*fa*	*fano*	*fao*, *fauwe*, *fawe*	*fale*
		mo	*mono*, *mokhwi*	*moo*, *mouwe*, *mowe*	*mola*
		šifa			

14(b) *šeba*$_{\text{CDEMCOP\_02}}$
here-is-`obj-3rd-cl2`
'here they are'

All demonstrative copulatives are listed in Table 2.12 on page 46, of which the data in columns 3 to 6 are copied from Poulos and Louwrens (1994, p. 88). As with the concords described in the previous paragraphs, the class 4/9, 8/10 and 15/LOC show homographous forms.

2.5 The adjective ($\mathbf{ADJ}_{\text{categ}}$)

The Northern Sotho adjective represents an unproductive class, i.e. one can list all of its elements in a lexicon. Table 2.13 on page 48 shows some examples.

Some linguists do not classify adjectives as a word class of Northern Sotho at all, Van Wyk et al. (1992, p. 73) name this category "adjectival noun", because of its nominal character. Another reason is inter alia explained by Poulos and Louwrens (1994, p. 91): adjectives are to be preceded by an "adjectival concord". This concord is called "qualificative particle"

Table 2.12: The demonstrative copulatives and their variants

class	annotation	demonstrative copulative			
		'here'	'here (next to)'	'there'	'yonder'
01 (01a)	CDEMCOP_01	*šo*	*šono*	*šoo*	*šola*, *šole*
02 (02b)	CDEMCOP_02	*šeba*	*šebano*	*šebao*	*šebala*, *šebale*
03	CDEMCOP_03	*šo*	*šo*	*šoo*	*šola*, *šole*
04	CDEMCOP_04	*še*	*šeno*	*šeo*	*šela*, *šele*
05	CDEMCOP_05	*šele*	*šeleno*	*šeleo*	*šelela*, *šelele*
06	CDEMCOP_06	*šea*	*šeano*	*šeao*	*šeala*, *šeale*
07	CDEMCOP_07	*sese*	*seseno*	*seseo*	*sesela*, *sesele*
08	CDEMCOP_08	*šidi*	*šidino*	*šidio*	*šidila*, *šidile*
09	CDEMCOP_09	*še*	*šeno*	*šeo*	*šela*, *šele*
10	CDEMCOP_10	*šidi*	*šidino*	*šidio*	*šidila*, *šidile*
14	CDEMCOP_14	*šebo*	*šebono*	*šeboo*	*šebola*, *šebole*
15	CDEMCOP_15	*šefa*	*šefano*	*šefao*	*šefala*, *šefale*
LOC	CDEMCOP_LOC	*šefa*	*šefano*	*šefao*	*šefala*, *šefale*

or "relative pronoun" by Lombard (1985, ibid.) and Van Wyk et al. (1992, ibid.). For its concordial properties, Ziervogel (1988, p. 55) and Taljard et al. (2008) on the other hand classify this POS as a demonstrative concord ("CDEM"). These concords are listed in the third column of Table 2.11.

A Northern Sotho adjective is thus formed by such a demonstrative concord and another element labeled ADJ which is described by some linguists as a noun, as it makes use of a noun class prefix and may appear in several such classes. Seeing that adjectives may replace the nouns that they refer to in noun phrases (cf. paragraph 3.8.4.2) which counts for nominal characteristics, we will therefore treat any units labeled 'ADJ' as similar to a noun. Semantically, these elements all indicate a property, comparable to properties described by adjectives of other languages (number, colour, and others, e.g. properties of size, like *–golo* 'big' or *–nyane* 'small'), cf. Table 2.13 for examples. However there are also a number of nouns which may fall into this category, like the locative nouns, e.g. $pele_{\text{NLOC}}$ 'first' or $godimo_{\text{NLOC}}$ 'high, above'. A number of nouns of class 14 can also be interpreted as having an adjectival content, e.g. $bohlale_{\text{N14}}$ 'clever', or $boleta_{\text{N14}}$ 'kind/soft' proving that the Northern Sotho word categories are rather based on the morphological structure of an element rather than its semantic content (cf. Louwrens (1991, p. 4) describing Doke's

classifications).

2.6 The enumerative (ENUM)

The term "enumerative" as it is used by linguists should not be confused with its usages in other areas such as mathematics, where it refers to counting or the exhaustive listing of objects. Instead, Poulos and Louwrens (1994, p. 112 et seq.) refer to it generally as a qualificative, because it consists of a concord agreeing with a noun and a stem.

As such, enumeratives show a number of properties suggesting them to be similar to adjectives. ENUM is a closed class containing only a few stems (*–šele* 'different, foreign, strange', *–tee* 'one', and *šoro* 'cruel'). Poulos and Louwrens (1994, ibid.) also list *–fe* 'which', which we however categorise as a question word (cf. Table 2.19). All enumeratives are preceded by a demonstrative concord. In this study we therefore follow Lombard (1985, p. 58–59) who classifies enumeratives as adjectives.

Table 2.13: Some examples of frequently used adjectives

root	**class** 01+03	02	04	05	06	07	08	09	10	14	15+LOC
–bedi 'two'		*babedi*	*mebedi*		*mabedi*		*(di)pedi*		*(di)pedi*		
–hlano 'five'		*bahlano*	*mehlano*		*mahlano*		*tlhano hlano*		*tlhano hlano*		
–ngwe 'another'	*mongwe*	*bangwe*	*mengwe*	*lengwe*	*mangwe*	*sengwe*	*(di)ngwe*	*ngwe*	*(di)ngwe*	*bongwe*	*gongwe*
–golo 'big'	*mogolo*	*bagolo*	*megolo*	*legolo*	*magolo*	*segolo*	*(di)kgolo*	*kgolo*	*(di)kgolo*	*bogolo*	*gogolo*
–nyane 'small'	*monyane*	*banyane*	*menyane*	*lenyane*	*manyane*	*senyane*	*(di)nnyane dinyane*	*nnyane*	*(di)nnyane dinyane*	*bonyane*	
–tala 'old'	*motala*	*batala*	*metala*	*letala*	*matala*	*setala*	*(di)tala*	*tala*	*(di)tala*	*botala*	
–so 'black'	*moso*	*baso*	*meso*	*leso*	*maso*	*seso*	*ntsho*	*ntsho*	*ntsho*	*boso*	
–be 'bad'	*mobe*	*babe*	*mebe*	*lebe*	*mabe*	*sebe*	*mpe*	*mpe*	*mpe*	*bobe*	*bobe*
–kae 'how much'	*mokae*	*bakae*	*mekae*	*lekae*	*makae*	*sekae*	*(di)kae*	*kae*	*(di)kae*	*bokae*	*gokae*
–bose 'sweet, nice'	*mobose* (cl. 3 only)	*babose*	*mebose*	*lebose*	*mabose*	*sebose*	*(di)bose*	*bose*	*(di)bosese*	*gobose*	

2.7 The verb stem (V)

2.7.1 Introduction

Northern Sotho is a disjunctively written language, which is particularly clearly demonstrated by its verbs. A Northern Sotho verb usually consists of several orthographic words (all of which are categorised as equal elements in the set of POS defined by Taljard et al. (2008), i.e. independent of their status as independent or dependent morpheme). Our word class 'V' should therefore not be confused with verbs of other languages, as it represents the verb stem only. Example (15) is taken from Van Wyk et al. (1992, p. 55) to demonstrate a complete verb of Northern Sotho. This verb consists of a subject concord of the 2nd person plural, followed by a progressive morpheme, an object concord referring to an object of class 9, followed finally by the verb stem.

(15)

le$_{\text{CSPERS}_2\text{pl}}$	*sa*$_{\text{MORPH\_prog}}$	*e*$_{\text{CO09}}$	*nyaka*$_{\text{V}}$
`subj-2nd-pl`	still	`obj-3rd-cl9`	want
you(pl)	still	it	want

'(all of) you still want it'

Note that in paragraph 3.2.1.7 we will reflect further on this part of speech, because due to its syntactic appearance as a head of verbal phrases, it will be necessary to add an additional level of annotation.

2.7.2 Notes on some verbal suffix clusters

As stated in the previous paragraph, the part of speech category 'V' describes verb stems only. A complete or so-called linguistic verb of Northern Sotho hence consists of a number of prefixal, infixal and suffixal morphemes. Some of the pre- and infixal morphemes are written separately, like *tlo*$_{\text{MORPH\_fut}}$ '`fut`', indicating future tense, others are fused with the verb stem, like *N*-$_{\text{COPERS\_1sg}}$ (`obj-1st-sg`), representing an object of the verb (1st person singular) and being fused with. The morphemes we will examine in this paragraph are all suffixal and always fused with the verb stem.

There are a number of issues concerning verb stems and verbal morphemes, however, we will concentrate on the few which are important for an identification of the morphosyntactic constellations as described in the next chapter:

- The verbal endings
- The perfect/past tense extensions

- The plurality marker or plural morpheme
- The relative suffixes

Verb stems of Bantu languages appear in a large number of derivations. Table 2.14 (on page 54) is an excerpt of Prinsloo et al. (2008, Table 2) who refer to Ziervogel and Mokgokong's Groot Noord-Sotho Woordeboek (GNSW) (Ziervogel and Mokgokong (1975)). It demonstrates that Northern Sotho verb stems are morphologically complex units. Altogether 350 possible suffixes and suffix clusters have been identified according to Prinsloo et al. (2008) so far. For the purpose of this study, these suffixes are not analysed, cf. Anderson and Kotzé (2006) for a discussion.

Northern Sotho verb stems usually end in *–a* or *–e*. These endings will be repeatedly mentioned in paragraphs 3.2.2 et seq., therefore we describe them in more detail here.

A basic, positive verb usually ends in *-a*, for example $rek_{\text{root}}–a_{\text{Vend}}$ 'buy', while *–e* can (together with the negative morphemes *ga se*) indicate a negation, like in *ga se ...* $rek_{\text{root}}–e_{\text{Vend}}$. However, the verbal ending *-e* can not only be one marker of a negation, it can also be the final element of the past tense forms. In this case, the past tense morpheme *-il-* is inserted between the root and this ending, like in $rek_{\text{root}}–il–_{\text{past}}–e_{\text{Vend}}$. This past tense morpheme appears in a number of allomorphs, Van Wyk et al. (1992, p. 47 et seq.) mention morpho-phonetic rules ("sound changes") as being the reason for this phenomenon. Some examples of the many allomorphs of *-il-* are demonstrated in 16, (a) to (d).

16 (a) Rule: ROOT-*tša* + *-ile* → *ditše*
bitša 'call' + *-il-* + *-e* → *biditše*

(b) Rule: ROOT-*nya* + *-ile* → *ntše*
senya 'ruin/destroy' + *-il-* + *-e* → *sentše*

(c) Rule: ROOT-*la* + *-ile* → *tše*
bula 'open' + *-il-* + *-e* → *bitše*

(d) Rule: ROOT-*sa* + *-ile* → *sitše*
lesa 'let loose/free' + *-il-* + *-e* → *lesitše*

Another ending will be mentioned in the next chapter (cf. 3.2.3): the plurality marker *-ng* which is suffixed to the verbal ending; it usually appears in an imperative (cf. paragraph 3.2.3), when more than one person is addressed, like in *Tšhabang!* 'Get out of the way!' (literally: $\text{you}_{\text{plural}}$ must get out of the way).

The relative suffix is usually fused with the verbal ending, too. It appears as one of the

allomorphs *-go* or *-ng*. However, the relative mood (cf. paragraph 3.2.7) is not only marked by a relative suffix (which might also be fused with the future tense morpheme), this constellation requires a subject concord and a demonstrative concord to appear as well, like in (17), taken from Van Wyk et al. (1992, p. 86).

(17) *lesogana* *le* *le* *segago* ...
young man `dem-3rd-cl5` `subj-3rd-cl5` laugh-`rel` ...
'(a) young man who is laughing ...'

Note that in chapter 3, rules describing verbal endings will not refer to the surface form of a verbal stem, but to the morphemes described here. For example, whenever a rule will be mentioned that entails a restriction like e.g. 'Vstem ends in *-a, -ang*', still there might be verb stems processable by this rule that do not appear with such endings on the surface. A fairly frequent verb can serve as an example to demonstrate this issue, the "defective" verb re_{V} 'say' which qualifies for all rules requiring a verb ending in *-a*. Other verbs, like the verbs that entail a "state of completion" (as Lombard (1985, p. 49) describes them) appear with a perfect tense extension but indicate a (currently active) state, like, e.g. *dutše* '$\mathrm{sit}_{\mathrm{pres}}$' (comparable to a present continuous tense).

As will be explained in more detail in paragraph 5.1.2.3, we will lexicalise information on the ending of a verb stem with a specific parameter, i.e. we will add a parameter 'verbal ending' to each lexicon entry describing a verb. A parser processing a condition on verbal suffix clusters will not check the verb's ending as such, but this parameter instead. As verbs like *re* or *dutše* will then be described as ending in *-a*, a parser will handle them like other legitimate verbs with this ending, not considering the surface ending with which they appear.

2.7.3 The auxiliary verb (V_aux)

Louwrens (1991, p. 19) defines an auxiliary word group as "a word group of which the first member is an auxiliary verb. The word or word group which follows the auxiliary verb is referred to as the complement". Like the verb stem V, the auxiliary verb stem V_aux presents only a part of the auxiliary verb. This verb is similar to main verbs as it also contains a subject concord and possibly tense markers. There is however also a significant difference found between auxiliary verbs and main verbs: Poulos and Louwrens (1994, p. 276) state that auxiliary verbs do not take objects.

The auxiliary verb as such is usually defined as part of a word group (consisting of the auxiliary verb and its verbal complement). This definition makes it different from e.g. the morpheme *tlo* which may only be a part of a verb (cf. the paragraph on morphemes, 2.9).

Unlike other languages, where auxiliaries are contained in a closed set of few words, Northern Sotho offers a variety of verb stems that can be used as auxiliaries. In this case, according to Poulos and Louwrens (1994, p. 273) "their basic meanings very often change somewhat and take a related figurative meaning", like *šetše* 'remain, stay', which, when appearing as an auxiliary, means 'already'.

Table 2.15 (on page 55) shows some auxiliaries, Table 2.16 (on page 56) some examples of main verbs used as auxiliaries[10].

2.7.4 The copulative (VCOP)

There are three categories of copulatives: identifying, descriptive and associative copulative. As Lombard (1985, p. 192 et seq.) states, these three groups (which will be defined in more detail in paragraph 3.3) are named on semantic grounds. Identifying copulatives give two elements equality, descriptive copulatives describe one element with another word or phrase, and associative copulatives relate one element with another, in the sense of the English 'to be with'. Syntactically, these three groups can each be divided into two subcategories, namely stative and inchoative.

Only few copulas exist in Northern Sotho that are not homographous with other parts of speech. The subject concords (described in Table 2.8 on page 41), for example, can appear with an additional copulative sense, like the subject concord $re_{\text{CSPERS\_1pl}}$ 'we' that can also occur as a copula $re_{\text{VCOP\_1pl}}$ 'we are' as shown in 18, (a) and (b). Note that subject concords occurring as copulas are to be labeled appropriately, i.e. they are not labeled as CS_categ, but as VCOP_categ.

18(a) $re_{\text{CSPERS\_1pl}}$ $reka_{\text{V}}$ $dijo_{\text{N10}}$
`subj-1st-pl` buy food
'we buy food'

(b) $re_{\text{VCOP\_1pl}}$ $barutiši_{\text{N02}}$
`subj-1st-pl-cop` teachers
'we are teachers'

[10]Note that a brief description of auxiliary verbal phrases can be found in paragraph 3.4.

Table 2.17 (on page 56) offers a brief overview of some copula of Northern Sotho. Detailed information on all copulative constellations will be given in section 3.3 (see also Prinsloo (2002) for schematic example driven representations of copulatives in Northern Sotho). Some of the copulatives have to agree with subject nouns, they contain class information which is to be added to the label. Others are negated and as will be shown in paragraph 3.3, no other negation marker then appears in the verbal phrase that contains them. For a correct syntactic analysis, these copulatives are marked 'neg' on a second level annotation.

Table 2.14: Derivations of the verb *gadika* 'roast, thrash' in GNSW

1 Suffix -A root + standard (std) modifications				
Structure	ROOTa	ROOTile	ROOTwa	ROOTilwe
Grammatical formula	VR	VRPer	VRPas	VRPerPas
Example	*gadika*	*gadikile*	*gadikwa*	*gadikilwe*
Translation	roast/thrash	roasted/thrashed	be roasted/thrased	was/were roasted/thrashed
2 Suffix -ANA root + reciprocal + std modifications				
Structure	ROOTana	ROOTane	ROOTanwa	ROOTanwe
Grammatical formula	VRRec	VRRecPer	VRRecPas	VRRecPerPas
Example	*gadikana*	*gadikane*	*gadikanwa*	*gadikanwe*
Translation	roast/thrash each other	roasted/thrashed each other	(theoretical form)	(theoretical form)
3 Suffix -EGA root + neutro passive+ std modifications				
Structure	ROOTega	ROOTegile		
Grammatical formula	VRNPas	VRNPasPer	(VRNPasPas)	(VRNPasPerPas)
Example	*gadikega*	*gadikegile*		
Translation	be roasted/thrashed	was/were roasted/thrashed		
4 Suffix -ELA root + applicative + std modifications				
Structure	ROOTela	ROOTetše	ROOTelwa	ROOTetšwe
Grammatical formula	VRApp	VRAppPer	VRAppPas	VRAppPerPas
Example	*gadikela*	*gadiketše*	*gadikelwa*	*gaditketšwe*
Translation	roast for	roasted for	be roasted for	was/were roasted for
5 Suffix -ELANA root + applicative + reciprocal + std modifications				
Structure	ROOTelana	ROOTelane	ROOTelanwa	ROOTelanwe
Grammatical formula	VRAppRec	VRAppRecPer	VRAppRecPas	VRAppRecPerPas
Example	*gadikelana*	*gadikelane*	*gadikelanwa*	*gaditkelanwe*
Translation	roast for each other	roasted for each other	be roasted for each other	was/were roasted for each other

Table 2.15: Examples of auxiliary verbs

V_aux	**tense**	**Translation**	**Example of use**
ba	n.a.	'furthermore', 'and so'	ba_{1CS02} ba_{V_aux} ba_{2CS02} $kitimela_{V}$ 'and so they ran' (De Schryver, 2007, p. 8)
be	past	'did/was/were'	o_{1CS01} be_{V_aux} a_{2CS01} sa_{MORPH_neg} $tsebe_{V}$ $gore_{CONJ}$ ke_{VCOP_1sg} $gona_{PROEMPLOC}$ 'he didn't know that I was here' (Louwrens, 1991, p. 52)
bego	past	'who did/was/were'	ba_{1CS02} $bego_{V\_aux}$ ba_{2CS02} $bolela_{V}$ 'those who were talking'
napile	n.a.	'then, subsequently, afterwards'	ba_{1CS02} $napile_{V\_aux}$ ba_{2CS02} mo_{CO01} $itia_{V}$ $gape_{ADV}$ 'they afterwards hit him again' (Louwrens, 1991, p. 52)
ke	n.a.	'should'	ba_{CO02} $kgopele_{V}$ $gore_{CONJ}$ ba_{1CS02} ke_{V_aux} ba_{2CS02} $homole_{V}$ $ganyane_{ADV}$ 'ask them (that they should) to be quiet a little (Louwrens, 1991, p. 52)
kago	past	'who once did/was/were	ba_{1CS02} $kago_{V}$ ba_{2CS02} $bolela_{V}$ 'they once used to talk'
ešo	n.a.	'not yet'	ga_{MORPH_neg} $ke_{1CSPERS\_1sg}$ $ešo_{V\_aux}$ $ka_{3CSPERS\_1sg}$ $bolela_{V}$ $nabo_{PART\_con02}$ 'I have not yet spoken to them' (Louwrens, 1991, p. 52)
tšama	n.a.	'continually'	o_{1CS01} $tšama_{V\_aux}$ a_{2CS01} ba_{CO02} $noša_{V}$ $sehlare_{N07}$ seo_{CDEM07} 'she continually lets them drink that medicine' (Louwrens, 1991, p. 52)
bilego	past	'on top of that'	ba_{1CS02} $bilego_{V\_aux}$ ba_{2CS02} $bolela_{V}$ 'on top of that they were talking'
ile, kile	past	'once upon a time'	o_{1CS01} ile_{V_aux} a_{2CS01} $kgaoga_{V}$ $pelo_{N09}$ ge_{CONJ} a_{2CS01} $lemoga_{V}$ seo_{CDEM07} 'he once was greatly troubled when he noticed/realised that' (Lombard, 1985, p. 188) $naga_{N09}$ e_{1CS09} $kile_{V}$ ya_{3CS09} $tlala_{V}$ $diphoogolo_{N10}$ 'Once upon a time the country was full of game' (Lombard, 1985, p. 189)

Table 2.16: Examples of verbs that may be used as auxiliaries

V_aux	***translation***	***Example of use***
šetše	'stayed, remained'	o_{1CS01} $šetše_{V}$ gae_{NLOC} 'he stayed/remained at home'
	'already'	o_{1CS01} $šetše_{V\_aux}$ a_{2CS01} $boile_{V}$ 'he has already returned (Louwrens, 1991, p. 51)
dula	'sit (down), live, stay'	$o_{1CSPERS\_2sg}$ $dula_{V}$ kae_{QUE_loc}? 'where do you stay? (De Schryver, 2007, p. 42)
	'keep on (doing)'	ba_{1CS02} $dula_{V\_aux}$ ba_{2CS02} *leta* 'they keep on waiting' (De Schryver, 2007, p. 42)
ehlwa	'spend the day'	$o_{1CSPERS\_2sg}$ be_{V_aux} $o_{2CSPERS\_2sg}$ $ehlwa_{V}$ gae_{LOC} $o_{1CSPERS\_2sg}$ bapala *V* 'you spent the whole day at home playing' (De Schryver, 2007, p. 44)
	'continue'	ba_{1CS02} $ehlwa_{V\_aux}$ ba_{2CS02} $bolela_{V}$ 'they continue talking'

Table 2.17: Some copula of Northern Sotho

Copula	**Annotation**	**Translation(s)**
ba, *eba*	VCOP	become
bago	VCOP	which is/are becoming
be	VCOP	was/were
bego	VCOP	which was/were
bile	VCOP	was/were
bilego	VCOP	which was/were
le	VCOP	is/are/am
lego	VCOP	which is/are
se	VCOP_neg	is/are not
seng, *sego*	VCOP_neg	which is/are not
ena, *na*	VCOP	have/has
nago	VCOP	which have/has
ne	VCOP	was/have

2.8 Adverbs (ADV)

The word class ADV can be divided into basic or derived adverbs (cf. e.g. (Lombard, 1985, p. 166 et seq.)). Reduplicated forms appear, for example *bjalebjale* 'in a moment, quickly', a reduplicated form of *bjale* 'now'. A number of nouns can also function as adverbs, e.g. all locative nouns, locativised forms (the *–ng* derivations of nouns), and nouns indicating time, like *lehono* 'today'.

The word class 'adverb' is open, therefore new forms can be derived by e.g. prefixing the locative particle (cf. paragraph 2.10.6) *ga–* to proper names, like in *GaSekhukhune* 'at Sekhukhune'. The locative particle *ga–* also indicates possession, as such *GaSekhukhune* refers to the place belonging to a person called *Sekhukhune*. More generally, according to Poulos and Louwrens (1994, p. 335), it can refer to a territory or a place name. Another possible derivation entails prefixing the possessive stem *gabo–*, like in *gabomogolo* 'at the elder brother's/sister's' (derived from *mogolo* 'elder brother/sister').

2.9 The morphemes (MORPH)

The bound morphemes of Northern Sotho occur only within verbs, where each of them provides an aspectual addition to a verbal meaning. One could argue that the concords are also morphemes and thus belong to this category. However, as the concords are class-dependent, they constitute a sub-category of morphemes and are as such described and labeled separately.

The word classes described in this paragraph often contain one or only few element(s). However, as will be shown in paragraph 3.2, each has a specific morphosyntactic function and occurs in a specific environment; therefore the definition of separate word classes seems justifiable.

2.9.1 The imperfect or present tense morpheme *a* (MORPH_pres)

The name of this morpheme is actually misleading because such a morpheme indicates "that the information which follows indicative verbs is old or redundant" (Louwrens (1991, p. 23) referring to Kosch (1985)). Other authors, like Van Wyk et al. (1992, p. 22) or Poulos and Louwrens (1994, p. 72 and 202 et seq.) follow the traditional point of view in describing this morpheme as marking the present tense in the "long" form of the verb, because this

morpheme only occurs in the present tense form of indicative verbs, like in example (19), cf. paragraph 3.2.5.1.

(19) *o*$_{1CS01}$ ***a***$_{MORPH\_pres}$ *ipshina*$_{V}$
subj-3rd-cl1 **pres** enjoy oneself
's(h)e is enjoying herself/himself'

2.9.2 The perfect or past tense morpheme *a* (MORPH_past)

The past tense morpheme *a* only appears in one constellation, a negated perfect indicative, cf. paragraph 3.2.5.2 and example (20). Lombard (1985, p. 146) labels this morpheme the "perfect/stative *a*".

(20) *ga*$_{MORPH\_neg}$ *o*$_{2CS03}$ ***a***$_{MORPH\_past}$ *tšhaba*$_{V\_itr}$
neg subj-3rd-cl3 **past** flee
'It did not flee'

2.9.3 The future tense morphemes (MORPH_fut)

There are two interchangable allomorphs adding a future aspect to a verb, viz. *tla* and *tlo*. To translate them into English, auxiliary structures containing 'shall' or 'will' are used, cf. (21), taken from Van Wyk et al. (1992, p. 55).

(21) *ba* ***tlo*** *gana*
subj-3rd-cl2 **fut** refuse
'they will refuse'

Both morphemes also occur in the relative form as *tlogo* and *tlago* 'who/which shall/will', as demonstrated in (22) by Lombard (1985, p. 143)

(22) *banna* *ba* *ba* ***tlogo*** *boa*
men cdem-3rd-cl2 subj-3rd-cl2 **fut-rel** return
'(the) men who will return'

2.9.4 The potential morpheme *ka* (MORPH_pot)

A number of verbal phrases express the possibility of an event. The potential morpheme *ka*, appearing between subject concord and verb stem marks this potential form, as in (23). Lombard (1985, p. 190) describes this morpheme as "potential deficient verb form".

(23) *di*$_{CS10}$ ***ka***$_{MORPH\_pot}$ *fula*$_{V\_itr}$
subj-3rd-cl10 **pot** graze
'they may (possibly) graze'

Secondly, the potential morpheme appears in negated future tense forms, as will be described e.g. in paragraph 3.2.5.3, cf. example 24, (Lombard, 1985, p. 147).

24	*mmutla*	*o*	***ka***	*se*	*tšhabe.*
	hare	`subj-3rd-cl3`	**`pot`**	`neg`	flee.
	'(a) hare will not flee.'				

2.9.5 The negation morphemes (MORPH_neg)

The use of the negation morphemes *ga*, *sa* and *se* will be described in detail in chapter 3.

2.9.6 The infinite morpheme *go* (MORPH_cp15)

This morpheme has been described in paragraph 2.2.2.5.

2.9.7 The deficient morphemes (MORPH_def) and the progressive morpheme *sa* (MORPH_prog)

Lombard (1985, p. 189) describes deficient verb forms as shortened auxiliaries that historically became part of the verb structure. As such, he and also Van Wyk et al. (1992, p. 55 et seq.) include the future morphemes (cf. paragraph 2.9.3) in this class. Our system's lexicon, where the elements are labeled alongside Taljard et al. (2008), only lists *fo* 'just', *no* 'simply' and *yo*, a fused variant of *ya go* 'go to' that appears to be similar to the English 'going-to future'.

Ziervogel (1988, p. 34) describes the progressive morpheme *sa* briefly as expressing the word sense 'still and to appear in certain verbal constellations, example (15) of page 49 is repeated as (25) here for the sake of convenience. However, these morphemes will not be further dealt with in this study; here, they are only mentioned for the sake of completeness.

(25)	$le_{\text{CSPERS}_2\text{pl}}$	$sa_{\text{MORPH\_prog}}$	e_{CO09}	$nyaka_{\text{V}}$
	`subj-2nd-pl`	still	`obj-3rd-cl9`	want
	you(pl)	still	it	want
	'(all of) you still want it'			

2.10 Particles (PART)

Unlike the bound morphemes listed in paragraph 2.9, particles are free morphemes that can be heads of phrases (containing nominal complements), in other words, the appearances

of some of them are comparable to that of prepositions of other languages. There are a number of particles of this kind in Northern Sotho, e.g. the agentive, the temporal and the instrumental. The connective particles can show both a preposition-like and a conjunction-like character, while the question particle added to any sentence marks it as a question. In the following paragraphs names and functions of these and other particles will be explained in more detail.

Note that like the morpheme classes, some of the particle classes contain one element only; however, again such elements must be considered unique in their use.

2.10.1 The agentive particle *ke* (PART_agen)

This particle introduces 'by'-phrases, like in *longwa ke mpša* 'bitten by the/a dog'. It is used with passive verbs only and usually requires a nominal as its complement (Lombard, 1985, p. 173).

2.10.2 The connective particles (PART_con)

There are a few connective particles in Northern Sotho, of which *le* appears fairly frequently. Two different uses of this particle are described, *le* 'and/with' when used comparably to a conjunction, like in $basadi_{\text{N02}}$ $le_{\text{PART\_con}}$ $bana_{\text{N02}}$ 'women and children'. If used after a verb, *le* appears to be similar to the English preposition 'with' in an associative sense, demonstrated in (26) by Van Wyk et al. (1992, p. 169).

(26)

o_{1CS01}	$sepela_{\text{V\_itr}}$	$le_{\text{PART\_con}}$	$mosadi_{\text{N01}}$
`subj-3rd-cl1`	walk	`con`	woman

'(s)he walks with (a) woman'

Louwrens (1991, p. 96) adds another example of the use of the connective[11] particle *le*: it is shown in (27). Some verbs of telling, like e.g. $boletše_{\text{V}}$ 'spoke' (the perfect tense form of $bolela_{\text{V}}$ 'speak'), require their complement to be a (connective) particle phrase headed by $le_{\text{PART\_con}}$.

(27)

o_{1CS01}	$boletše_{\text{V\_itr}}$	$le_{\text{PART\_con}}$	$morutiši_{\text{N01}}$
`subj-3rd-cl1`	spoke	`con`	teacher

'(s)he spoke to (a) teacher'

[11] Louwrens (1991) calls this particle "associative".

Table 2.18 contains the other connective particles, which are fused forms of *na*+pronoun 'with him/her/it/them'. These particles appear with an associative function. Note that except for *nago* 'with you' no form exists referring to first or second person(s) to our knowledge. Also, no forms of class 15 or LOC appear to exist, hence there is no full paradigm given in Table 2.18.

Table 2.18: The connective particle *na* fused with pronouns

Connective particle	**Annotation**	**Translation(s)**
nago	PART_con2sg	'with you'
nae, *naye*	PART_con01	'with him/her'
nabo	PART_con02	'with them'
nawo	PART_con03	'with him/her/it/them'
nayo	PART_con04	'with him/her/it/them'
nalo	PART_con05	'with him/her/it/them'
naso	PART_con06	'with him/her/it/them'
natšo	PART_con08	'with him/her/it/them'
nayo	PART_con09	'with him/her/it/them'
natšo	PART_con10	'with him/her/it/them'
nabjo	PART_con14	'with him/her/it/them'

2.10.3 The copulative particle *ke* (PART_cop)

In paragraph 2.7.4, $ke_{\text{VCOP\_1sg}}$ was introduced as an identifying copulative semantically containing a subject of the first person singular (to be translated as 'I am'). Note that $ké_{\text{PART\_cop}}$ (high tone) 'it is' represents the class-independent copulative, while *kè* (low tone) 'I am' represents the subject concord of the first person singular used as a copulative. However, their orthographic forms are identical, as 28(a) and (b) demonstrate. The use of this particle will be demonstrated in more detail in paragraph 3.3.1.

28(a) *ke* *morutiši*
`subj-3rd-cop` teacher
'it is (a) teacher'

(b) *ke* *morutiši*
`subj-1st-sg` teacher
'I am (a) teacher'

2.10.4 The hortative particles (PART_hort)

In most cases found in the corpus, it is *a* 'let' that is used as hortative particle, usually followed by a subject concord of a person, like in $a_{\text{PART\_hort}}$ $re_{\text{CSPERS\_1pl}}$ nwe_{V} $teye_{\text{N09}}$ 'let us drink tea' or $a_{\text{PART\_hort}}$ $ke_{\text{CSPERS\_1sg}}$ ... 'let me ...'. Other hortative particles are *ake*, *anke* and *ga*, a variant of *a*. These forms are usually translated into the English 'please' (Lombard, 1985, p. 155 et seq.).

2.10.5 The instrumental particle *ka* (PART_ins)

From a syntactical perspective, the instrumental particle *ka* can be treated similarly to the English instrumental preposition 'with', as it requires one complement (the instrument), which is usually nominal. Like all particles (or English prepositions) of this kind, it appears as an adjunct to the verb, though there are exceptions, as in (29).

(29) $ke_{\text{PART\_cop}}$ $ka_{\text{PART\_ins}}$ $lebaka_{\text{N05}}$ la_{CPOSS05} $eng_{\text{QUE\_N09}}$
`subj-3rd-cop` with reason of what
'what is (a) reason for'

2.10.6 The locative particles (PART_loc)

Several locative particles inform about directions of actions/states described by the preceding verb. Lombard (1985, p. 170) lists the following examples: *go* '(in the direction) to', *ka* 'in(side)', *mo* 'on', or *kua* 'there(in)'. Poulos and Louwrens (1994, p. 335) add *ga* 'to/at (the place of)' demonstrating its use in (30).

(30) *ke* *ya* *ga* *kgoši* *Matlala*
`subj-1st-sg` go to king Matlala
'I go to chief Matlala's place'

2.10.7 The question particles (PART_que)

Question particles can be added to any sentence, *na* at its beginning or at its end, *a* only at its beginning, to mark a question. In some cases, the variant *naa* is used.

2.10.8 The temporal particle *ka* (PART_temp)

According to Louwrens (1991, p. 27), this particle is not to be confused with the instrumental, as its homograph is to be understood to 'specify a particular point in time which the process expressed by the verb is associated', like shown in (31).

31(a) *ke*$_{\text{CSPERS\_1sg}}$ *tla*$_{\text{MORPH\_fut}}$ *boa*$_{\text{V\_itr}}$ *ka*$_{\text{PART\_temp}}$ *moswana*$_{\text{N03}}$
`subj-1st-sg` `fut` return by tomorrow
'I shall return by tomorrow'

(b) *re*$_{\text{CSPERS\_1pl}}$ *tla*$_{\text{MORPH\_fut}}$ *thoma*$_{\text{V\_itr}}$ *go*$_{\text{MORPH\_cp15}}$ *šoma*$_{\text{V\_itr}}$ *ka*$_{\text{PART\_temp}}$
`subj-1st-pl` `fut` start to work by
Mošupulogo$_{\text{N03}}$
Monday
'we will start to work by Monday'

2.10.9 The question words (QUE$_{\text{categ}}$)

A number of class-independent interrogative or question words exist in Northern Sotho, like *bjang* 'how' or *gakae* 'how often', like in (32).

(32) *wena*$_{\text{PROEMPPERS\_2sg}}$ *o*$_{\text{1CS01}}$ *phela*$_{\text{V\_itr}}$ *bjang*$_{\text{QUE}}$ *?*
`emp-2nd-sg` `subj-3rd-cl1` live how ?
'how are you doing?'

Other question words are used in cases where the requested answer should contain a noun of a certain class, like *mang*$_{\text{QUE\_N01a}}$ 'who', that appears when asking a person's name (names are contained in class N01a), cf. (33) and paragraph 2.2.2.1. However, only a few forms seem to occur in the language (classes 01a, 02b, 9 and 14).

33(a) *o*$_{\text{VCOP\_2sg}}$ *mang*$_{\text{QUE\_N01a}}$ *?*
you are who ?
'who are you?'

(b) *ke*$_{\text{VCOP\_1sg}}$ *Mahlatse*$_{\text{N01a}}$.
I am Mahlatse.
'I am Mahlatse.'

Although the elements of the last category described here do not necessarily appear requiring nominals as answers, they are still used in a class-specific context, like *kae* 'how many' and *–fe* 'which'. The list of common question words is contained in Table 2.19.

Table 2.19: Question words of Northern Sotho

Question word	**Annotation**	**Translation**
bjang	QUE	'how'
gakae	QUE	'how often'
goreng	QUE	'why'
hleng	QUE	'(if you) please'
kae	QUE	'where'
neng	QUE	'when'
mang	QUE_N01a	'who' (sg)
bomang	QUE_N02b	'who' (plural or respect sg. form)
eng	QUE_N09	'what', 'why'
bokae	QUE_N14	'how many' (noun of class 14)
ofe	QUE_01	'which' (of class 1)
bafe	QUE_02	'which' (of class 2)
ofe	QUE_03	'which' (of class 3)
efe	QUE_04	'which' (of class 4)
lefe	QUE_05	'which' (of class 5)
afe	QUE_06	'which' (of class 6)
sefe	QUE_07	'which' (of class 7)
difе	QUE_08	'which' (of class 8)
efe	QUE_09	'which' (of class 9)
difе	QUE_10	'which' (of class 10)
bofe	QUE_14	'which' (of class 14)
gofe	QUE_15	'which' (of class 15)
gofe	QUE_loc	'which' (of class LOC)

2.10.10 Miscellanous

All other parts of speech defined in the tagset are summarised in Table 2.20. Some of them will be explained and/or their use demonstrated with the examples mentioned throughout the following paragraphs on morphosyntactic rules. Others are found in our corpora seldomly and need more examination before they can be described in more detail.

Table 2.20: Excerpt of the tagset: Miscellanous part of speech

Short description	**Annotation**	**Example of use**
Conjunction	CONJ	*gore* 'that'
Interjection	INT	*hle* 'please'
Ideophone	IDEO	*tserr* 'suffocating hot'
Abbreviation	ABBR	*SABC* (South African Broadcasting Corporation)
Numeral	NUM	*šupa* 'be seven in number'
		1,2,3, ...
Ordinal	ORD	1., 2., a), b)
Clause separating punctuation	$.	, : ; . ! ?
Quoting punctuation	$"	' " ( )
Other punctuation	$'-	/ & %

2.11 Summary

This chapter has given a brief overview of the parts of speech used in the proposed system's lexicon. All definitions of word classes were taken from the literature, especially from Lombard (1985); Van Wyk et al. (1992); Poulos and Louwrens (1994) and Taljard et al. (2008). The examples (and their translations) stated in this chapter were found either in this literature, in text collections (e.g. Thobakgale (2005) and Matsepe (1974)), and/or in dictionaries, like in Endemann (1911); Ziervogel and Mokgokong (1975) and De Schryver (2007).

As Tables 2.20 and 2.21 (taken from Taljard et al. (2008)) illustrate, besides punctuation, basically 18 word classes are distinguished: nouns, pronouns (emphatic, possessive, quantitative), concords (subject, object, possessive, demonstrative, demonstrative copulative), adjectives, verbs (stems and auxiliaries), copulas, adverbs, morphemes (present tense, past, future, negation, deficient, potential, class 15 prefix), particles (agentive, connective, copulative, hortative, instrumental, locative, question, temporal), question words, conjunctions,

interjections, ideophones, abbreviations, enumeratives, numerals and ordinals.

The noun classes are additionally labeled onto all parts of speech that are class specific, i.e. that have to show agreement with other parts of speech. Some are assigned on the first level of annotation; this leads to a total amount of altogether 141 possible POS labels on this level. The tagset size further amounts to 263 possible annotations if the second level of annotation is taken into account.

The following chapter will contain definitions of grammar rules, i.e. deliver a morphosyntactic description of how a number of these word classes appear in Northern Sotho text.

Table 2.21: The tagset of Northern Sotho 1 / 2

Description	**tag 1st level**	**tag 2nd level**
concords		
subject class 1 – 10,14,15	CS01 – CS10, CS14, CS15	–
personal subject	CSPERS	1sg,2sg,1pl,2pl
locative subject	CSLOC	–
indefinite subject	CSINDEF	–
neutral subject	CSNEUT	–
object class 1 – 10,14,15	CO01 – CO10, CO14, CO15	–
personal object	COPERS	2sg,1pl,2pl
locative object	COLOC	–
possessive class 1 – 10, 14, 15	CPOSS01 – 10, CPOSS14, CPOSS15	–
possessive locative	CPOSSLOC	–
demonstrative class 1 – 10, 14	CDEM01 – CDEM10, CDEM14	–
demonstrative copulative	CDEMCOP	01 – 10, 14, 15, loc
pronouns		
emphatic class 1 – 10, 14, 15	PROEMP01 - 10, 14, 15	–, loc
emphatic personal	PROEMPPERS	1sg,2sg,1pl,2pl
emphatic locative	PROEMPLOC	–
possessive class 1 – 10, 14, 15	PROPOSS01 – 10, 14, 15	–
possessive personal	PROPOSSPERS	1sg,2sg,1pl,2pl
possessive locative	PROPOSSLOC	–
quantitative class 1 – 10, 14, 15	PROQUANT01 – 10, 14, 15	–
quantitative locative	PROQUANTLOC	–
nouns		
class 1 – 10, 14	N01 – N10, N14	–, dim, aug, loc
locative	NLOC	–, dim
names of persons singular	N01a	–
names of persons plural / respect form	N02b	–
names of places	NPP	loc

Table 2.22: The tagset of Northern Sotho 2 / 2

Description	**tag 1st level**	**tag 2nd level**
adjectives		
class 1 – 10, 14, 15	ADJ01 – 10, ADJ14, ADJ15	–, dim
locative	ADJLOC	–
verbals		
verb stem	V	–, aux
copula	VCOP	–, N01 – N10, N14
morphemes		
deficient	MORPH	def
negation	MORPH	neg
potential	MORPH	pot
future	MORPH	fut
present	MORPH	pres
past	MORPH	past
progressive	MORPH	prog
class 15 marker	MORPH	cp15
particles		
agentive	PART	agen
connective	PART	con
copulative	PART	cop
hortative	PART	hort
instrumental	PART	ins
locative	PART	loc
question	PART	que
temporal	PART	temp
question words		
nominal	QUE	N01 – N10, N14
others	QUE	–, 01 – 10, 14, 15, loc
others	see Table 2.20	