A morphosyntacic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English

Gertrud Faaß

A thesis submitted in accordance with the requirements
for the degree of Ph.D. in African Languages at
the University of Pretoria.
June 2010 (final version).

Supervisor: Prof. D.J. Prinsloo
Co-Supervisor: Prof. U. Heid

Summary

This PhD thesis provides a morpho-syntactic description of Northern Sotho from a computational perspective. While a number of publications describe morphological and syntactical aspects of this language, may it be in the form of prescriptive study books (inter alia Lombard (1985); Van Wyk et al. (1992); Poulos and Louwrens (1994)) or of descriptive articles in linguistic journals or conference proceedings (inter alia Anderson and Kotzé (2006); Kosch (2006); De Schryver and Taljard (2006)), so far no comprehensive description is available that would provide a basis for developing a rule-based parser to analyse Northern Sotho on sentence level. This study attempts to fill the gap by describing a substantial grammar fragment. Therefore, Northern Sotho morpho-syntactic phenomena are explored which results in the following descriptions:

- language units of Northern Sotho are identified, i.e. the tokens and words that form the language. These are sorted into word class categories (parts of speech), using the descriptions of Taljard et al. (2008) as a basis;
- the formal relationships between these units, wherever possible on the level of parts of speech, are described in the form of productive morpho-syntactic phrase grammar rules. These rules are defined within the framework of generative grammar.

Additionally, an attempt is made to find generalisations on the contextual distribution of the many items contained in verbs which are polysemous in terms of their parts of speech. The grammar rules described in the preceding chapter are now explored in order to find patterns in the co-occurrence of parts of speech leading towards a future, more general linguistic modelling of Northern Sotho verbs. It is also shown how a parser could work his way step-by-step doing an analysis of a complete sentence making use of a lexicon and the rules developed here.

We have also implemented some relevant phrase grammar rules as a constraint-based grammar fragment, in line with the theory of *Lexical-Functional Grammar* (Kaplan and Bresnan, 1982). Here, we utilized the Xerox Linguistic Environment (XLE) with the friendly permission of the Xerox Palo Alto Research Centre (PARC).

Lastly, the study contains some basic definitions for a proposed machine translation (MT) into English attempting to support the development of MT-rules. An introduction to MT and a first contrastive description of phenomena of both languages is provided.

Key terms

Northern Sotho, English, language units, word classes, electronic grammars, morpho-syntactic analysis, Northern Sotho Verbal Phrase features, word class distribution, automated translation, machine translation, grammar implementation.

Declaration

I declare that **A morphosyntacic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English** is to the best of my knowledge and belief, my original work. All the sources that I have used or quoted have been indicated and acknowledged by means of complete references. The material has not been submitted, either in whole or part, for a degree at this or any other university.

Dipl. Ling. Gertrud Faaß

Acknowledgements

I should like to express my gratitude towards my promoters, Prof. Daan J. Prinsloo and Prof. Ulrich Heid. I freely admit that without both of their support, this work would have never been finished.

Many thanks to the other members of the department of African Languages at the University of Pretoria, especially Prof. Elsabé Taljard.

I am indebted to the DAAD, the German Academic Student Exchange Service, as their financial contribution made the project possible.

Thanks to the National Research Foundation of South Africa, who supported the funding of my working environment.

My thanks also go to friends and family, who all supported me throughout the time it took to complete this book.

There were a number of other people, for whom the space here is not enough to list them all. They helped me with translations and contributed to my general understanding of the languages and cultures of South Africa. *Ke a leboga*, *baie dankie* and thank you to all of you, as you offered me a grip of what is called *Ubuntu*.

List of Abbreviations

1st	First person (inflectional element)
2nd	Second person (inflectional element)
3rd	Third person (inflectional element)
ADVP	Adverb Phrase
AP	adjective phrase
c-structure	constituent structure
CALL	Computer Aided Language Learning
CFG	Context–Free (Phrase Structure) Grammar
CONJ	conjunction
DM	Distributed Morphology
f-structure	functional structure
HLT	Human Language Technology
LFG	Lexical Functional Grammar
MT	Machine Translation
NP	Nominal Phrase
NPs	Nominal Phrases
OBJ-TH	thematic object
PARC	Palo Alto Research Centre
past	Subscript: past tense (inflectional element
PF	Phonological Form
pl	Subscript: plural (inflectional element)
POS	Part of Speech
PP	Particle Phrase
PSC	University of Pretoria Sepedi Corpus
refl	reflexive
sg	Subscript: singular (inflectional element)

SMT	Statistical Machine Translation
TNS-ASP	TENSE-ASPECT
TTS	Text to Speech
VBP	Basic Verbal Phrase
Vend	The suffix(es) a verb stem ends in
VIE	Verbal Inflectional Element
VIEimp	imperative verbal inflectional phrase
VP	Verbal Phrase
VPimp	imperative VP
VPpred	predicative VP
VPs	Verbal Phrases
XLE	Xerox Linguistic Environment

Contents

List of Tables

List of Figures

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?'.

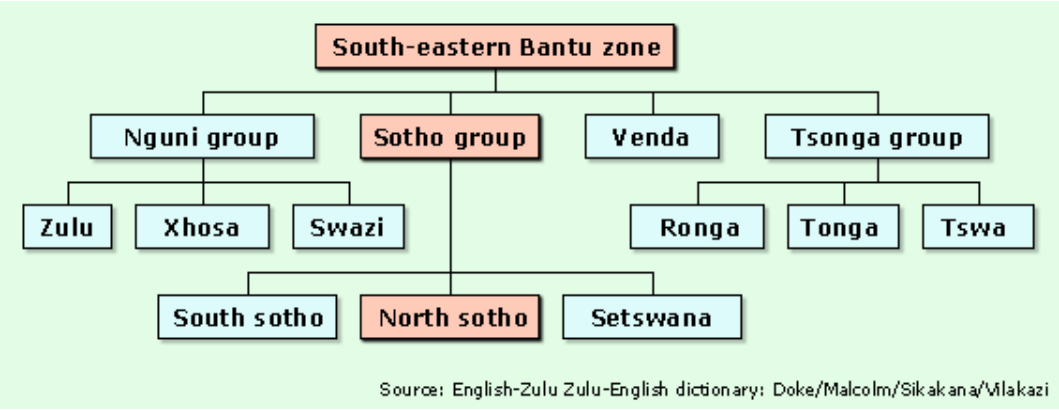


Figure 1.2: Northern Sotho as a part of the South-Eastern Bantu languages

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).

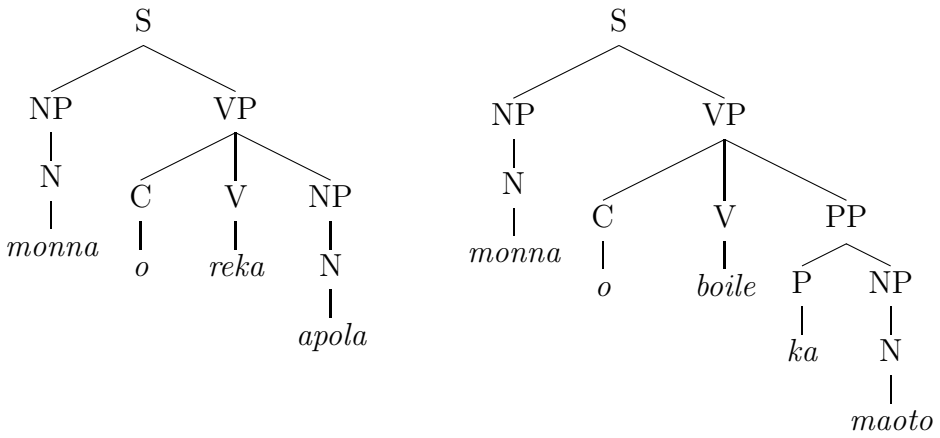


Figure 1.3: Example parse trees in CFG

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'				

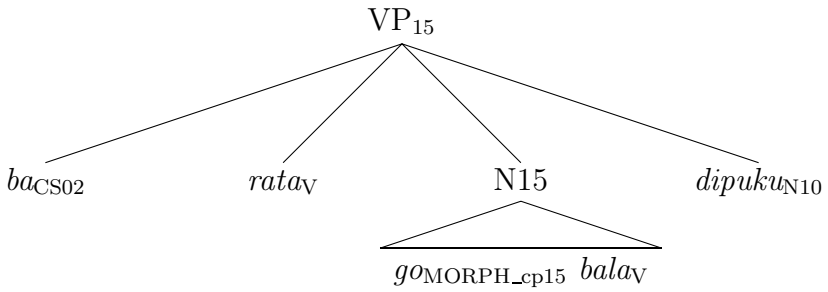


Figure 2.1: First analysis: *ba rata go bala dipuku* 'they like to read books'

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).

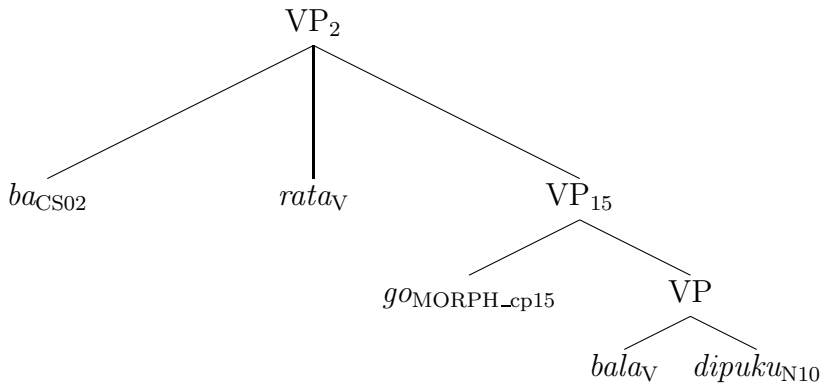


Figure 2.2: Second analysis: *ba rata go bala dipuku* 'they like to read books'

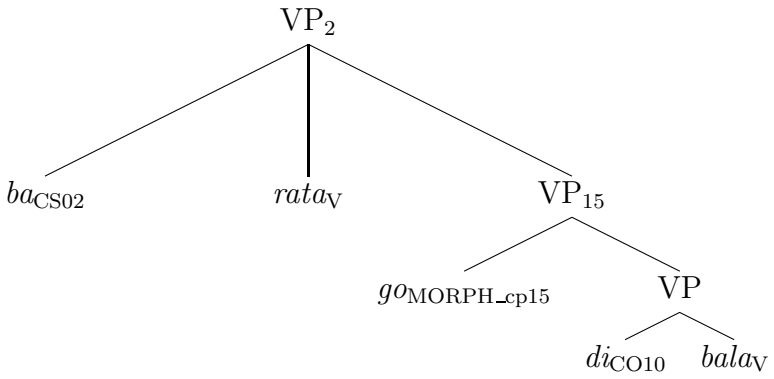


Figure 2.3: *ba rata go di bala* 'they like to read them'

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.

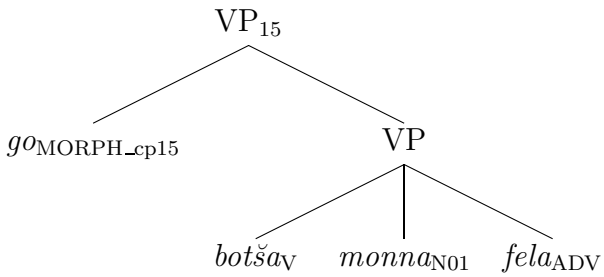


Figure 2.4: *go botša monna fela* 'to tell (the) man only'

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'

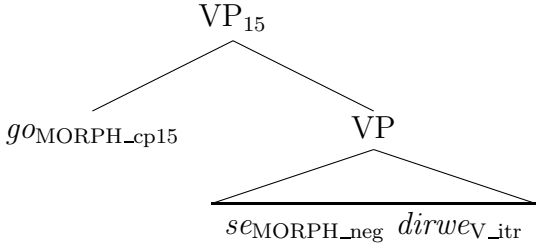


Figure 2.5: *go se dirwe* 'not to be done'

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'

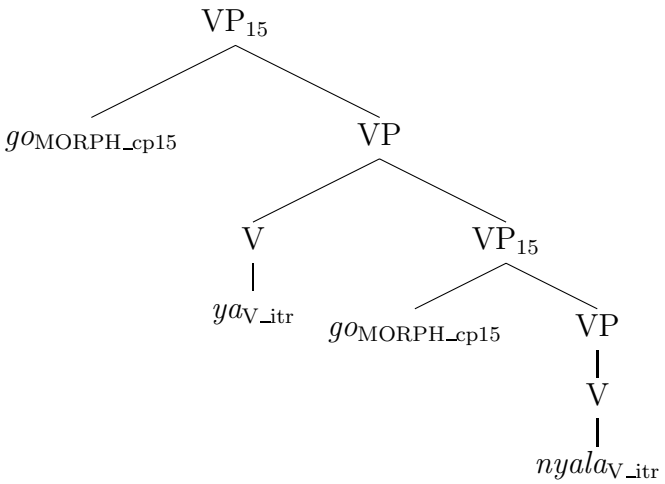


Figure 2.6: *go ya go nyala* 'going to marry'

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	

Chapter 3

A fragment of the grammar of Northern Sotho

3.1 Introduction

This chapter will describe a significant fragment of the Northern Sotho grammar, i.e. the definition of Northern Sotho words and phrases in terms of the order in which the POS can appear in them. The rules on how the phrases then combine to form permitted phrases and sentences are also explained.

Distributed Morphology (henceforth DM), as described in Embick and Noyer (2007), offers – at first glance – a similar perspective. Embick and Noyer (2007, p. 290) describe that in DM, a word "is not a privileged derivational object as far as the architecture of the grammar is concerned, since all complex objects, whether words or phrases, are treated as the output of the same generative system (the syntax)"[1]. In describing a token-based grammar system for a disjunctively written language, we seem to follow this principle of DM and like in DM, the grammar fragment described in this study utilizes morphemes, not linguistic words in its rules.

However, this study is not in line with DM, as a token of Northern Sotho still might present a complex word formed by morphological processes not described here. Therefore, we assume a system lexicon to be present, while in DM, there is no lexicon. Moreover, we do not describe a Phonological Form (PF) which is indeed described in DM.

[1]On Rolf Noyer's webpages on DM at the University of Pennsylvania (`http://www.ling.upenn.edu/~rnoyer/dm/`), he describes DM inter alia as "Syntactic Hierarchical Structure All the Way Down".

This section begins with an overview of the constellations described by Lombard (1985). It will attempt to map these onto structures we find in Poulos and Louwrens (1994). This consequently means that we might add to the rules of Lombard whenever it is necessary by using rules defined by Poulos and Louwrens (1994), if such structures appear in the analysed sentences. There are significant differences in their definitions of verbal moods which will be explored later in this chapter. However, as we aim to consistently retain the computational perspective in this study, we cannot side with either of the authors and opt for viewing the described constellations only in terms of their processability.

In order to gain a wider overview of possible constellations, Van Wyk et al. (1992) and Louwrens (1991) will be considered as well. This is not a trivial task because different word class systems and categorisations are used by these authors, of which some are of a contradictionary kind, hence the different approaches are all described wherever possible. In any other cases, there will be a reference to the appropriate literature.

Note that some of the phenomena described by several of these authors will not be considered, like the insertion of some rather rare "aspect prefixes" Poulos and Louwrens (1994, p. 289 et seq.) or "prefixal morphemes" (Lombard, 1985, p. 148) into verbal elements. This and other issues might be accounted for at a later stage of the project. No distinction will be made between a morphological and a syntactic rule, hence the term 'morphosyntactic' rule is used. Generally, all elements constituting the verb, for example, will equally form one verbal phrase, regardless of whether they are bound morphemes (cf. paragraph 1.1) or e.g. a nominal phrase appearing as the verb's object. We opted for this unusual method mainly because of the object concord. On the one hand, this concord functions as a pronominal object of the verb whenever it appears whereas on the other hand it remains a bound morpheme and constitutes a morphological part of the verb. In Northern Sotho there are therefore cases of bound morphemes with a syntactic function, including that of the object.

Like the object nominal that can be represented by an object concord, the subject nominal in Northern Sotho clauses can be omitted. In this case, the subject concord, another bound morpheme, will acquire the function of the grammatical subject[2].

[2]The object concord usually only appears whenever the object noun is deleted, while the subject concord is present in all predicative verb constellations where it usually is responsible for the agreement with the subject.

As shown in the previous chapter, some definitions of Taljard et al. (2008), on which our tagset is based, may be different from the descriptions found in the literature referenced above. To avoid confusion, only the tagset of Taljard et al. (2008) will be used (for an overview, consider Tables 2.20 and 2.21 (a) and (b) on pages 65, 67, and 68) in the following rule definitions.

The section begins with verbal phrases (VPs) of Northern Sotho, followed by noun phrases (NP) and continues describing adjunctive phrases like the particle phrase (abbreviated PP for their similarity with prepositional phrases of English or German), adjective phrase (AP), and adverb phrase (ADVP). It will end with a brief discussion of Northern Sotho clauses and sentences.

3.2 The Verbal Phrase (VP)

3.2.1 Introduction

3.2.1.1 Basic Verbal Phrase (VBP) versus Verbal Inflectional Element (VIE)

For sake of convenience, we repeat the example (1) on page 9 as (34). It contains the disjunctively written verb *ke tlo apea* within a clause. It demonstrates that a verb in Northern Sotho usually contains (a number of) bound morphemes written separately, but in a certain order (these morphemes precede the verb stem). The word class V (representing only the verb stem (cf. paragraph 2.7)) should therefore not be confused with complete verbs[3]. Figure 3.1 shows a morphosyntactic analysis of (34).

(34)

Nna$_{\text{PROEMPPERS\_1sg}}$	*ke*$_{\text{CSPERS\_1sg}}$	*tlo*$_{\text{MORPH\_fut}}$	*apea*$_{\text{V\_tr}}$	*dijo*$_{\text{N10}}$
`emp-1st-sg`	`subj-1st-sg`	`fut`	cook	food

'I (personally) will cook (the) food'

When examining Northern Sotho verbal phrases, it can be observed that the constellations of morphemes preceding the verb stem determine the mood, the tense, and the actuality[4]

[3]For a more detailed argument on this definition, cf. Poulos and Louwrens (1994, p. 165 et seq.)

[4]Lombard (1985, p. 139 et seq) introduces the distinction between the sub-categories mood, tense and actuality of the verbs of Northern Sotho.

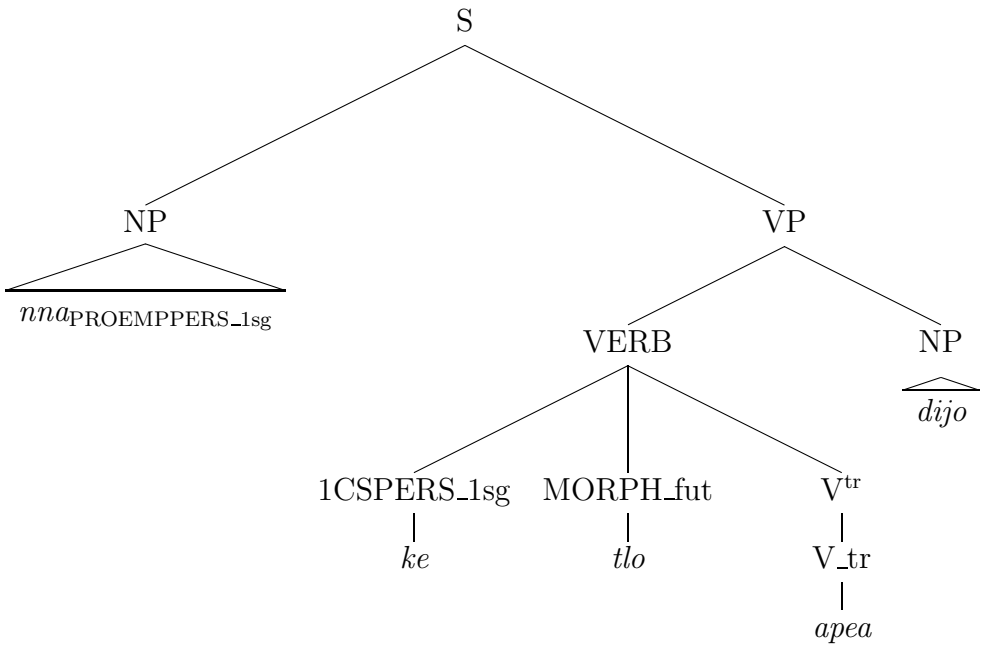


Figure 3.1: First analysis of *nna ke tlo apea dijo* 'I (personally) will cook (the) food'

(i.e. the positiveness or the negativeness) of the verb as a whole. The category of the subject concord(s) appearing in a predicative verb must be identical with the category of its subject (agreement in class and person). The verb stem's ending is also determined by the tense and the actuality.

The verb stem's lexical semantics on the other hand determines the kind and number of its necessary arguments. In other words, the type and number of syntactic functions which should appear in the clause or sentence are determined by the verb stem's valency only.

The position of these functional arguments is dependent on i.a. emphasis; topicalisation of an object, for example, may occur. However, usually the subject precedes the verb (like *nna* does in example 34) even if represented by a pronoun or by another pronominal (e.g. a demonstrative concord), while the functional object follows the verb stem and precedes possible adverbial extensions. The object can also be represented by an object concord preceding the verb stem (cf. paragraph 3.2.1.8). If there are two objects (required by double transitive verb stems), the indirect object will precede the direct object[5]. Unlike

[5]This ordering of the two objects is described by Ziervogel (1988, p. 82) as a rule solely for the applied verbal extension. However, all other sample sentences that were examined in the scope of this study show

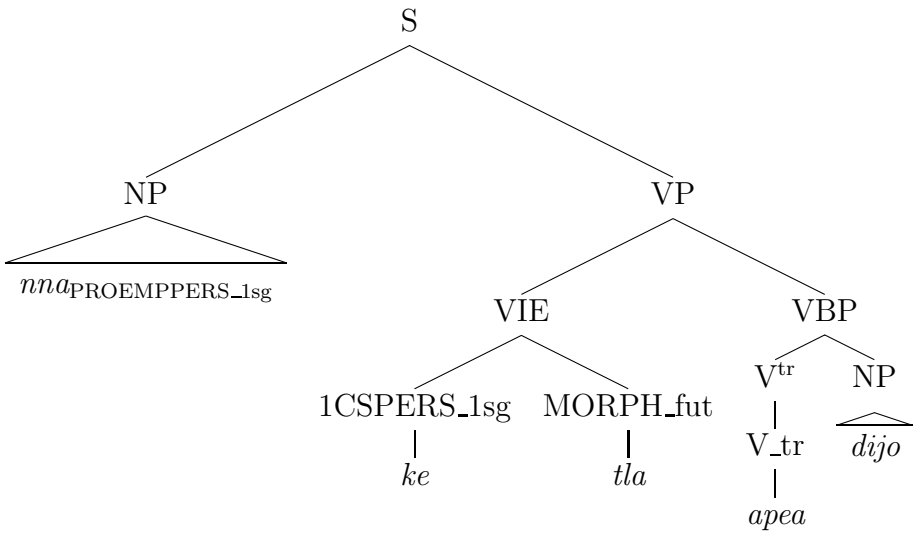


Figure 3.2: Second analysis of *nna ke tlo apea dijo* 'I (personally) will cook (the) food'

in other Bantu-languages (e.g. Tswana), in Northern Sotho, only one of the two objects of the double transitive verb – namely the indirect object – is usually represented by an object concord.[6]

To appropriately describe these phenomena, an independent structure is defined, the basic verbal phrase (VBP), which contains the verb stem and its object(s) (cf. paragraph 3.2.2). All other elements of the verb will be grouped separately into the verbal inflectional element (VIE). Figure 3.2 reflects this view.

3.2.1.2 Terminology used in this chapter

The terms 'verbal inflectional element' and 'basic verbal phrase' should not be confused with the term "verbal element" by Louwrens (1991, p. 17) who describes this term as a "main verb or an auxiliary word group" (without adjuncts). Louwrens (1991, ibid.) furthermore defines the term 'predicate' as including the verbal element and its adjuncts. Some linguists do not agree with this point of view, e.g. Bußmann (2002, p. 527) who explicitly describes

this order, therefore we take it as a general rule.

[6]When examining sample sentences, only one example of a direct object being represented by a object concord was found with a verb subcategorising direct and indirect object: *ke le ngwalela tate*, 'I write father it (a letter)'. No respective rules were found in the literature. Because *tate* 'father' could also be understood as an (adverbial) addressee, it was decided not to consider this case further.

the predicate as only containing the verb itself (proper main verb or copulative) or a verbal word group (e.g. auxiliary word group)[7]. In this study, the following terms are used (note that in the absence of adjuncts a predicate may constitute a VP):

→VBP	:	The verb stem and its objects;
→VIE	:	Inflectional elements of the verb;
→Predicate	:	VIE and VBP;
→VP	:	Verbal Phrase: VBP, VIE (optional), and adverbial constituents.

3.2.1.3 Introduction to the modal system

Lombard (1985, p. 144) gives an overview of verbal constellations. We present the data of Table 7.5.4 (Lombard (1985, ibid.)) in our Tables 3.2 and 3.3. These tables are augmented with information of the respective definitions of Poulos and Louwrens (1994). For each of those subcategories, constellations and examples will be shown in the following paragraphs. Table 3.1 however introduces the terms used in the modal system according to Lombard (1985, p. 139 et seq.).

[7]For a more detailed discussion on this issue, cf. e.g. Glück (2000).

Table 3.1: Lombard's modal system

General	**Dependency**	**Ind./Mod.**	**Mood**	**Comments**
Predicative				refers to a subject
	Independent			not dependent on other information, distinguishes tenses
		Indicating	Indicative	in main clauses
		Modifying		not in main clauses
			Situative	modifies the verb
			Relative	modifies the noun
	Dependent			dependent on other information, does not distinguish tenses
			Consecutive	chronologically dependent
			Subjunctive	causatively dependent
			Habitual	habitually dependent
Non-predicative				does not refer to a subject
			Imperative	
			Infinitive	

Table 3.2: Lombard's definition of independent moods compared with the respective constellations described by Poulos and Louwrens

Lombard	**Tense**	**Actuality**	**Poulos and Louwrens**
Non-predicative moods			**(no subject concords)**
IMPERATIVE		positive	√
		negative	√
INFINITIVE		positive	√
		negative	√
Predicative moods			**Incorporation of subject concords**
Independent moods			
INDICATIVE	Imperfect	positive	√INDICATIVE PRINCIPAL
(indicating mood)		negative	√
	Perfect	positive	√
		negative	√
	Future	positive	√
		negative	√
SITUATIVE	Imperfect	positive	√PARTICIPIAL (dependent)
(modifying mood)		negative	√
	Perfect	positive	√
		negative	√
	Future	positive	√
		negative	√
RELATIVE	Imperfect	positive	
(modifying mood)			not a mood
	Perfect	positive	
		negative	
	Future	positive	
		negative	

Table 3.3: Lombard's definition of dependent moods compared with the respective constellations described by Poulos and Louwrens

Lombard	**Actuality**	**Poulos and Louwrens**
Dependent moods		
CONSECUTIVE	positive	√ CONSECUTIVE
	negative	√
SUBJUNCTIVE	positive	√ SUBJUNCTIVE
	negative	√
HABITUAL	positive	√ HABITUAL
	negative	√
described as deficient auxiliary verb form		POTENTIAL

3.2.1.4 The slot system

As described in paragraph 3.2.1.1, the verb stem is usually followed by up to two objects of which one may be replaced by an object concord directly preceding the stem. We define the basic verbal phrase, VBP as containing the verb stem and its subcategorised arguments.

This VBP may appear together with bound morphemes to its left forming the VIE. Some of these morphemes appear in complementary distribution, i.e. the presence of some prevents the presence of others. The future morpheme MORPH_fut, for example, never appears together with the present tense morpheme, MORPH_pres. On the other hand, some morphemes in certain constellations have to occur together, like, for example, $ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$, forming a negation cluster. Such distributionary issues will be examined and summarised in chapter 4.

In order to simplify the graphical representation and to give a better overview of the many different VIEs, a slot-system is designed, that is, positions of certain parts of speech or parts of speech clusters as parts of phrases are defined. The slot system is then utilised for building morphosyntactic rules aiming at unambiguous analyses which can later be translated into e.g. phrase grammar rules.

The VBP is defined as 'slot zero' representing the core element of the VP. It makes use of one to four fields or positions. These VBP positions are numbered from the leftmost pos-1 to the rightmost pos+2, as Table 3.4 demonstrates. In each of the positions pos-1 to pos-0, only one token of a specific part of speech (an object concord and a verb stem respectively) may appear while the positions pos+1 and pos+2 are defined to contain the object(s) of (double) transitive verbs which can be nouns, nominals, noun phrases or even clauses. The central position contained in the VBP, 'pos-0', contains the verb stem. Slot zero forms part of all further descriptions of the verbal moods as it remains unchanged. The VIE slots are then built to the left of the VBP numbered as zero-1 to zero-2 (from right to left). Slot zero-1 may only contain one tense marker, i.e. the present tense or one of the future morphemes, while slot zero-2 contains the constellations of subject concord and/or negation marker(s). Except for slot zero, pos-0 containing the verb stem, all other positions are permitted to be empty, as in *Boeletša*! 'Repeat!', an imperative.

Table 3.4: A schematic representation of the slot system

The slot system					
VIE		**VBP**			
zero-2 subject and/or negation marker	**zero-1** tense marker	**slot zero** **verb stem and its object(s)**			
		pos-1 object concord	**pos-0** **verb stem**	**pos+1** object 1	**pos+2** object 2

3.2.1.5 Labels used on nodes

In this study, up to four labels will be attached to the nodes of trees and elements of morphosyntactic rules. Some of these labels are retrieved from the lexicon, e.g. the data that forms part of the word class label (e.g. V or N01_loc), others are added by means of the rules that are defined in this chapter (e.g. a VIE_{01} containing a subject concord of class 1, CS01). The appearance of labels is demonstrated in (35): the superscript left to the node is reserved for information on the verbal mood of a node, e.g. ^{IND}VP, to describe an indicative VP. The subscript to the left of the node shows a syntactic function, e.g. $_{OBJ}$NP (an object NP). When being used at lexical items, it can however also mark the perfect tense form of a verb stem, i.e. $_{perf}$V. The superscript to the right of the node is used whenever it is necessary to know the transitivity of an element, e.g. V^{itr}(i.a. an intransitive verb), and the subscript to the right of the node will show the noun class of the node, e.g. NP_{01} (NP of noun class 01), if necessary.

(35)

grammatical mood		transitivity
	NODE	
syntactic function		noun class

The category of a lexical item, as already shown in a number of examples, is annotated as sub to the right, like in nwa_{V} '[to] drink'.

3.2.1.6 Labelling information on the transitivity of verbs

A parser should be aware of the transitivity of a verb to avoid ambiguous and/or incorrect analyses. The problem arises especially in the case of Northern Sotho where punctuation is often used sparsely[8]. As a subject concord might represent an omitted subject nominal, any nominal placed between a verb and a subject concord might therefore function either as the object of the preceding, or as the subject of the following clause (cf. Figures 3.3 to 3.7 below). Therefore, whenever the respective punctuation is not present, it is problematic to identify the correct position of the sentence border. Moreover, nouns may generally be used either as arguments or as adverbials; only for locative or locativised nouns it can be said that they probably appear more often with an adverbial function that as an object of a verb, cf. example (36) of Van Wyk et al. (1992, p. 41).

(36) *ke*$_{\text{1CSPERS\_1sg}}$ *ya*$_{\text{V\_itr}}$ *sekolong*$_{\text{ADV}}$
`subj-1st-sg` go to-school
'I go to (a) school'

Without a lexicon containing information on the kind and number of arguments a verb requires, all these cases will be analysed ambiguously if the noun in question and the subject concord are of the same noun class. Example (37) shows an ambiguous case: two clauses, of which one contains a demonstrative concord. These demonstratives may accompany nouns, however, they might also occur with a pronominal function. Without information on the transitivity of the appearing verbs, *rata* '[to] like' and *tlile* 'arrived', a number of wrong analyses will ensue.

(37) *ke*$_{\text{1CSPERS\_1sg}}$ *rata*$_{\text{V}}$ *mošemane*$_{\text{N01}}$ *yo*$_{\text{CDEM01}}$ *o*$_{\text{1CS01}}$ *tlile*$_{\text{V}}$
`subj-1st-sg` like boy `dem-3rd-cl1` `subj-3rd-cl1` arrived
maabane$_{\text{N06}}$
yesterday
(see the analyses below)

Figures 3.3 to 3.7 show the possible analyses of the first clause, of which analysis (a) (assuming an intransitive verb) and (c) (assuming a double transitive verb) are incorrect because the verb *rata* '[to] like' is a transitive verb requiring one object. Note that *maabane*$_{\text{yesterday}}$ is labelled ADV to mark its adjunctive status in (a) to (d), the incorrect analysis (e) is based on the wrong assumption that *tlile* 'arrived' is a transitive verb, hence *maabane* would be

[8]To the best of the author's knowledge, official rules on where to use punctuation in Northern Sotho sentences have not yet been formulated.

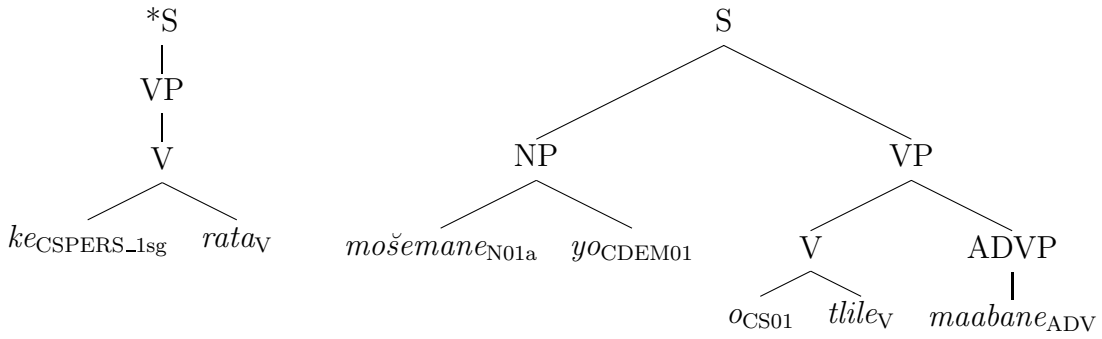


Figure 3.3: Analysis (a): *ke rata mošemane yo o tlile maabane* *'I like, this boy arrived yesterday'

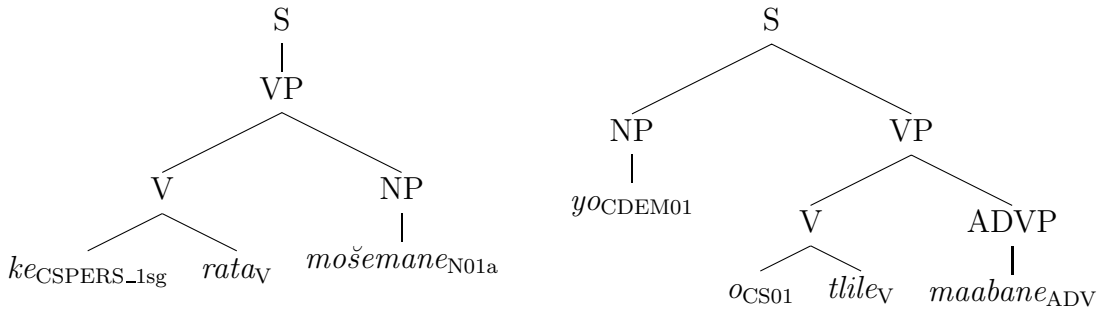


Figure 3.4: Analysis (b): *ke rata mošemane yo o tlile maabane* 'I like (the) boy, this one arrived yesterday'

analysed as a noun in this case.

In labelling the correct transitivity of a verb together with the introduction of rules that take these labels into account, the incorrect analyses (a), (c) and (e) can be avoided. The transitivity of the verb is therefore annotated on the second level of annotation, i.e. $rata_{V\_tr}$ '[to] like' to mark that this verb is transitive, $tlile_{V\_itr}$ 'arrived' to mark intransitivity, etc. Paragraph 3.2.3 will describe such rules in detail.

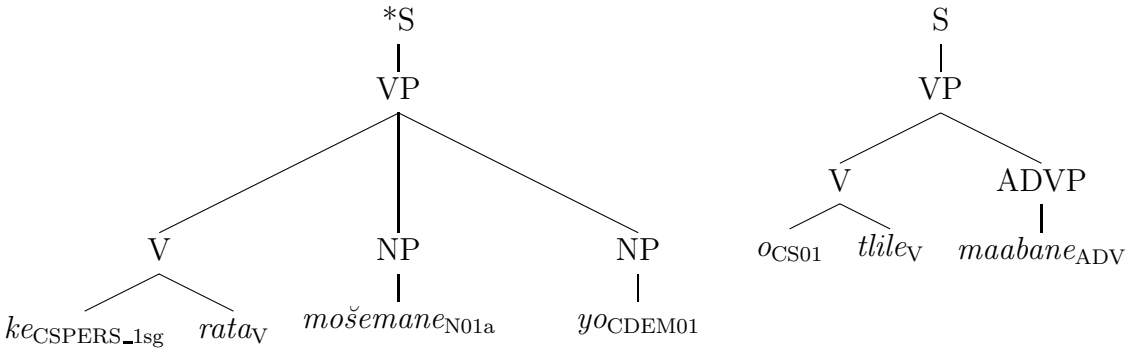


Figure 3.5: Analysis (c): *ke rata mošemane yo o tlile maabane* *'I like this boy, (some)one arrived yesterday'

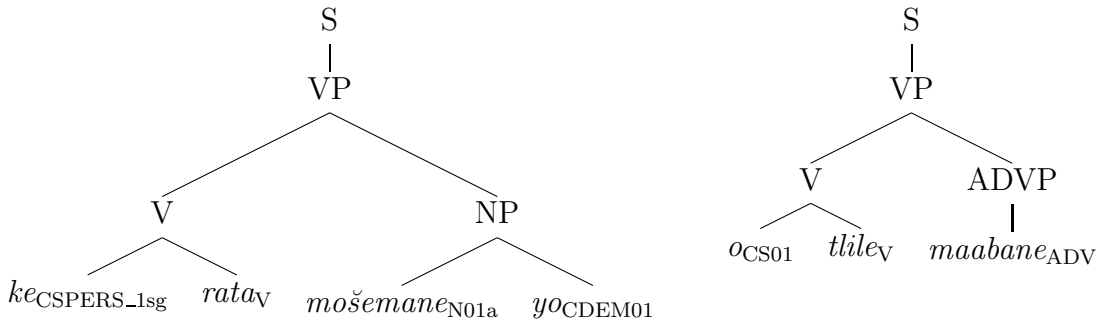


Figure 3.6: Analysis (d): *ke rata mošemane yo o tlile maabane* 'I like this boy, (some)one arrived yesterday'

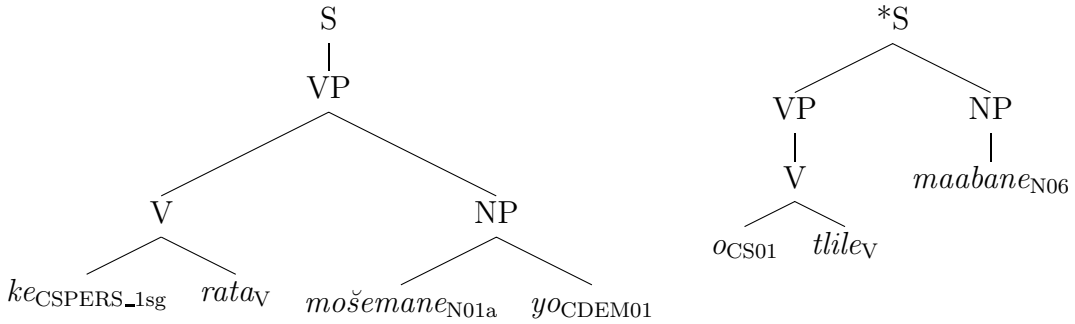


Figure 3.7: Analysis (e): *ke rata mošemane yo o tlile maabane* *'I like this boy, (some)one arrived yesterday'

3.2.1.7 Saturated verb forms

Another problem arises if an object concord (cf. paragraph 2.4.3) is fused with a transitive or double transitive verb stem, i.e. when both form one single token. This is the case in (38), where the object concord of the first person singular, *N-* '`obj-1st-sg`' is merged to the verb stems *thuša* '[to] help' and *fa* '[to] give', forming *nthuše* '`obj-1st-sg`-help' and *mphe* '`obj-1st-sg`-give'[9].

38 (a) *Nthuše*$_{\text{V\_tr}}$*!*
`obj-1st-sg`-help!
'Help me!'

38 (b) *Mphe*$_{\text{V\_dtr}}$ *puku!*
`obj-1st-sg`-give book!
'Give me the book!'

Due to the fact that *thuša* '[to] help' is transitive, a parser would require one overt argument, i.e. an object to be present in the sentences in which it appears. As *fa* '[to] give' is double transitive, two overt arguments would be expected to appear with this verb. As this study is however not concerned with the morphological analysis of fused forms, the fact that these verb forms already contain one functional object (represented by the proclitic object concord) has to be marked in their lexicon entry. A further second level annotation is therefore added to the part of speech label of these verb forms, namely a label indicating this information.

Our solution is based on the observation that the merged object concord figuratively 'saturates' the verb's requirement for an external object, we therefore suggest adding the labels **saturated transitive** (**sat-tr**) and **half-saturated** double transitive (**hsat-dtr**) verbs to our set[10] Other saturated forms are e.g. the reflexive forms, where the proclitic *i-* is fused with the stem, like in *ipona* '[to] see oneself' (derived from *bona* '[to] see'), these will be annotated accordingly.

Additional labels on words usually add to the number of morphosyntactic rules necessary

[9]Note that these verb stems when containing object concords or the reflexive morpheme end in *–e* or *–eng*.

[10]The term 'saturated' is chosen for its similarity with its use by Pollard and Sag (1994, p. 38), describing a "saturated phrase".

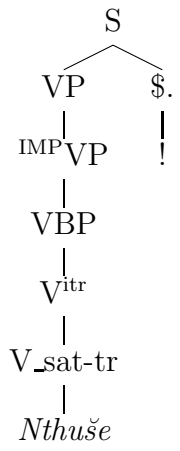


Figure 3.8: *Nthuše!* 'Help me!'

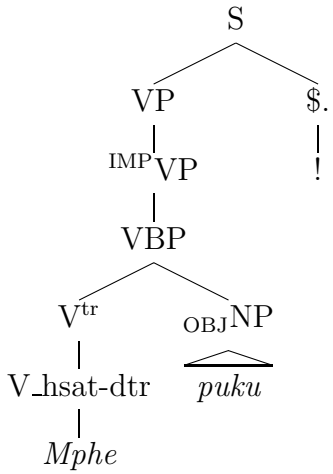


Figure 3.9: *Mphe puku!* 'Give me the book!'

to describe the verb as they extend the tagset. However, from the perspective of a grammarian, rules containing these labels do not have to be formulated explicitly. A V_sat-tr behaves identically to a V_itr in requiring no overt object phrase, while a V_hsat-dtr behaves identically to a V_tr in requiring one overt object to appear. Two simple general rules, syntactically equalising these verb forms (V_sat-tr = V_itr, V_hsat-dtr = V_tr) will allow us to ignore the saturated forms in our morphosyntactic rules, as demonstrated in Figures 3.8 and 3.9.

3.2.1.8 The object concord as part of the verb

We now come back to the phenomenon briefly introduced in paragraph 3.2.1.1, namely that a pronominal object concord representing an object can be inserted directly in front of the verb stem, as in (39)[11]. It has been indicated that the way that object(s) occur with the verb stem is independent of the inflectional morphemes preceding these constellations. In this paragraph, the linguistic background for our solution is provided, the phenomenon is described in more detail and proof is provided for a reasoning that the separation of the verb into VIE and VBP indeed results in correct morphosyntactic analyses.

39 (a) *monna* *o* *nwa* ***bjalwa***
man `subj-3rd-cl1` drink beer
'(a) man drinks (a) beer'

(b) *monna* *o* *a* ***bo*** *nwa*
man `subj-3rd-cl1` `pres` **`obj-3rd-cl14`** drink
'(a) man drinks it'

From a traditional syntactic perspective, object(s) of verbs form phrases (or clauses if their head is a verb) on their own and as such they fulfil a functional role towards the verb. Both verb and (usually) nominal phrase with an object function are then combined as daughter nodes of the verbal phrase (VP). Therefore, a description of the term 'verb' usually should not include its object but only the verb itself (For Northern Sotho that is the verb stem with its inflectional morphemes). However, in all respective literature examined (including (Van Wyk et al., 1992, p. 25) or Anderson and Kotzé (2006)), the object concord is described not as a separate grammatical unit but as part of the verb; fused forms of verb stem and object, like `mpona` '[to] see him/her' seem to prove this assumption.

Following traditional grammar rule systems, the insertion of the pronominal object concord ${}_{\text{OBJ}}bo_{\text{CO14}}$ '`obj-3rd-cl14`' into sentence (39) splits the verb into two discontinuous elements, one containing the subject concord o_{CS01} '`subj-3rd-cl1`' and the present tense morpheme $a_{\text{MORPH\_pres}}$ '`pres`', the other containing the verb stem $nwa_{\text{V\_tr}}$ 'drink', cf. figures 3.10, showing a respective analysis of 39 (b).

[11]Note that 39 (b) demonstrates an independent imperfect indicative sentence which ends in a verb stem. In such a case the present tense morpheme *a* has to be present (cf. 3.2.5.1).

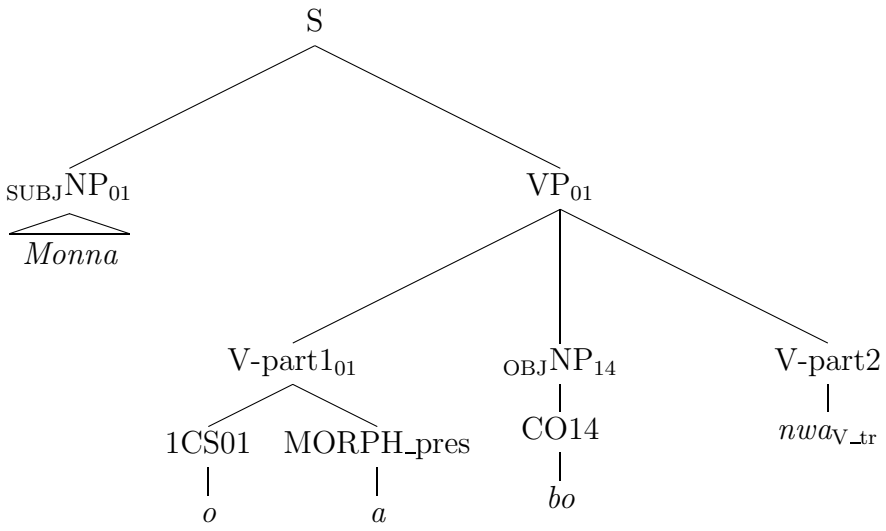


Figure 3.10: A discontinuous verb in *monna o a bo nwa* '(a) man drinks it'

Poulos and Louwrens (1994, p. 185 et seq. and p. 196) follow this methodology[12]. However, they only explicitly define the placement of the object concord into the verb for the imperative, before stating that such methodology is "used for all verb roots". At this stage, they ignore the possibility of a double transitive where an indirect object can also be represented by an object concord to be inserted at the same position[13]. If one followed their approach, the number of necessary morphosyntactic rules would multiply, as for every possible constellation of the elements of the verb, two additional rules would have to be defined describing the split forms.

Instead of splitting the verb, i.e. instead of adding morphosyntactic rules in order to describe the cases where object concords occur for all possible verbal constellations, we have opted for a new perspective on the verbal phrase for the use of disjunctively written languages like Northern Sotho, by defining a basic verb phrase (VBP) as an intermediate structure. When intransitive, the verb stem can occur on is own (formulated as a phrase rule: VBP → V), whenever transitive (or double transitive) it forms a basic verbal phrase together with its object(s). The possible constellations forming a VBP will be described in

[12]We refer only to Poulos and Louwrens (1994) here because Lombard does not explicitly define a set of rules, the issue of object concords however is briefly discussed in (Lombard, 1985, p. 103 et seq.), and it is demonstrated with some examples in the chapter on verbal constellations (Lombard, 1985, p. 139 et seq.).

[13]Note that in Northern Sotho, only one object concord may be used at a time in contrast to, for example, Setswana

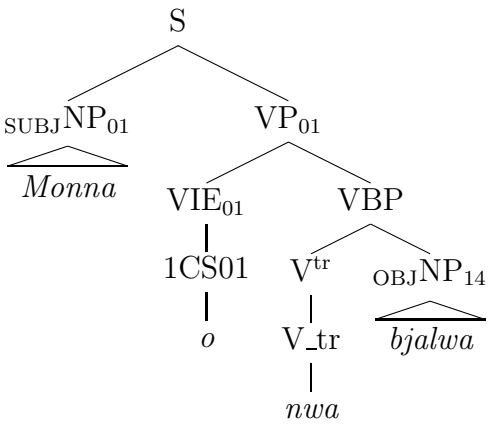


Figure 3.11: *monna o nwa bjalwa* '(a) man drinks (a) beer'

more detail in paragraph 3.2.3 (on imperatives), note that all of them may be contained in all VPs containing main verbs. Whenever inflectional morphemes occur, they are grouped in the verbal inflectional element (VIE).

As the VBP can form an independent phrase (the positive imperative), we see this perspective as both a reasonable and convenient approach to describe the verbs of Northern Sotho. Our point of view is further supported when looking at the morphosyntactic rules forming verbal phrases: The number of necessary arguments is determined by the lexical semantics of the verb root and its suffixes or suffix clusters. At the same time, this issue is independent of the other elements of the verb (VIE) which provide morphological and subject-verb agreement information. The constellations forming the VIE are the only data necessary for categorizing the VP as a whole as e.g. belonging to a certain mood and/or tense. The sentences in (39) are therefore analysed as in Figures 3.11 and 3.12.

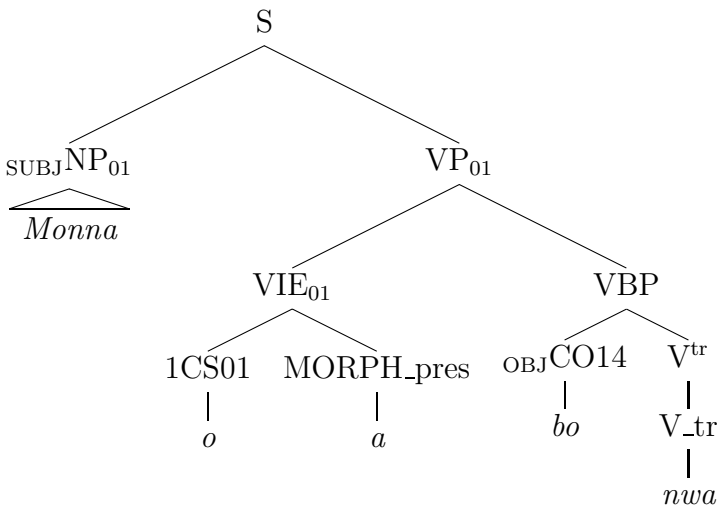


Figure 3.12: *monna o a bo nwa* '(a) man drinks it'

Table 3.5: The intransitive VBP (imperative VP)

slot zero				comments
pos-1	**pos-0**	**pos+1**	**pos+2**	**Vstem ends in**
	$\mathbf{V}^{\text{itr}}$			*–a, –ang*
	$Bolela_{\text{V\_itr}}$ speak			
	'Speak(!)'			

3.2.2 The Basic Verbal Phase (VBP)

3.2.3 The non-predicative moods: The imperative

Poulos and Louwrens (1994) distinguish between two kinds of verbs; those incorporating a subject concord and those without (cf. Tables 3.2 and 3.3 on pages 76 and 77). Lombard labels these two categories as predicative and non-predicative moods, the non-predicative moods do not contain a subject (and thereby no subject concord). The positive imperative is the simplest structure of a verb in a non-predicative mood. In the case of a positive intransitive verb, the imperative consists only of a VBP which in turn contains the verb stem, as shown in Table 3.5.
Table 3.5 demonstrates the first of the basic verbal phrases, where only the central slot pos-0 is filled with an intransitive verb. The second row shows an example.

Table 3.6 shows the use of the pre-defined slots of a transitive verb followed by its object, usually a nominal phrase (NP). Table 3.7 extends the definition of the VBP with a double transitive verb. Slot pos+1 is reserved for the use of the indirect object, slot pos+2 for the use of the direct object. Both NPs can contain one noun only, however also all other nominal structures described in section 3.8 (on nominal phrases) are possible. Some verbs subcategorise whole clauses as their object, like, e.g. the verbs of saying. These cases will be discussed at a later stage (cf. section 3.10).
Table 3.8 reflects the object noun being represented by an object concord in the slot pos-1. In the case of double transitivity, only the indirect (oblique) object may be represented by an object concord which is demonstrated in Table 3.9 (see also the discussion in paragraph 3.2.1.1).

Table 3.6: A transitive VBP (imperative VP)

slot zero				comments
pos-1	**pos-0**	**pos+1**	**pos+2**	**Vstem ends in**
	$\mathbf{V}^{\text{tr}}$		$_{\text{OBJ}}\mathbf{NP}_{\text{class}}$	*–a, –ang*
	$bulang_{\text{V\_tr}}$ open		$lemati_{\text{N05}}$ door	
'open the door(!)'				

Table 3.7: A double transitive VBP (imperative VP)

slot zero				comments
pos-1	**pos-0**	**pos+1**	**pos+2**	**Vstem ends in**
	$\mathbf{V}^{\text{dtr}}$	$_{\text{OBJIND}}\mathbf{NP}$	$_{\text{OBJ}}\mathbf{NP}$	*–a, –ang*
	$fa_{\text{V\_dtr}}$ give	$dimpša_{\text{N10}}$ dogs	$dijo_{\text{N10}}$ food	
'give the dogs food(!)'				

Table 3.8: A transitive VBP containing an object concord (imperative VP)

slot zero				comments
pos-1	**pos-0**	**pos+1**	**pos+2**	**Vstem ends in**
$_{\text{OBJ}}\mathbf{CO}_{\text{categ}}$	$\mathbf{V}^{\text{tr}}$			*–e, –eng*
le_{CO05} `obj-3rd-cl5`	$buleng_{\text{V\_tr}}$ open			
'open it(!)'				

Table 3.9: A double transitive VBP containing an object concord (imperative VP)

slot zero				comments
pos-1 $_{\text{OBJIND}}\mathbf{CO}_{\text{categ}}$	**pos-0** $\mathbf{V}^{\text{dtr}}$	**pos+1**	**pos+2** $_{\text{OBJ}}\mathbf{NP}$	**Vstem ends in** *–e, –eng*
di_{CO10} obj-3rd-cl10	$fe_{\text{V\_dtr}}$ give		$dijo_{\text{N10}}$ food	
‘give them food(!)’				

Table 3.10: The non-predicative negative imperative VP

zero-2	**zero-1**	**zero**	**Vstem ends in**
$se_{\text{MORPH\_neg}}$		**VBP**	*–e, –eng*
examples:			
$se_{\text{MORPH\_neg}}$ neg		$šome_{\text{V\_itr}}$ work	
‘do not work(!)’			
$se_{\text{MORPH\_neg}}$ neg		$reke_{\text{V\_tr}}$ $puku_{\text{N09}}$ buy book	
‘do not buy the book(!)’			

The basic verbal phrases (cf. Tables 3.5 to 3.9) are identical to the positive imperative verbal phrases. Table 3.10 describing the negative imperative is the first table where slots pos-1 to pos+2 (i.e. all contents of the VBP) are fused and shown as slot zero, demonstrating that any of the five basic verbal phrases described can occur in this position. This table shows the negative form of the imperative which is marked by the use of the negative morpheme $se_{\text{MORPH\_neg}}$, preceding the VBP and filling slot zero-2. Slot zero-1 will be used in other constellations, it is therefore left empty at the moment.

The imperative in its negative form (like the other non-predicative mood, the infinitive, cf. paragraph 3.2.4) can only fill one more slot: zero-2. Table 3.11 shows all forms of the imperative verbal phrase, $^{\text{IMP}}$VP, the optional imperative VIE, and summarises the VBPs.

Table 3.11: The imperative mood (table contains all VBPs)

	^{IMP}VIE		^{IMP}VP				
			VBP				
descr.	**zero-2**	**zero-1**	**slot zero**				**Vstem ends in**
imp.pos.			all VBPs as below				*–a, –ang*
			pos-1	**pos-0**	**pos+1**	**pos+2**	
				$\mathbf{V}^{itr}$			
				$\mathbf{V}^{tr}$		$_{OBJ}\mathbf{NP}$	
				$\mathbf{V}^{dtr}$	$_{OBJIND}\mathbf{NP}$	$_{OBJ}\mathbf{NP}$	
			$_{OBJ}\mathbf{CO}_{categ}$	$\mathbf{V}^{tr}$			
			$_{OBJIND}\mathbf{CO}_{categ}$	$\mathbf{V}^{dtr}$		$_{OBJ}\mathbf{NP}$	
imp.neg.	*se*$_{MORPH\_neg}$		all VBPs as above				*–e, –eng*

3.2.4 The non-predicative moods: The infinitive

The infinitive is introduced by the class prefix of the noun class 15 (MORPH_cp15), *go*. In terms of the morphological structure of the infinitive, Northern Sotho does not differ from the syntactic structure of English, where the infinitive also has to be introduced by a marker, the infinitive particle 'to', as was pointed out in paragraph 2.2.2.5 on page 28.

As in other languages, like, e.g. German[14], the infinitive can appear as a nominalized VP (here, of class 15). In this case, there will be the syntactic function *SUBJ* assigned by a verb, as shown in Figures 3.13 and 3.14 that demonstrate respective analyses of examples (40) and (41) . The respective verb will contain the subject concord of class 15, *go*.

(40) *go* *sepela* *go* *re* *bontšha* *mafase*
to walk `subj-3rd-cl15` `obj-2nd-pl` let-see countries
'traveling lets us see (the) world'

(41) *go* *sepela* *go* *a* *lapiša*
to walk `subj-3rd-cl15` `pres` make tired
'walking is exhausting'

[14]In German, a nominalised infinitive (*substantivierter Infinitiv*) is described as a noun derived from a (verbal) infinitive. Morphologically, this phenomenon is explained as a result of conversion (a derivation process that does not add or delete affixes), e.g. in *Schwimmen ist gesund*, 'Swimming is healthy'. Such conversion is known to happen also in other European languages, e.g. Italian. As can be seen in our example, it is the gerund that appears with a grammatical function in English.

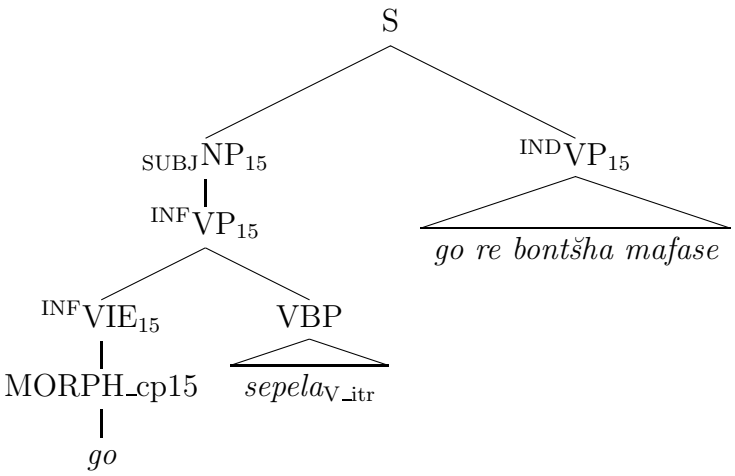


Figure 3.13: *go sepela go re bontšha mafase* 'travelling lets us see the world'

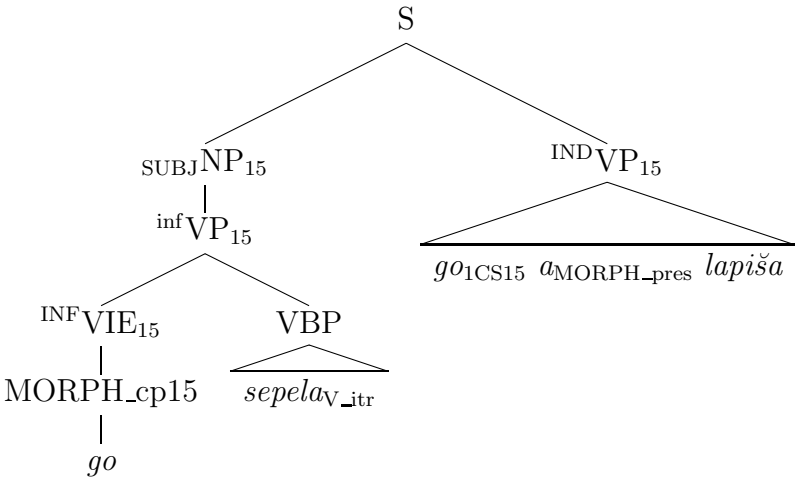


Figure 3.14: *go sepela go a lapiša* 'walking is exhausting'

Table 3.12: The infinitive

	^{INF}VP			
	^{INF}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
inf.pos.	MORPH_cp15		VBP	*–a*
	example:			
	$go_{\text{MORPH\_cp15}}$		*šoma*	
	to		work	
	'to work'			
inf.neg.	MORPH_cp15 $se_{\text{MORPH\_neg}}$		VBP	*–e*
	example:			
	$go_{\text{MORPH\_cp15}}$ $se_{\text{MORPH\_neg}}$		*šome*	
	to neg		work	
	'not to work'			

The infinitive is negated like the imperative, using the negative morpheme *se*. Lombard (1985, p. 159) states that the negative morpheme *sa* could occur "in some northern dialects" and therefore he allows this negation to be used as well. However, for the fragment of the grammar developed in this study, dialectical issues are not considered. Table 3.12 demonstrates the constellations of the infinitive with examples.

3.2.5 The predicative independent indicating mood: The indicative

The verb constellations contained in this predicative mood are indicative, as Lombard states (cf. table 3.2), i.e. these verbs form – together with their subjects, object and adverbials belonging to them – independent clauses, while the VPs of the modifying and of the dependent moods are part of dependent clauses that cannot appear on their own. In contrast to the non-predicative moods, the predicative moods appear in tenses of which two – the present tense and the future tense – use tense markers. Slot zero-1 will be filled with these.

We now come back to the idea of the three sets of subject concords described in the paragraph on subject concords (2.4.2). Each set of subject concords marks specific constellations in terms of mood, tense and actuality. Therefore the set that the subject concord belongs

to should be labelled in the morphosyntactic rules to avoid concords of inappropriate sets to be used. As a reminder, Table 2.8 of paragraph 2.4.2 is repeated as Table 3.13.
Poulos and Louwrens (1994, p. 206 et seq.) explain that there are two indicatives: the principal and participial. The principal indicatives form independent verbs, while the participial verbs cannot appear alone but refer to independent verbs in a sense that they are 'subordinate' verbs. Usually the independent principals use the concords of sets 1 and 2, the dependent participials in general use set 3. Lombard (1985, p. 152 et seq.) describes a dependent mood, the consecutive, to generally use set 3.

However, contrary to the general statement of Poulos and Louwrens (1994, p. 206 et seq.) there are some negated constellations of the principal that also use set 3 subject concords. Some are e.g. illustrated by Lombard (1985, p. 146). Consider for example sentence (42), that demonstrates one of the possible negative forms of the indicative in the perfect tense. Example sentence (43) is taken from Poulos and Louwrens (1994, p. 214) to demonstrate the same issue. Note that the tense of the examples in (42) and (43) is marked by the use of the negation morpheme cluster *ga se* occurring together with the subject concord of set 3, wa_{3CS03}, or ya_{3CS09} respectively. As the classification of the 'consecutive subject concord' is no clear-cut case, we opt for consequently calling these concords 'set 3' concords instead.

(42) *mmutla* ***ga*** ***se*** ***wa*** ***tšhaba***
hare **neg** **neg** **subj-3rd-cl3** ***flee***
'(the) hare did not flee'

(43) *mpša* ***ga*** ***se*** ***ya*** ***loma*** *mošemane*
dog **neg** **neg** **subj-3rd-cl9** **bite** boy
'(the) dog has not bitten (the) boy'

Table 3.13: The three sets of subject concords

	subject concords			
	set 1 $1CS_{categ}$	**set 2** $2CS_{categ}$	**set 3** $3CS_{categ}$	**fused forms**
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*	

3.2.5.1 The predicative independent indicating mood: The imperfect indicative

The independent moods appear in two tenses, the imperfect (Lombard, 1985, p. 141) or present tense (Poulos and Louwrens, 1994, p. 208), and the perfect tense ((Lombard, 1985, p. 141) and (Poulos and Louwrens, 1994, p. 213)). The future tense is encoded in a more aspectual way in Northern Sotho making use of a 'future tense morpheme'. The indicative 'imperfect' form is described first, i.e. the present tense, which is the only constellation which appears in two positive forms, the 'long' and the 'short' form, as shown in example (44). The 'long' form contains the present tense morpheme $a_{\text{MORPH\_pres}}$.

44(a) o_{1CS01} $\boldsymbol{a}_{\text{MORPH\_pres}}$ $ipshina_{\text{V\_sat-tr}}$
subj-3rd-cl1 pres enjoy oneself
'(s)he is enjoying herself/himself'

44(b) o_{1CS01} $tseba_{\text{V\_tr}}$ $dikgomo_{\text{N10}}$
subj-3rd-cl1 know cattle
'(s)he knows cattle'

Poulos and Louwrens (1994, p. 209) mention that 'the long form is used when the verb is the last word in the sentence'. However, there are a number of examples in the corpus data where this morpheme occurs with a verb stem that is followed by a comma, marking the end of a clause. On the other hand, this morpheme does not appear if the verb stem is not the last element of the clause or sentence.

An implementation of Northern Sotho grammar (cf. paragraph 5.2.4.2) should indeed be expected to be able to determine the occurences of *a* as a present tense morpheme. However, for the sake of clarity of definition, the VBP is marked with a label indicating whether the verb stem was the final element. The superscript 'p' (meaning that the present tense morpheme may appear) is therefore introduced and the VBP is updated, as in Table 3.14. As far as the slot system is concerned, the $a_{\text{MORPH\_pres}}$ is inserted in slot zero-1 which is explicitly reserved for one morpheme; indicating tense. Additionally, we should however take a possible clause border following the VBP (the part of speech $. indicates the appropriate punctuation), into account as adjuncts might appear which would inhibit the present tense morpheme from appearing. Table 3.15 which describes the long form, contains information about such a clause border occuring after the VBP in slot zero+1, to prevent verbal adjuncts, i.e. adverbs from appearing after the verb.

Table 3.14: Update on names of the constellations forming the VBP

description	**VBP**			
	pos-1	**pos 0**	**pos+1**	**pos+2**
VBP		$\mathbf{V}^{\text{itr}}$		
VBP		$\mathbf{V}^{\text{tr}}$		$_{\text{OBJ}}\mathbf{NP}$
VBP		$\mathbf{V}^{\text{dtr}}$	$_{\text{OBJIND}}\mathbf{NP}$	$_{\text{OBJ}}\mathbf{NP}$
$\mathbf{VBP}^{\text{p}}$	$_{\text{OBJ}}\mathbf{CO}_{\text{categ}}$	$\mathbf{V}^{\text{tr}}$		
VBP	$_{\text{OBJIND}}\mathbf{CO}_{\text{categ}}$	$\mathbf{V}^{\text{dtr}}$		$_{\text{OBJ}}\mathbf{NP}$

Table 3.15: The long form of the predicative imperfect indicative

	$^{\text{INF}}$**VIE**	$^{\text{INDPRES}}$VP **VBP**			
descr.	**zero-2**	**zero-1**	**zero**	**zero+1**	**Vstem ends in**
ind.pres. pos.long.	$1\text{CS}_{\text{categ}}$	MORPH_pres	VBP^{p}	$. (clause border)	*–a*
	example:				
	$o_{1\text{CS01}}$	$a_{\text{MORPH\_pres}}$	$ipshina_{\text{V\_sat-tr}}$		
	`subj-3rd-cl1`	`pres`	enjoy oneself		
	'(s)he is enjoying herself/himself'				

The short form of the indicative present tense and its negation is presented in Table 3.16. The positive short form appears whenever the clause does not end after the verb stem. The positive forms of the indicative present tense make use of the subject concords of the first set, the negative form uses the subject concords of the second, i.e. *a* instead of *o* in class 1.

Table 3.16: The short and the negated form of the predicative imperfect indicative

	$^{\text{INDPRES}}$VP				
	$^{\text{INDPRES}}$**VIE**		**VBP**		
descr.	**zero-2**	**zero -1**	**zero**	**zero +1**	**Vstem ends in**
ind.pres. pos.short	$1CS_{categ}$		VBP	¬$.	*−a*
	example:				
	o_{1CS01}		$tseba_{V\_tr}$ $dikgomo_{N10}$		
	`subj-3rd-cl1`		knows cattle		
	‘(s)he knows cattle’				
ind.pres. neg.	ga_{MORPH_neg} $2CS_{categ}$		VBP		*−e*
	example:				
	ga_{MORPH_neg} a_{2CS01}		$tsebe_{V\_tr}$ $dikgomo_{N10}$		
	`neg` `subj-3rd-cl1`		know cattle		
	‘(s)he does not know cattle’				

3.2.5.2 The predicative independent indicating mood: The perfect indicative

The positive perfect tense is solely identified by certain verbal endings, as demonstrated by the examples in (45), cf. ((Lombard, 1985, p. 146)). This tense can therefore only be determined automatically when the ending of the verb stem is taken into account (cf. paragraph 5.1.2.3 beginning on page 229 for a sample analysis). As described in paragraph 2.7.2, the ending *–ile* will be mentioned in the morphosyntactic rules of this chapter as referring to all its possible allomorphs.

45 (a) *mmutla* *o* *tšhab**ile***
hare `subj-3rd-cl3` fled
'(a) hare fled'

(b) *lesogana* *le* *bole**tše***
young man `subj-3rd-cl5` spoke
'(a) young man spoke'

There are several ways to negate a perfect tense indicative, as Table 3.17 demonstrates. Note that perfect form 4 shown in Table 3.17 is the only described constellation containing the token *a* as MORPH_past (cf. paragraph 2.9.2), classified by Lombard as the 'perfective/stative *–a–*'. Poulos and Louwrens (1994, p. 214) mention the occurrence of *a* in this constellation, too, however, they do not give it a label.

Table 3.17: The perfect indicative

	$^{\text{INDPERF}}$VP		
	$^{\text{INDPERF}}$**VIE**	**VBP**	
descr.	**zero-2(slot zero-1 not used)**	**zero**	**Vstem ends in**
ind.perf.pos.	$1\text{CS}_{\text{categ}}$	VBP	e.g. *–ile*
	example: $o_{1\text{CS}03}$ `subj-3rd-cl3` 'it fled'	$\text{tšhabile}_{\text{V\_itr}}$ fled	
indicative perf.neg. 1	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$	VBP	*–a*
	example: $ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $wa_{3\text{CS}03}$ `neg neg subj-3rd-cl3` 'it did not flee'	$tšhaba_{\text{V\_itr}}$ flee	
indicative perf.neg. 2	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $2\text{CS}_{\text{categ}}$	VBP	*–e*
	example: $ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $o_{2\text{CS}03}$ `neg neg subj-3rd-cl3` 'it did not flee'	$tšhabe_{\text{V\_itr}}$ flee	
indicative perf.neg. 3	$ga_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$	VBP	*–a*
	example: $ga_{\text{MORPH\_neg}}$ $wa_{3\text{CS}03}$ `neg subj-3rd-cl3` 'it did not flee'	$tšhaba_{\text{V\_itr}}$ flee	
indicative perf.neg. 4	$ga_{\text{MORPH\_neg}}$ $1\text{CS}_{\text{categ}}$ $a_{\text{MORPH\_past}}$	VBP	*–a*
	example: $ga_{\text{MORPH\_neg}}$ $o_{2\text{CS}03}$ $a_{\text{MORPH\_past}}$ `neg subj-3rd-cl3 perf` 'it did not flee'	$tšhaba_{\text{V\_itr}}$ flee	

3.2.5.3 The predicative independent indicating mood: The future indicative

The positive form of the future 'tense' is clearly marked in Northern Sotho by the future tense morphemes $tlo_{\text{MORPH\_fut}}$, or $tla_{\text{MORPH\_fut}}$, as demonstrated in 46 (a). Either one of these morphemes appears between the subject concord and the verb root (according to Lombard (1985, p. 147) and Poulos and Louwrens (1994, p. 220)).

The negation of the future tense makes use of the potential morpheme $ka_{\text{MORPH\_pot}}$ (as described by Lombard (1985, ibid.)); for the potential forms, cf. 3.5.2). It is demonstrated in 46 (b) (for sake of convenience, we repeat example (24) of paragraph 2.9.4 here). Both examples are taken from Lombard (1985, ibid.).

46 (a) *mmutla* *o* ***tlo*** *tšhaba*
hare `subj-3rd-cl3` `fut` flee
'(a) hare will flee'

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

Table 3.18 shows how the future tense constellations fill the predefined slots. The future tense morpheme occupies slot zero-1. The negative form uses subject concords of set 2, as described by Lombard (1985, p. 147) and Poulos and Louwrens (1994, p. 212), who also mention auxiliary constructions containing a future aspect, which will be referred to at a later stage. Subject concord and negation cluster fill slot zero-2.

Table 3.18: The future indicative

	$^{\text{INDFUT}}$VP			
	$^{\text{INDFUT}}$**VIE**		**VBP**	comments
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
ind.fut. pos.	$1\text{CS}_{\text{categ}}$	$tlo/tla_{\text{MORPH\_fut}}$	VBP	$-a$
	example:			
	$o_{1\text{CS03}}$	$tlo_{\text{MORPH\_fut}}$	$tšhaba_{\text{V\_itr}}$	
	`subj-3rd-cl3`	`fut`	flee	
	'It will flee'			
ind.fut. neg.	$2\text{CS}_{\text{categ}}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$		VBP	$-e$
	example:			
	$o_{2\text{CS03}}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$		$tšhabe_{\text{V\_itr}}$	
	`subj-3rd-cl3` `pot` `neg`		flee	
	'It will not flee'			

Table 3.19: A summary of the independent indicative forms

	$^{\text{INDPRES}}$VP					
	VIE			**VBP**		
descr.	**zero-2**		**zero-1**	**zero**	**zero+1**	**Vstem ends in**
pres.pos.long	$1CS_{categ}$		MORPH_pres	VBP^{p}	\$.	*–a*
pres.pos.short	$1CS_{categ}$			VBP	¬\$.	*–a*
pres.neg.	ga_{MORPH_neg}	$2CS_{categ}$		VBP		*–e*
perf.pos.	$1CS_{categ}$			VBP		*–ile*
perf.neg. 1	ga_{MORPH_neg} $3CS_{categ}$	se_{MORPH_neg}		VBP		*–a*
perf.neg. 2	ga_{MORPH_neg} $2CS_{categ}$	se_{MORPH_neg}		VBP		*–e*
perf.neg. 3	ga_{MORPH_neg}	$3CS_{categ}$		VBP		*–a*
perf.neg. 4	ga_{MORPH_neg} a_{MORPH_past}	$1CS_{categ}$		VBP		*–a*
fut.pos	$1CS_{categ}$		*tlo/tla* $_{MORPH\_fut}$	VBP		*–a*
fut.neg	$2CS_{categ}$ se_{MORPH_neg}	ka_{MORPH_pot}		VBP		*–e*

3.2.5.4 The predicative independent indicating mood: A summary

Table 3.19 summarises all inflectional elements of the independent indicative. In some cases, only the verb ending provides information on the tense of the constellation (like e.g. the perfect positive form). In others, it is solely the verbal inflectional element (VIE) which provides information on the tense of the VP (like e.g. the long present tense form or the future tense forms).

3.2.6 The predicative independent modifying moods: The situative

Lombard (1985, p. 147) classifies the situative mood as "independent". He states that it "indicates the situation or circumstances under which another action or other actions take place" and lists example (47). Note however that though this mood is classified as 'independent', clauses appearing in this mood should not be confused with independent main clauses, as Lombard (1985, p. 140) explicitly states that the modifying moods "do not appear in main clauses"' (cf. paragraph 3.2.1 and Table 3.1). Poulos and Louwrens (1994, p. 221) define these constellations as 'participial' and as using the subject concords from set 2. These 'participials' occur in subordinate (embedded) sentences. As such, Poulos and Louwrens (ibid.) describe the participial as being dependent in nature.

The situative clauses are often introduced by the conjunctions *ge*$_{\text{CONJ}}$ 'when'[15] or *le ge* 'although'.

(47)

ge	*ba*	*bona*	*noga*	*ba*	*a*	*e*	*bolaya*
when	`subj-3rd-cl2`	see	snake	`subj-3rd-cl2`	`pres`	`obj-3rd-cl9`	kill

'when they see (a) snake, they kill it'

In another example, (48), (cf. Poulos and Louwrens (1994, p. 222)) the situative *ba tsena* 'they entered' is also introduced by the conjunction *ge*$_{\text{CONJ}}$ and preceded by an indicative clause, *ke ba bone* 'I saw them'. Figure 3.15 shows its possible morphosyntactic analysis. Situative clauses may however also be introductory sentences, as in (47) and (49), a possible analysis of (49) is demonstrated in Figure 3.16.

If the situative occurs without an introductory conjunction, it is no trivial task for a grammar to identify it as such, because it can easily be confused with the indicative. The only obvious difference between e.g. a positive indicative in the present tense and its respective situative is the use of the subject concords of set 1 vs. set 2 as used by the situative, however, these are in most cases identical. Example (50) and Figure 3.17 (from the Bible) show a situative without an introducing conjunction.

In summary, a verb in the situative mood is categorised as an adjunct verbal clause, follow-

[15]Poulos and Louwrens (1994, p. 222) in some cases translated *ge*$_{\text{CONJ}}$ as 'if' or 'if/when', in this study, however, it is solely translated as 'when' according to the dictionary De Schryver (2007).

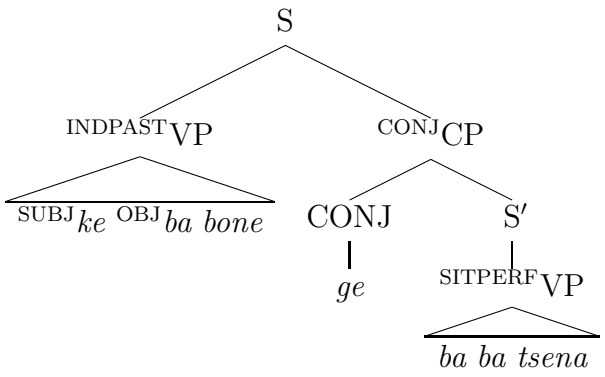


Figure 3.15: *ke ba bone ge ba tsena* 'I saw them when they entered'

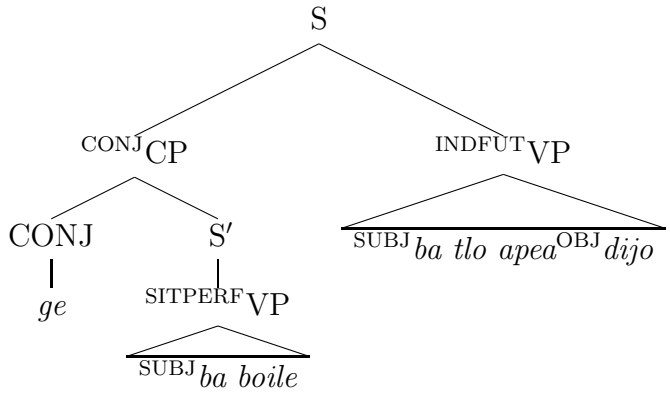


Figure 3.16: *ge ba boile ba tlo apea dijo* 'when they have returned they will cook the food'

ing or preceding a matrix clause which is in the indicative mood. It usually is embedded in a conjunctive clause ($^{\text{CONJ}}$CP) . There might however occur cases where there is no overt conjunction, here, a detection of the sentence as a situative proves difficult, as the surface forms are in most cases identical to the indicative. In terms of its semantics, the situative supplies additional information to the situation/action described by the matrix clause, i.e. it modifies the verb of the matrix clause, like any other adverbial clause.

(48) *ke* `subj-1st-sg` *ba* `obj-3rd-cl2` *bone* saw *ge* when ***ba*** `subj-3rd-cl2` ***tsena*** entered
'I saw them (the moment) when they entered'

(49) *ge* ***ba*** `subj-3rd-cl2` ***boile*** returned *ba* `subj-3rd-cl2` *tlo* `fut` *apea* cook *dijo* food
'when they have returned they will cook (the) food'

(50) *Jesu* o_{1CS01} *tlo* *boa* *ka* *wona* *mokgwa* *wo*
Jesus `subj-3rd-cl1` `fut` return by `emp-3rd-cl3` manner `dem-3rd-cl3`
le *mmonego* ***a***$_{2CS01}$ ***eya*** ***legodimong***
`subj-2nd-pl` who-him-see `subj-3rd-cl1` **go** **heaven**–`loc`
'Jesus will return (the) very same way as you saw him when going to heaven'

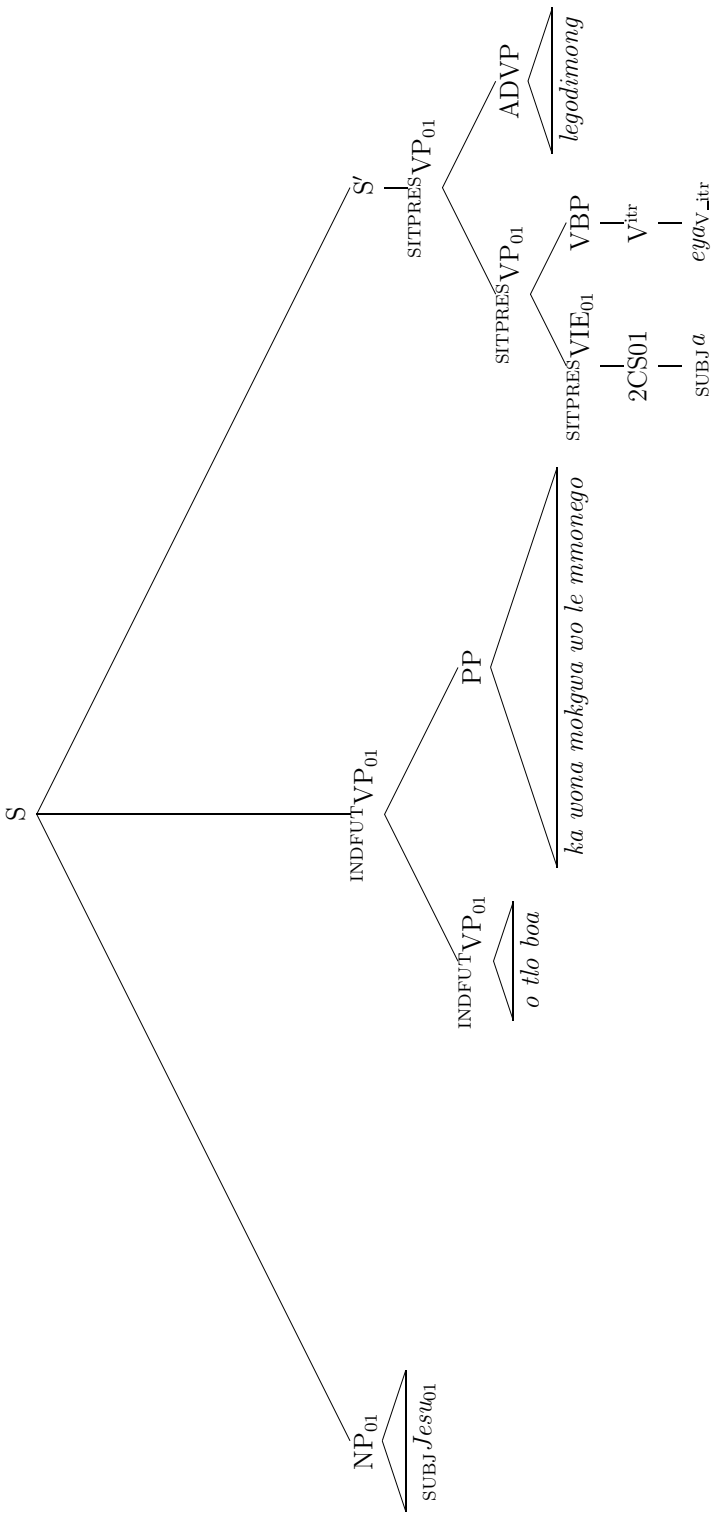


Figure 3.17: *Jesu o tlo boa ka wona mokgwa wo le mmonego a eya legodimong* 'Jesus will return (the) very same way as you saw him when going to heaven'

3.2.6.1 The predicative independent modifying moods: The imperfect situative

The positive imperfect situative contains a subject concord of the second set and the verb stem ends in *–a* as is also the case with the indicative. Table 3.20 demonstrates the constellations of the imperfect tense and shows that the negation is formed by using the negation morpheme $\text{sa}_{\text{MORPH\_neg}}$ (the verb stem in this case ends in *–e*). Again, we must stress the fact that all negation morphemes are labelled identically, hence our definition must explicitly state the type(s) used.

Table 3.20: The present tense of the situative

	$^{\text{SITPRES}}$VP			
	$^{\text{SITPRES}}$**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
sit.pres.pos.	$\text{2CS}_{\text{categ}}$		VBP	*–a*
	example:			
	a_{2CS01}		$boa_{\text{V\_itr}}$	
	`subj-3rd-cl1`		return	
	'(s)he returns'			
sit.pres.neg.	$\text{2CS}_{\text{categ}}$ $sa_{\text{MORPH\_neg}}$		VBP	*–e*
	example:			
	a_{2CS01} $sa_{\text{MORPH\_neg}}$		$boe_{\text{V\_itr}}$	
	`subj-3rd-cl1 neg`		return	
	'(s)he does not return'			

3.2.6.2 The predicative independent modifying moods: The perfect situative

The perfect tense of the situative is, according to Lombard (1985, p. 149) "identical to that of the indicative". However, in our definition, this mood always uses the subject concords of set 2 ($2CS_{categ}$), hence there is a difference, although it only shows with the subject concords of class 1. Lombard (ibid.) moreover describes three ways to negate the situative of the perfect tense as shown in Table 3.21. The first makes use of the subject concords of set 3 ($3CS_{categ}$) preceded by the negative morpheme *se*. The second uses subject concords of set 2 ($2CS_{categ}$), one preceding the negative morpheme, one following it, while the third form of the negated perfect tense also begins with a subject concord of set 2, however, it uses the negation morpheme *sa* and no second subject concord occurs following the negative morpheme. Poulos and Louwrens (1994, p. 220) describe this third form as the only one applicable to participial negated verbs.

Table 3.21: The perfect tense of the situative

	$^{\text{SITPERF}}$VP			
	$^{\text{SITPERF}}$**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
sit.perf.pos.	$2CS_{categ}$		$_{perf}$VBP	*ile*
	example: di_{2CS10} `subj-3rd-cl10` 'they grazed'		$futše_{V\_itr}$ grazed	
sit.perf.neg.1	$2CS_{categ}$ se_{MORPH_neg} $3CS_{categ}$		VBP	*–a*
	example: di_{2CS10} se_{MORPH_neg} $tša_{3CS10}$ `subj-3rd-cl10 neg subj-3rd-cl10` 'they did not graze'		$fula_{V\_itr}$ graze	
sit.perf.neg.2	$2CS_{categ}$ se_{MORPH_neg} $1CS_{categ}$		VBP	*–a*
	example: di_{2CS10} se_{MORPH_neg} di_{1CS10} `subj-3rd-cl10 neg subj-3rd-cl10` 'they did not graze'		$fule_{V\_itr}$ graze	
sit.perf.neg.3	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–a*
	example: di_{2CS10} sa_{MORPH_neg} `subj-3rd-cl10 neg subj-3rd-cl10` 'they did not graze'		$fula_{V\_itr}$ graze	

Table 3.22: The future situative

	$^{\text{SITFUT}}$VP			
	$^{\text{SITFUT}}$**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
sit.fut.pos.	$2CS_{categ}$	$tlo/tla_{\text{MORPH\_fut}}$	VBP	*–a*
	example: di_{2CS10} `subj-3rd-cl10`	$tla_{\text{MORPH\_fut}}$ `fut`	$fula_{\text{V\_itr}}$ graze	
	‘they will graze’			
ind./sit. fut.neg	$2CS_{categ}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$		VBP	*–e*
	example: di_{2CS01} $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$ `subj-3rd-cl10 pot neg`		$fule_{\text{V\_itr}}$ graze	
	‘they will not graze’			

3.2.6.3 The predicative independent modifying moods: The future situative

The future tense/aspect is described by Lombard (1985, p. 149) as being identical to the indicative, “but the subject concord of class 1 is *a*”. However, note that this is also the case with the negative form of the future indicative. The negative form of the future indicative cannot therefore be distinguished from the negative form of the future situative unless a conjunction is present. Table 3.22 details the possible forms.

3.2.7 The predicative independent modifying moods: The relative

Like the situative, the relative mood is to be categorised as a dependent clause because it adds information (to a noun) and cannot stand alone. The relative is marked twice, first by an introductory demonstrative concord, called “relative particle” by Lombard (1985, p. 150), secondly by the verb stem itself which ends in one of the verbal endings *–ago/–ang* for the positive and *–ego* or *–eng* for the negative constellations.

Poulos and Louwrens (1994, p. 103) do not consider the relative to be a mood, but rather categorise it as a "qualificative". They distinguish between a verbal and a nominal relative where a introductory demonstrative is followed by a noun, cf. example (51). Other qualificatives, according to (Poulos and Louwrens, 1994, p. 90 et seq.) are adjectives, possessives, and enumeratives.

(51) *monna* ***yo*** ***maatla***
man `dem-3rd-cl1` strength
'(a) strong man'

However, the constellation Lombard describes (ibid.) is mirrored in their verbal relative, where the demonstrative concord is called a "basic demonstrative". In using an introductory element in its relative clauses, Northern Sotho is similar to other languages, cf. example (52) from Poulos and Louwrens (1994, p. 105).

(52) *monna* ***yo*** ***a*** ***go*** ***bitšago***, *o* *tseba*
man `dem-3rd-cl1` `subj-3rd-cl1` `obj-2nd-sg` who-call, `subj-3rd-cl1` know
tate
father
'(the) man who is calling you, knows (my) father'

A few sentences appear in the text collection of the *University of Pretoria Sepedi Corpus* (PSC), (cf. De Schryver and Prinsloo (2000)) where there is no subject concord as part of the relative verb (cf. (53). Such sentences are usually considered acceptable by native speakers. However, all authors of the consulted literature see the subject concord as being a mandatory part of all predicative verbs, meaning a closer examination of this phenomenon would be necessary before taking any further steps towards its acceptance. The grammar described in this study will not include such constellations for the moment.

(53) $motho_{\text{N01}}$ yo_{CDEM01} $tsebago_{\text{V\_itr}}$ $therešo_{\text{N09}}$
man `dem-3rd-cl1` knows-rel truth
'(the) man who knows the truth'

The Northern Sotho relative clause constellations are defined similarly to a possible analysis of the English relative clause, as it is demonstrated in example (54) which is illustrated in Figures 3.18 and 3.19 .

(54) *kgoši* ***ye*** ***e*** ***bušago***
chief `dem-3rd-cl9` `subj-3rd-cl9` reigns-rel
'(a) chief who reigns'

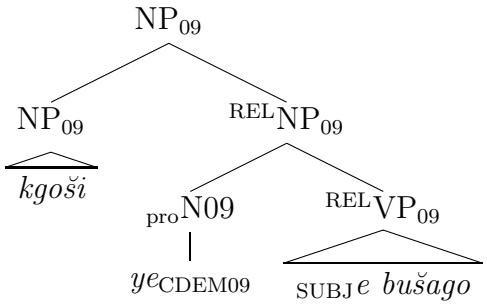


Figure 3.18: *kgoši ye e bušago* '(a) chief who reigns'

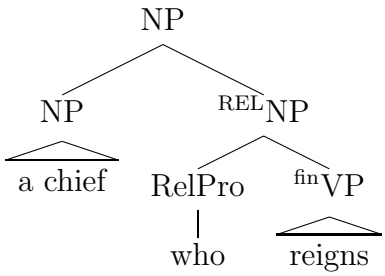


Figure 3.19: Analysis of the English relative: 'a chief who reigns'

Poulos and Louwrens (1994, p. 104 et seq.) distinguish between a "direct" and an "indirect" relative. In (54), the subject of the relative clause, represented by its subject concord, e_{1CS09}, refers to the same entity as the preceding noun ($kgoši_{N09}$), a certain chief or king. These relative clauses are therefore called direct relatives. In the case of the indirect relative, the relative clause refers to the preceding noun as its object, i.e. its subject concord refers to another entity, as in (55), illustrated in Figure 3.20. The constellation as a whole constitutes an NP.

(55) *thuthuthu* ***ye*** ***monna*** ***a*** ***e*** ***ratago***
Motorbike `dem-3rd-cl9` man `subj-3rd-cl1` `obj-3rd-cl9` who-like
'(a) motorbike that (a) man likes'

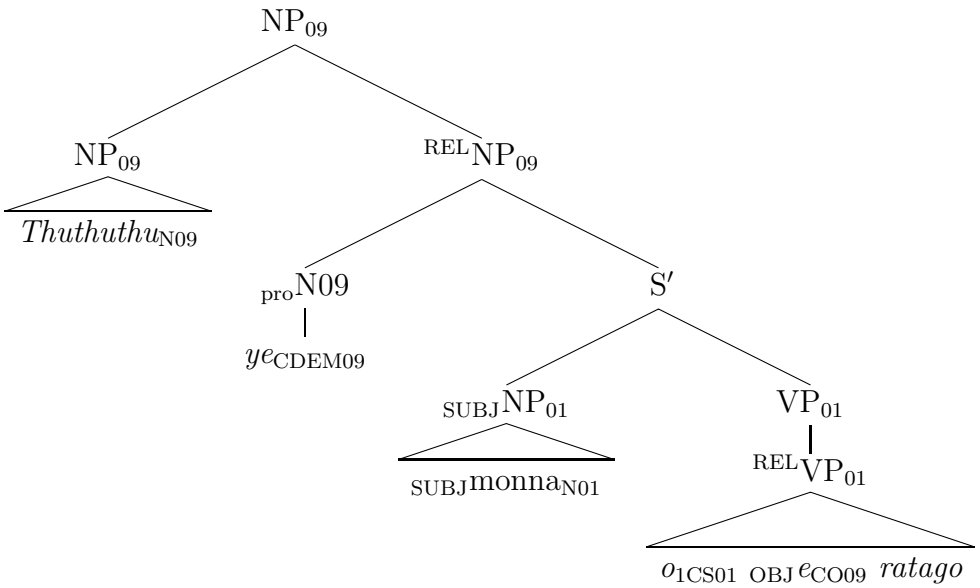


Figure 3.20: *Thuthuthu ye monna o e ratago* '(a) motorbike that (a) man likes'

3.2.7.1 The predicative independent modifying moods: The imperfect relative

Table 3.23 shows the possible constellations of the imperfect or present tense relative. As a dependent clause it makes use of the subject concord of set 2. In a smiliar way to the situative being introduced by a conjunction, the relative is usually introduced by a demonstrative concord. As the demonstrative can also introduce constellations other than the verbal relative (cf. example 51) it is seen as an independent element, separate from the morphosyntactic rules in Table 3.23. The clause as a whole will be defined in paragraph 3.8.4.1.

Table 3.23: The present tense of the relative

	$^{\text{RELPRES}}$**VIE**		$^{\text{RELPRES}}$VP **VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
rel.pres.pos.	**2CS**$_{\text{categ}}$		VBP	*−a* + ‘relative’
	example:			
	a_{2CS01}		$ikemelago_{\text{V\_sat-tr}}$	
	`subj-3rd-cl1`		who-defend-oneself	
	‘(s)he who defends himself/herself’			
rel.pres.neg.	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$		VBP	*−e* + ‘relative’
	example:			
	a_{2CS01} $sa_{\text{MORPH\_neg}}$		$ikemelego_{\text{V\_sat-tr}}$	
	`subj-3rd-cl1 neg`		who-defend-oneself	
	‘(s)he who does not defend himself/herself’			

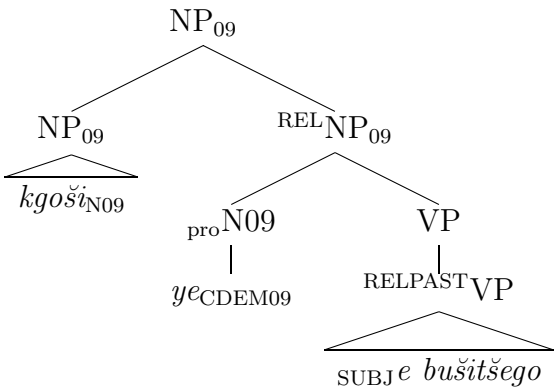


Figure 3.21: *kgoši ye e bušitšego* '(a) chief who reigned'

3.2.7.2 The predicative independent modifying moods: The perfect relative

The positive perfect tense can only be distinguished from the imperfect tense by determining a perfect/perfect tense suffix of the relative. Lombard (1985, p. 151) mentions the perfect tense form of (54) in example (56) (Figure 3.21).

(56)	*kgoši*	***ye***	***e***	***bušitšego***
	chief	`dem-3rd-cl9`	`subj-3rd-cl9`	who-reigned
	'(a) chief who reigned'			

The negative of the perfect tense appears in two constellations. The first one uses the subject concord of set 2, followed by one of the negation morphemes *sego* or *seng*, followed by another subject concord of the same class contained in set 3. The verb stem concluding this form does not show the relative ending. The second form of the negated relative uses the subject concord of set 2, followed again by one of the negation morphemes *sego* or *seng*, followed by another subject concord of the same class again belonging to set 2, the verb stem here ends in *e*. Both negative forms are therefore clearly distinguishable from the negative forms of the imperfect. Examples 57 (a) and (b) are the negated forms of (56). Table 3.24 on page 118 summarises all perfect tense constellations of the relative.

57 (a)	*kgoši*	***ye***	***e***	***sego***	***ya***	***buša***
	chief	`dem-3rd-cl9`	`subj-3rd-cl9`	`rel-neg`	`subj-3rd-cl9`	reign
	'(a) chief who did not reign'					
(b)	*kgoši*	***ye***	***e***	***sego***	***e***	***buše***
	chief	`dem-3rd-cl9`	`subj-3rd-cl9`	`rel-neg`	`subj-3rd-cl9`	reign
	'(a) chief who did not reign'					

Table 3.24: The perfect tense of the relative

	$^{\text{RELPERF}}$VP		
	$^{\text{RELPERF}}$**VIE**	**VBP**	
descr.	**zero-2**	**zero**	**Vstem ends in**
rel.perf.pos.			
	$2CS_{categ}$	$_{perf}$VBP	'$-a$ + relative'
	example: di_{2CS10} `subj-3rd-cl10`	$fulago_{V\_itr}$ who-grazed	
	'they who grazed'		
rel.perf.neg. 1			
	$2CS_{categ}$ $sego/seng_{MORPH\_neg}$ $3CS_{categ}$	VBP	$-a$
	example: di_{1CS10} $sego_{MORPH\_neg}$ $t\check{s}a_{3CS01}$ `subj-3rd-cl10 rel-neg subj-3rd-cl10`	$fula_{V\_itr}$ graze	
	'they who did not graze'		
rel.perf.neg. 2			
	$2CS_{categ}$ $sego/seng_{MORPH\_neg}$ $2CS_{categ}$	VBP	$-e$
	example: di_{1CS10} $seng_{MORPH\_neg}$ di_{2CS01} `subj-3rd-cl10 rel-neg subj-3rd-cl10`	$fule_{V\_itr}$ graze	
	'they who did not graze'		

3.2.7.3 The predicative independent modifying moods: The future relative

There are two options for the future relative constellations, both make use of the relative form *-go*, however, in the first one it appears with the future tense morpheme (as $tlago_{\text{MORPH\_fut}}$ or $tlogo_{\text{MORPH\_fut}}$), in the second, it appears with the verb like in the other relative constellations previously described. Again, the subject concords of set 2 ($2\text{CS}_{\text{categ}}$) are used.

The negation of the future tense contains the potential morpheme and the future tense morpheme which ends in one of the relative endings. In other cases, negations of the future tense of the relative clause also appear without the future tense morpheme, comparable to the negated forms of the future tense indicative. Table 3.25 on page 120 shows all forms.

3.2.8 A summary of the indepedent modifying moods

Table 3.26 summarises all verbal constellations of the modifying moods in order to complement the information given in Table 3.19 summarising the independent moods. With this summary, all basic verbal independent constellations are described according to Lombard (1985). The following paragraphs will focus on the dependent moods of Northern Sotho.

Table 3.25: The future relative

		$^{\text{RELFUT}}$**VIE**		$^{\text{RELFUT}}$VP **VBP**	
descr.	**zero-2**		**zero-1**	**zero**	**Vstem ends in**
rel.fut.pos 1					
	$2CS_{categ}$		*tlago/tlogo* $_{MORPH\_fut}$	VBP	*–a*
	example:				
	di_{2CS10}		*tlogo* $_{MORPH\_fut}$	$fula_{V\_itr}$	
	`subj-3rd-cl10`		`fut`	graze	
	‘they who will graze’				
rel.fut.pos 2					
	$2CS_{categ}$		*tla/tlo* $_{MORPH\_fut}$	VBP	*–a* + ‘relative’
	example:				
	di_{2CS10}		*tlo* $_{MORPH\_fut}$	$fulago_{V\_itr}$	
	`subj-3rd-cl10`		`fut`	who-graze	
	‘they who will graze’				
rel.fut.neg 1					
	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}		*tlago/tlogo* $_{MORPH\_fut}$	VBP	*–a*
	example:				
	di_{2CS01} ka_{MORPH_pot} se_{MORPH_neg}		*tlogo* $_{MORPH\_fut}$	$fula_{V\_itr}$	
	`subj-3rd-cl10 pot neg`		`fut`	graze	
	‘they who will not graze’				
rel.fut.neg 2					
	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}		*tla/tlo* $_{MORPH\_fut}$	VBP	*–a* + ‘relative’
	example:				
	di_{2CS01} ka_{MORPH_pot} se_{MORPH_neg}		*tlo* $_{MORPH\_fut}$	$fulago_{V\_itr}$	
	`subj-3rd-cl10 pot neg`		`fut`	who-graze	
	‘they who will not graze’				
rel.fut.neg 3					
	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}			VBP	*–a* + ‘relative’
	example:				
	ba_{2CS02} ka_{MORPH_pot} se_{MORPH_neg}			$fulago_{V\_itr}$	
	`subj-3rd-cl10 pot neg`			who-graze	
	‘they who will not graze’				

Table 3.26: A summary of the modifying moods

	^{MOD}VP			
	^{MOD}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
sit.pres.pos.	$2CS_{categ}$		VBP	*–a*
sit.pres.neg.	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–e*
sit.perf.pos.	$2CS_{categ}$		VBP	*–ile*
sit.perf.neg.1	$2CS_{categ}$ se_{MORPH_neg} $3CS_{categ}$		VBP	*–a*
sit.perf.neg.2	$2CS_{categ}$ se_{MORPH_neg} $1CS_{categ}$		VBP	*–a*
sit.perf.neg.3	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–a*
sit.fut.pos.	$2CS_{categ}$	*tlo/tla* $_{MORPH\_fut}$	VBP	*–a*
sit.fut.neg.	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}		VBP	*–e*
rel.pres.pos.	$2CS_{categ}$		VBP	*–a* + 'relative'
rel.pres.neg.	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–e* + 'relative'
rel.perf.pos.	$2CS_{categ}$		VBP	*–ile* + *–a* + 'relative'
rel.perf.neg.1	$2CS_{categ}$ *sego/seng* $_{MORPH\_neg}$ $3CS_{categ}$		VBP	*–a*
rel.perf.neg.2	$2CS_{categ}$ *sego/seng* $_{MORPH\_neg}$ $2CS_{categ}$		VBP	*–e*
rel.fut.pos.1	$2CS_{categ}$	*tlago/tlogo* $_{MORPH\_fut}$	VBP	*–a*
rel.fut.pos.2	$2CS_{categ}$	*tla/tlo* $_{MORPH\_fut}$	VBP	*-a* + 'relative'
rel.fut.neg.1	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	*tlago/tlogo* $_{MORPH\_fut}$	VBP	*–a*
rel.fut.neg.2	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	*tla/tlo* $_{MORPH\_fut}$	VBP	*–a* + 'relative'
rel.fut.neg.3	$2CS_{categ}$ ka_{MORPH_pot}	se_{MORPH_neg}	VBP	*–a* + 'relative'

Table 3.27: The consecutive

	^{CONS}VP			
	^{CONS}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
cons.pos.	$3CS_{categ}$		VBP	*–a*
	example:			
	$ka_{3CSPERS\_sg}$		$reka_{V\_tr}$ $maswi_{N06}$	
	`subj-1st-sg`		buy milk	
	'I buy/bought/will buy milk'			
cons.neg.	$3CS_{categ}$ se_{MORPH_neg}		VBP	*–e*
	example:			
	$ka_{3CSPERS}$ se_{MORPH_neg}		$reke_{V\_itr}$ $maswi_{N06}$	
	`subj-1st-sg neg`		buy milk	
	'I do not / did not / will not / buy milk '			

3.2.9 The predicative dependent moods: The consecutive

The consecutive is described by Lombard (1985, p. 152) as a dependent mood, indicating "an action or process which follows another action/other actions." Poulos and Louwrens (1994, p. 240 et seq.) describe the consecutive accordingly. Its identifying element is the subject concord of set 3, and its chronological placement is dependent on the matrix sentence it refers to as it does not contain tense marker itself. Example (58) (cf. (Lombard, 1985, p. 152)) demonstrates a consecutive referring to an event in the past. The consecutive itself does not show any tense, the matrix sentence however is an indicative of the perfect tense.

(58) *ke* *ile* *toropong* ***ka*** ***reka*** ***maswi***
`subj-1st-sg` went `town-loc` `subj-1st-sg` buy milk
'I went to town (and then) I bought milk'

Lombard (1985, p. 154) demonstrates one possible negation of the consecutive containing the negative morpheme se_{MORPH_neg}. Poulos and Louwrens (1994, p. 241) add a second possibility to negate the consecutive, a compound form containing an auxiliary. However, we do not expect this form to be frequent, as in our corpus no occurrences were found. Therefore this form does not appear in Table 3.27 demonstrating the consecutive forms.

3.2.10 The predicative dependent moods: The subjunctive and the habitual

The subjunctive and the habitual are, like the consecutive, described by both sources as forming dependent clauses. The subjunctive, according to Lombard (1985, p. 154), is 'conditioned causally', i.e. the action described in this clause was caused by the action described in the matrix clause, while the habitual, according to Lombard (1985, ibid.), 'indicates an action/process which proceeds as a habit from previous actions/processes'.

Example 59 (a) shows a subjunctive, (b) a habitual. The elements of 59 (a) and (b) are annotated with the relevant labels to demonstrate that the forms in which both these moods appear are orthographically identical[16]. However, the subjunctive may be preceded by a conjunction, *gore* 'that' (cf. 59 (c)), while the habitual might be part of an auxiliary structure, as shown in 59(d). All examples in (59) are taken from Lombard (1985, p. 154 et seq.) ((b) is simplified for the sake of clarity).

59 (a) *ke*$_{\text{2CScateg}}$ *mmoditše*$_{\text{VBP−V−end−e}}$ ***a*** ***lahle*** ***selo*** ***seo***
`subj-1st-sg` `obj-3rd-cl1-told` **`subj-3rd-cl1`** **must throw away** **thing** **that**
'I told him to throw that thing away (that he should throw that thing away)'

(b) *re* *tsoga* ***re***$_{\text{2CScateg}}$ ***apare***$_{\text{VBP−−V−−end−−e}}$
`subj-1st-pl` get up **`subj-1st-pl`** **get dressed**
'we get up (and usually) we get dressed'

(c) *Ke* *nyaka* ***gore*** ***o***$_{\text{2CScateg}}$ ***nthuše***$_{\text{VBP−−V−−end−−e}}$
`subj-1st-sg` want **that** **`subj-2nd-sg`** **`obj-1st-sg-help`**
'I want you to help me (that you should help me)'

(d) ***o***$_{\text{1CScateg}}$ ***tle***$_{\text{V\_aux}}$ ***a***$_{\text{2CScateg}}$ ***fihle***$_{\text{VBP−−V−−end−−e}}$ *ka* *Mokibelo*
`subj-3rd-cl1` **usually** **`subj-3rd-cl1`** arrive by Saturday
'he usually arrives on (a) Saturday'

The negated forms of the subjunctive and the habitual entail both the negation morpheme *se*$_{\text{MORPH\_neg}}$ and, like the positive, subject concords of the second set. All constellations of the subjunctive/habitual forms are shown in Table 3.28, page 124.

[16]Poulos and Louwrens (1994, p. 244) however point out that there is indeed a difference in tone: while the verb stem ends in *–e* in the habitual, it ends in *-ê* in the subjunctive. However, tones are not marked in the official orthography of the language, therefore such information cannot be taken into account.

Table 3.28: The subjunctive/habitual

	^{SUHA}VP			
	^{SUHA}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
subjunct. / habitual pos.				
	$2CS_{categ}$		VBP	*–e*
	example:			
	$ke_{2CSPERS\_sg}$		$reke_{V\_tr}$ $maswi_{N06}$	
	`subj-1st-sg`		buy milk	
subjunct.	‘(that) I buy / bought / will buy milk’			
habitual	‘I (usually) buy / bought / will buy milk’			
subjunct. / habitual neg.				
	$2CS_{categ}$ se_{MORPH_neg}		VBP	*–e*
	example:			
	$ke_{2CSPERS}$ se_{MORPH_neg}		$reke_{V\_itr}$ $maswi_{N06}$	
	`subj-1st-sg neg`		buy milk	
subjunct.	‘(that) I do not / did not / will not / buy milk’			
habitual	‘I (usually) do not / did not / will not / buy milk’			

3.2.11 The dependent moods: A summary

The dependent moods are summarised in Table 3.29. Note that these verbal phrases are not expected to occur without a matrix clause present in the sentence. The following section will be dedicated to the copulatives and demonstrate some auxiliary constellations.

Table 3.29: Summary of the dependent moods

		^{DEP}VP		
	^{DEP}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
cons.pos.	$3CS_{categ}$		VBP	*–e*
cons.neg.	$3CS_{categ}$ se_{MORPH_neg}		VBP	*–e*
suha.pos.	$2CS_{categ}$		VBP	*–e*
suha.neg.	$2CS_{categ}$ se_{MORPH_neg}		VBP	*–e*

3.3 The copulative verbal phrase (VP_cop)

3.3.1 Introduction

In linguistic descriptions of word classes, copulas are usually described as a subset of the sets of verbs. They are often called "linking" verbs. Taljard et al. (2008) categorise the Northern Sotho copulas as "VCOP" (compare the class "V" containing main verb stems). Lyons (1968, p. 322 et seq.) states that this word class can be clearly distinguished from main verbs from a syntactic perspective because copulatives accept – amongst other items – adjectives as their complement. We have therefore opted to categorise copula as linking a subject with a complement which is either an entity itself or describes properties of this subject[17].

Lyons (ibid.) furthermore argues that, as there are have been a number of languages (e.g. Russian, Greek or Latin) which did not use copulas earlier, this linguistic element is therefore "not itself a constituent of deep structure, but semantically empty **dummy verbs**

[17]One should however note that there are indeed main verbs in Northern Sotho that semantically contain copula and complement, like, for example, the verb *thaba* '**be** happy' in $ke_{1CSPERS\_1sg}$ ile_{V_aux} $ka_{3CSPERS\_1sg}$ $thaba_{V\_itr}$ ge_{CONJ} $wena_{PROEMPPERS\_2sg}$ o_{CSPERS_2sg} $boile_{V\_itr}$ $maabane_{N06}$ 'I **was** happy when you arrived yesterday'.

generated by the grammatical rules" of a language. Lyons (1968, p. 389 et seq.) continues in demonstrating two categories of copulas by listing the examples "apples are sweet" (copula as a characterising element) and "apples are fruit" (copula as a sorting element). While in English, both categories make use of the same copula, 'are' (here as the present tense plural form of '[to] be'), this kind of distinction is more overt in Northern Sotho, and expressed by two different copulatives, the identifying and the descriptive copulative, which make use of different copula.

While the identifying copulatives do not fit with third person subjects, the descriptives do (cf. Lombard (1985, p. 194 and 196)). Northern Sotho moreover makes use of a third category: the associative copulative, expressing the sense 'to be with', which can also mean 'to have', like in 60(a). Here, the possession is usually linked to the copula *na* (or the dynamic *ba* 'become' in the future tense) by a connective particle, *le* 'with'. The negation may occur without this particle, cf. 60(b).

60(a) *o*$_{1CS01}$ ***na***$_{VCOP}$ *le*$_{PART\_con}$ *tšhelete*$_{N09}$ *na*$_{PART\_que}$*?*
`subj-3rd-cl1` is `con` money `que`?
'do you have money (with you)?'

(b) *ga*$_{MORPH\_neg}$ *ke*$_{2CSPERS\_1sg}$ ***na***$_{VCOP}$ *tšhelete*$_{N09}$
`neg` subj-1st-sg is money
'I do not have money'

Like the main verbs of Northern Sotho, copulatives occur in a number of moods[18] and tenses, however, not all copulative categories (identifying, descriptive and associative) occur in all moods and tenses, as the paragraphs of this section will show. As far as the perfect tense is concerned, copulatives – like main verbs – usually utilise auxiliary verbs, they however select them from a closed set: *be, bego, bile, bilego*. Note however that – amongst others – *be* (as the past form of *ba*), *bile* and *bilego* also appear as copulatives themselves, indicating perfect tense.

As the following paragraphs will show, there are various constellations of the Northern Sotho copulatives. These have already been described extensively in the DLitt thesis of Elsabé Taljard (cf. Taljard (1999)), and it would exceed the scope of this study to describe them in as much detail as others have done before me.

[18]For an overview of all moods described by Lombard, cf Tables 3.2 and 3.3 on pages 76 and 77.

This section will therefore specifically focus on relevant morphosyntactic distinctions. Lombard (1985, p. 192 et seq.) and Poulos and Louwrens (1994, p. 289 et seq.) also offer an overview of the three categories identifying, descriptive and associative copula.

Poulos and Louwrens (1994, p. 291 et seq.), describe a fourth category; the "locational" copula, like in *bana*$_{\text{N02}}$ *ba*$_{\text{VCOP\_02}}$ *sekolong*$_{\text{N07\_loc}}$ '(the) children are at school'. The respective semantics in such a case, however, are expressed by the (locativised) object noun, namely *sekolong* 'at school' (cf. paragraph 2.2.2 on page 24). In conclusion, we do not find the copula *ba* contributing to the locational character of the statement as a whole. Therefore, such a category of copula is not considered in our study.

This section mainly relies on Louwrens (1991, p. 71 et seq.), Taljard (1999), and Prinsloo (2002), who all state that one should distinguish between two sub-categories "stative"[19] and "dynamic" for each of the three categories mentioned above. The static copula basically express the sense '[to] be', while the dynamic copula in general mean '[to] become'. There are therefore six basic categories of copula in total, of which each appears in certain tenses, moods, and actualities, as Table 3.30 demonstrates.
Prinsloo (2002, p. 28 et seq.)[20] summarises by stating that the following issues must be taken into account in the categorisation of copulas:

- copulas express relations between a subject and a complement, namely identification, description or association;
- there are two types of copulas for each of the defined relations: a stative and a dynamic type;
- copulas can contain the copulative particle *ke*$_{\text{PART\_cop}}$, or a subject concord referring to a person or class;
- copulas can be multiword expressions like *e le*, *e se*, *o ba*, *o na*, etc., (we should like to add: of which some contain auxiliaries, like *e*$_{\text{CSNEUT}}$ *be*$_{\text{V\_AUX}}$ *e*$_{\text{CSNEUT}}$ *le*$_{\text{VCOP}}$);

[19] Louwrens uses the term "static"

[20] Note that the progressive copula, as described by e.g. Prinsloo (2002), which makes use of the progressive morpheme *sa*$_{\text{MORPH\_prog}}$ will not be described in the scope of this study; these constellations might however be added at a later stage.

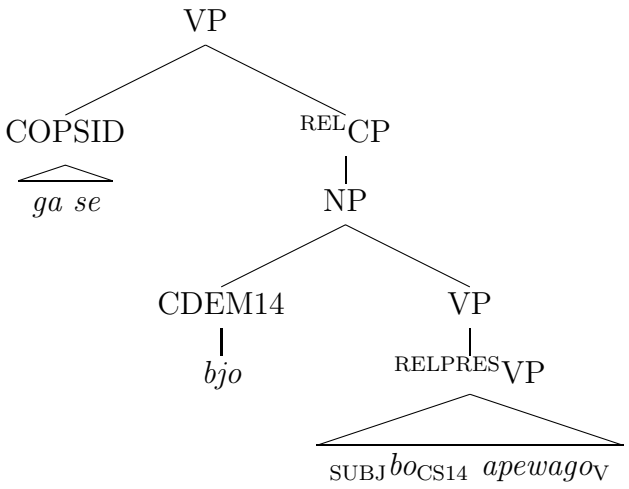
- copulas occur in moods.

There are variable and invariable copulatives, i.e. some agree with their subject and others appear to be independent of it. The class-independent (invariable) copulatives usually make use of the neutral subject concord (or the copulative particle) and, according to Louwrens (1991, p. 74 et seq.), always belong to the identifying copulatives. Nevertheless, there are variable copulatives which also appear in this category. Figures 3.22 to 3.28, are based on (Louwrens, 1991, p. 72), and list the copulative categories which are described in further detail in the accompanying paragraphs. They also contain some word-for-word translations to illustrate the differences in meaning as well as information on the expected complement for each of the copulative constellations (NOM for nominal, ADJ for adjectival). The following paragraphs will describe each of the categories in more detail. Note that the object of the copulative cannot be represented by an object concord, therefore the definition of an intermediate structure describing a 'basic copulative phrase' is deemed unnecessary.

Table 3.30: Overview of copulative constellations

Copulative	Identifying		Descriptive		Associative	
Category	stative	dynamic	stative	dynamic	stative	dynamic
Tense						
pres	×	×	×	×	×	×
perfect	×	×	×	×	×	×
fut		×		×		×
Mood						
indicative (pos/neg)	×	×	×	×	×	×
situative (pos/neg)	×	×	×	×	×	×
relative (pos/neg)	×	×	×	×	×	×
consecutive (pos/neg)		×		×		×
subjunctive (pos/neg)		×		×		×
habitual (pos/neg)		×		×		×
infinitive (pos/neg)		×		×		×
imperative (pos/neg)		×		×		×

Figure (63)



3.3.2 The identifying copulative

Lombard (1985, p. 194) states that this copulative is used when "a person or thing is identified with another one", like in the example (61). As only nouns express entities like person or things, the identifying copulas should require subject and object to be nominal, cf. example (62), taken from Poulos and Louwrens (1994, p. 307).

(61) *ke*$_{\text{VCOP\_1sg}}$ *moithuti*$_{\text{N01}}$
I am student
'I am (a) student'

(62) *mošemane*$_{\text{N01}}$ *ke*$_{\text{PART\_cop}}$ *yo*$_{\text{CDEM01}}$ *mošweu*$_{\text{ADJ}}$
boy is `dem-3rd-cl1` white
'(a) boy is white'

Example (and figure) (63) of Lombard (1985, p. 194) show that these copulatives also occur with clausal complements.

(63) *ga*$_{\text{MORPH\_neg}}$ *se*$_{\text{VCOP\_neg}}$ *bjo*$_{\text{CDEM14}}$ *bo*$_{\text{1CS14}}$ *apewago*$_{\text{V}}$
`neg` is not `dem-3rd-cl14` `subj-3rd-cl14` which-is cooked
'it is not that which is cooked'

The copulas that appear in these constellations are either identical to the personal subject concords (as described in paragraph 3.3.1, they will be categorised as copulas in this case)

or the impersonal copulative particle, $ke_{\text{PART\_cop}}$ which is used for all subjects belonging to a noun class. The negation makes use of the negative morpheme *ga*.

Just like the main verbs, the future tense ('will be') is formed by adding a preceding *tla* or *tlo* to the present tense. However, there are only dynamic forms of the future tense (e.g. *ba* 'become' will form *tlo ba* 'will become/be'). The negation of this tense, again similar to main verbs, makes use of the word group *ka se*. The stative is described for the indicative, situative (principal and participial by Poulos and Louwrens (1994)), and for the relative mood (e.g. Prinsloo (2002)). Furthermore, the dynamic describes dependent moods like e.g. the consecutive and the non-predicative categories, which contain infinitive and imperative.

To simplify the morphosyntactic rules defined in this chapter, a number of their units are described as elements of certain sets, cf. Table 3.31. Whenever a set name appears in a rule, it stands for one of its elements. If the set name occurs again in the same rule, it represents a repetition of the same element. As in the previous chapter, the abbreviation "categ" stands for an extension, however, the rule will store this information (e.g. '1st-sg' as an extension for persons or '3rd-cl2' as an extension for class 2) if its elements need to agree with an external subject. This issue is demonstrated by example (64), which shows an identifying stative copulative constellation in the perfect tense. The second line of this example shows the rule to be applied. In this example, the rule contains `1CSPCSN` (this set contains all `1CSPERS`, i.e. all subject concords of the first set referring to the 1st or 2nd person, and the neutral subject concord e_{CSNEUT}, cf. Table 3.31) . As `1CSPCSN` appears twice it means that identical subject concords should appear.

(64) $ke_{\text{1CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{1CSPERS\_1sg}}$ le_{VCOP} *(morutiši)*
`1CSPCSN` $be_{\text{V\_aux}}$ `1CSPCSN` le_{VCOP}
'I was (a teacher)'

An overview of the identifying copulatives is shown in the seven figures 3.22 to 3.28. For each of these figures[21]; a table is provided to demonstrate the rules for the respective constellations, illustrated by examples.

[21]In order to save space, only the relative pronoun 'that' appears in the constellations described by figures 3.22 to 3.28. It stands however for 'that', 'who' or 'which'. The same applies to 'is' representing 'is', 'am', or 'are', and 'was', representing 'was' or 'were'.

Table 3.31: groups of copulas

Set name	Set description	Elements of this set
$\text{CSPCSN}_{\text{categ}}$	CSPersonal and CSNeutral	$\boldsymbol{ke}_{\text{1CSPERS\_1sg}}$, $\boldsymbol{o}_{\text{1CSPERS\_2sg}}$ $\boldsymbol{re}_{\text{CSPERS\_1pl}}$,$\boldsymbol{le}_{\text{CSPERS\_2pl}}$, $\boldsymbol{e}_{\text{CSNEUT}}$
$\text{VCPC}_{\text{categ}}$	VCOPPersonal and cop. particle	$\boldsymbol{ke}_{\text{VCOP\_1sg}}$, $\boldsymbol{o}_{\text{VCOP\_2sg}}$ $\boldsymbol{re}_{\text{VCOP\_1pl}}$, $\boldsymbol{le}_{\text{VCOP\_2pl}}$, $\boldsymbol{ke}_{\text{PART\_cop}}$
$\text{VCNEG}_{\text{categ}}$	VCOPPersonal and negative VCOP	$\boldsymbol{ke}_{\text{VCOP\_1sg}}$, $\boldsymbol{o}_{\text{VCOP\_2sg}}$ $\boldsymbol{re}_{\text{VCOP\_1pl}}$, $\boldsymbol{le}_{\text{VCOP\_2pl}}$, $\boldsymbol{se}_{\text{VCOP\_neg}}$
$\text{1VCOP}_{\text{categ}}$	VCOPPersonal and classes	$\text{VCOP}_{\text{categ}}$ ($\boldsymbol{o}$ for class 1)
$\text{2VCOP}_{\text{categ}}$	VCOPPersonal and classes	$\text{VCOP}_{\text{categ}}$ ($\boldsymbol{a}$ for class 1)

3.3.2.1 The stative

Tables 3.32 and 3.33 show an overview of the morphological rules forming the stative constellations (demonstrated in Figures 3.22 and 3.23, and contain examples.

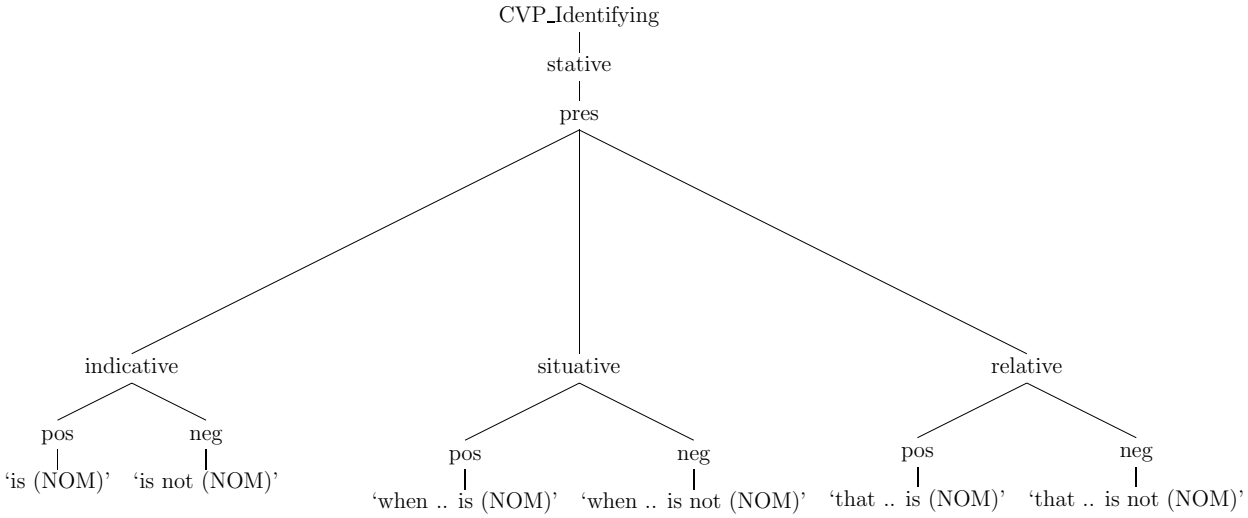


Figure 3.22: Identifying stative present tense (cf. Table 3.32)

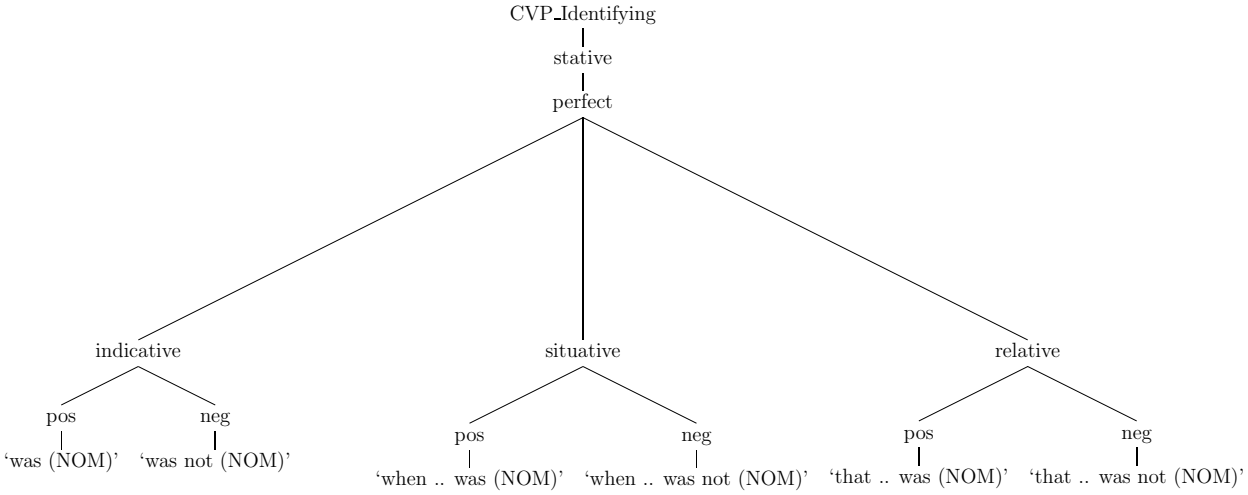


Figure 3.23: Identifying stative perfect tense (cf. Table 3.33)

Table 3.32: The **stative** forms of the **identifying** copulative (COPSID) (**present tense**, figure (3.22))

$^{\text{COPSID}}\text{V}$		
tense	**mood and actuality**	**elements**
pres.	**ind pos**	**VCPC**$_{\text{categ}}$
	examples:	
	(monna)	*ke*$_{\text{PART\_cop}}$ *(morutiši*$_{\text{N01}}$*)* '(the man) is (a teacher)' *ke*$_{\text{VCOP\_1sg}}$ *(morutiši*$_{\text{N01}}$*)* 'I am (a teacher)'
	ind neg	***ga***$_{\text{MORPH\_neg}}$ **VCNEG**$_{\text{categ}}$
	(monna)	*ga*$_{\text{MORPH\_neg}}$ *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)* '(the man) is not (a teacher)' *ga*$_{\text{MORPH\_neg}}$ *ke*$_{\text{VCOP\_1sg}}$ *(morutiši*$_{\text{N01}}$*)* 'I am not (a teacher)'
	sit pos	**CSPCSN**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$
	(ge monna)	*e*$_{\text{CSNEUT}}$ *le*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)* '(as the man) is (a teacher)'
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *le*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)* '(as) I am (a teacher)'
	sit neg	**CSPCSN**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$
	(ge monna)	*e*$_{\text{CSNEUT}}$ *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)* '(as the man) is not (a teacher)'
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)* '(as) I am not (a teacher)'
	rel pos	**CSPCSN**$_{\text{categ}}$ ***lego***$_{\text{VCOP}}$
	(monna yo)	*e*$_{\text{CSNEUT}}$ *lego*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)* '(the man who) is (a teacher)'
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *lego*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)* '(I who) am (a teacher)'
	rel neg	**CSPCSN**$_{\text{categ}}$ ***sego***$_{\text{VCOP\_neg}}$
	(monna yo)	*e*$_{\text{CSNEUT}}$ *sego*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)* '(the man who) is not (a teacher)'
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *sego*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)* '(I who) am not (a teacher)'

Table 3.33: The **stative** forms of the **identifying** copulative (COPSID) (**perfect tense**, figure (3.23))

		$^{\text{COPSID}}$V	
tense	**mood and actuality**	**tense marker**	**elements**
perfect	**ind/sit pos**	**CSPCSN**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$	**CSPCSN**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$
	examples:		
	((ge) monna)	e_{CSNEUT} *be*$_{\text{V\_aux}}$	e_{CSNEUT} *le*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'((as) the man) was (a teacher)'	
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *be*$_{\text{V\_aux}}$	*ke*$_{\text{1CSPERS\_1sg}}$ *le*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(as) I was (a teacher)'	
	ind/sit neg	**CSPCSN**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$	**CSPCSN**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$
	((ge) monna)	e_{CSNEUT} *be*$_{\text{V\_aux}}$	e_{CSNEUT} *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)*
		'((as) the man) was not (a teacher)'	
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *be*$_{\text{V\_aux}}$	*ke*$_{\text{1CSPERS\_1sg}}$ *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)*
		'(as) I was not (a teacher)'	
	rel pos	**CSPCSN**$_{\text{categ}}$ ***bego***$_{\text{V\_aux}}$	**CSPCSN**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$
	(monna yo)	e_{CSNEUT} *bego*$_{\text{V\_aux}}$	e_{CSNEUT} *le*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) was (a teacher)'	
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *bego*$_{\text{V\_aux}}$	*ke*$_{\text{1CSPERS\_1sg}}$ *le*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(I who) was (a teacher)'	
	rel neg	**CSPCSN**$_{\text{categ}}$ ***bego***$_{\text{V\_aux}}$	**CSPCSN**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$
	(monna yo)	e_{CSNEUT} *bego*$_{\text{V\_aux}}$	e_{CSNEUT} *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) was not (a teacher)'	
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *bego*$_{\text{V\_aux}}$	*ke*$_{\text{1CSPERS\_1sg}}$ *se*$_{\text{VCOP\_neg}}$ *(morutiši*$_{\text{N01}}$*)*
		'(I who) was not (a teacher)'	

3.3.2.2 The dynamic

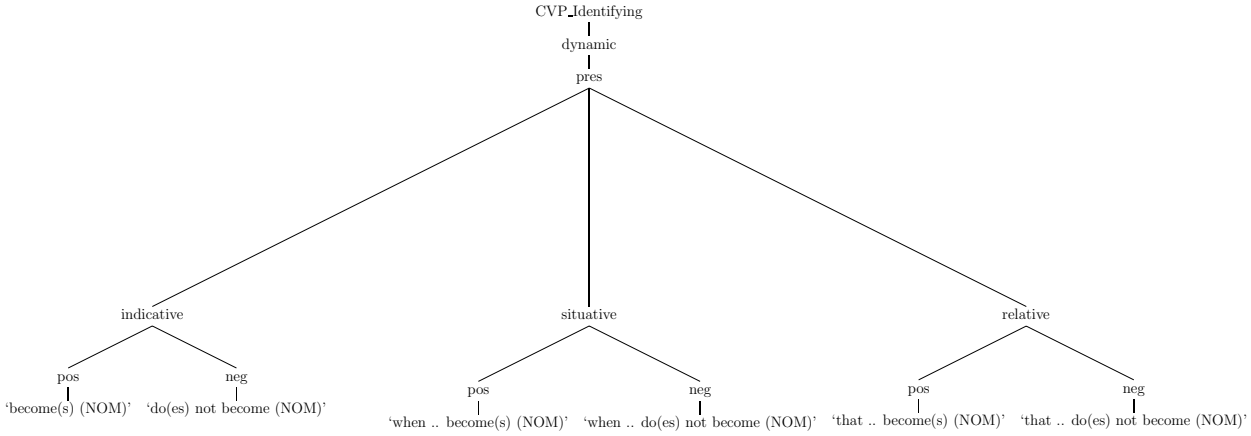


Figure 3.24: Identifying dynamic present tense (cf. Table 3.34)

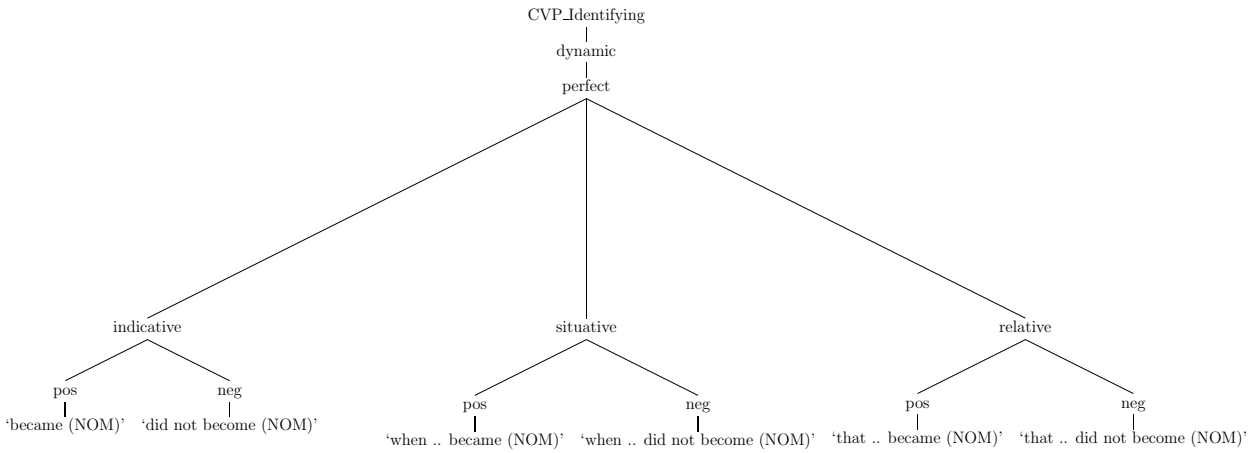


Figure 3.25: Identifying dynamic perfect tense (cf. Table 3.35)

Tables 3.34 to 3.39 contain the morphological rules forming the dynamic constellations (demonstrated in Figures 3.24 to 3.28. Note that in the perfect tense of the identifying dynamic copulative (Table 3.35), the tense itself is overt only in the positive form of the indicative/situative mood (the perfect tense form *bile* of *ba* 'become' is used). The negative forms show no tense marker and make use of the present tense copula *ba* and *be* 'become'. Therefore, the fact that these constellations represent the perfect tense cannot be deduced from a single component, rather from the constellation as a whole.

The dependent forms are translated into present tense English in tables 3.37 and 3.38 though consecutive, subjunctive and habitual could appear as other tenses as well. The tense of these constellations is determined by the tense of the main clause they depend upon.

Table 3.34: The **dynamic** forms of the **identifying** copulative (COPDID) (**present tense**, Figure 3.24)

		$^{\text{COPDID}}$V
tense	**mood and actuality**	**elements**
pres.	**ind/sit pos**	**CSPCSN**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$
	examples:	
	((ge) monna)	e_{CSNEUT} *ba*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'((as) the man) becomes (a teacher)'
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *ba*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(as) I become (a teacher)'
	ind neg	***ga***$_{\text{MORPH\_neg}}$ **CSPCSN**$_{\text{categ}}$ ***be***$_{\text{VCOP}}$
	(monna)	*ga*$_{\text{MORPH\_neg}}$ e_{CSNEUT} *be*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man) does not become (a teacher)'
		ga$_{\text{MORPH\_neg}}$ *ke*$_{\text{1CSPERS\_1sg}}$ *be*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'I do not become (a teacher)'
	sit neg	**CSPCSN**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***be***$_{\text{VCOP}}$
	(ge monna)	e_{CSNEUT} *sa*$_{\text{MORPH\_neg}}$ *be*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(as the man) does not become (a teacher)'
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *sa*$_{\text{MORPH\_neg}}$ *be*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(as) I do not become (a teacher)'
	rel pos	**CSPCSN**$_{\text{categ}}$ ***bago***$_{\text{VCOP}}$
	(monna yo)	e_{CSNEUT} *bago*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) becomes (a teacher)'
	(nna yo)	*ke*$_{\text{1CSPERS\_sg}}$ *bago*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(I who) become (a teacher)'
	rel neg	**CSPCSN**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***bego***$_{\text{VCOP}}$
	(monna yo)	e_{CSNEUT} *sa*$_{\text{MORPH\_neg}}$ *bego*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) does not become (a teacher)'
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *sa*$_{\text{MORPH\_neg}}$ *bego*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(I who) do not become (a teacher)'

Table 3.35: The **dynamic** forms of the **identifying** copulative (COPDID) (**perfect tense**, Figure 3.25)

		$^{\text{COPDID}}$V
tense	**mood and actuality**	**elements**
perfect	**ind/sit pos**	**CSPCSN**$_{\text{categ}}$ ***bile***$_{\text{VCOP}}$
	examples:	
	((ge) monna)	e_{CSNEUT} *bile*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'((when) the man) became (a teacher)'
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *bile*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(when) I became (a teacher)'
	ind neg	***ga***$_{\text{MORPH\_neg}}$ ***se***$_{\text{MORPH\_neg}}$ **3CS**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$
	(monna)	*ga*$_{\text{MORPH\_neg}}$ *se*$_{\text{MORPH\_neg}}$ *ya*$_{\text{3CS01}}$ *ba*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man) did not become (a teacher)'
		ga$_{\text{MORPH\_neg}}$ *se*$_{\text{MORPH\_neg}}$ *ka*$_{\text{3CSPERS\_1sg}}$ *ba*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'I did not become (a teacher)'
	sit neg	**CSPCSN**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***ba***$_{\text{VCOP}}$
	(ge monna)	e_{CSNEUT} *sa*$_{\text{MORPH\_neg}}$ *ba*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'((when) the man) did not become (a teacher)'
	(ge)	*ke*$_{\text{1CSPERS\_1sg}}$ *sa*$_{\text{MORPH\_neg}}$ *ba*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(when) I did not become (a teacher)'
	rel pos	**CSPCSN**$_{\text{categ}}$ ***bilego***$_{\text{VCOP}}$
	(monna yo)	e_{CSNEUT} *bilego*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) became (a teacher)'
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *bilego*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(when) I became (a teacher)'
	rel neg	**CSPCSN**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***bago***$_{\text{VCOP}}$
	monna yo	e_{CSNEUT} *sa*$_{\text{MORPH\_neg}}$ *bago*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) did not become (a teacher)'
	(nna yo)	*ke*$_{\text{1CSPERS\_1sg}}$ *sa*$_{\text{MORPH\_neg}}$ *bago*$_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(I who) did not become (a teacher)'

Table 3.36: The **dynamic** forms of the **identifying** copulative (COPDID) (**future tense**, Figure 3.26)

		$^{\text{COPDID}}\text{V}$
tense	**mood and actuality**	**tense marker and other elements**
fut.	**ind/sit pos**	**CSPCSN**$_{\text{categ}}$ ***tlo/tla***$_{\text{MORPH\_fut}}$ ***ba***$_{\text{VCOP}}$
	examples:	
	((ge) monna)	e_{CSNEUT} *tlo* ba_{VCOP} *(morutiši*$_{\text{N01}}$*)*
		'((when) the man) will become (a teacher)'
	(ge)	$ke_{\text{1CSPERS\_1sg}}$ *tlo* ba_{VCOP} *(morutiši*$_{\text{N01}}$*)*
		'(when) I will become (a teacher)'
	ind/sit neg	**CSPCSN** ***ka***$_{\text{MORPH\_pot}}$ ***se***$_{\text{MORPH\_neg}}$***be***$_{\text{VCOP}}$
	((ge) monna)	e_{CSNEUT} $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)*
		'((as) the man) will not become (a teacher)'
	(ge)	*nka* ($ke_{\text{1CSPERS\_1sg}}$ + $ka_{\text{MORPH\_pot}}$) $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)*
		'(as) I will not become (a teacher)'
	rel pos 1	**CSPCSN**$_{\text{categ}}$ ***tlo/tla***$_{\text{MORPH\_fut}}$ ***bago***$_{\text{VCOP}}$
	rel pos 2	**CSPCSN**$_{\text{categ}}$ ***tlogo/tlago***$_{\text{MORPH\_fut}}$ ***ba***$_{\text{VCOP}}$
	(monna yo)	e_{CSNEUT} *tlo* $bago_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		e_{CSNEUT} *tlogo* ba_{VCOP}
		'(the man who) will become (a teacher)'
	(nna yo)	$ke_{\text{1CSPERS\_sg}}$ *tlo* $bago_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		$ke_{\text{1CSPERS\_sg}}$ *tlago* ba_{VCOP}
		'(I who) become (a teacher)'
	rel neg	**CSPCSN**$_{\text{categ}}$ ***ka***$_{\text{MORPH\_pot}}$ ***se***$_{\text{MORPH\_neg}}$ ***bego***$_{\text{VCOP}}$
	(monna yo)	e_{CSNEUT} $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$ $bego_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(the man who) will not become (a teacher)'
	(nna yo)	*nka* ($ke_{\text{1CSPERS\_1sg}}$ + $ka_{\text{MORPH\_pot}}$) $se_{\text{MORPH\_neg}}$ $bego_{\text{VCOP}}$ *(morutiši*$_{\text{N01}}$*)*
		'(I who) will not become (a teacher)'

Table 3.37: The **dynamic** forms of the **identifying** copulative (COPDIDD) (**dependent constellations** part 1 of 2, Figure 3.27)

$^{\text{COPDIDD}}$V	
mood and actuality	**elements**
consecutive pos	**3CS**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$
	examples: *(monna)* ya_{3CS01} ba_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(then the man) becomes (a teacher)' $ka_{\text{3CSPERS\_1sg}}$ ba_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(then I) become (a teacher)'
consecutive neg	**3CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$ ***be***$_{\text{VCOP}}$
	(monna) ya_{3CS01} $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(then the man) does not become (a teacher)' $ka_{\text{3CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(then I) do not become (a teacher)'
subjunctive pos	**CSPCSN**$_{\text{categ}}$ ***be***$_{\text{VCOP}}$
	(gore monna) e_{CSNEUT} be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(so that the man) becomes (a teacher)' *(gore)* $ke_{\text{1CSPERS\_1sg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(so that) I become (a teacher)'
subjunctive neg	**CSPCSN**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$ ***be***$_{\text{VCOP}}$
	(gore monna) e_{CSNEUT} $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(so that the man) does not become (a teacher)' *(gore)* $ke_{\text{1CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(so that) I do not become (a teacher)'

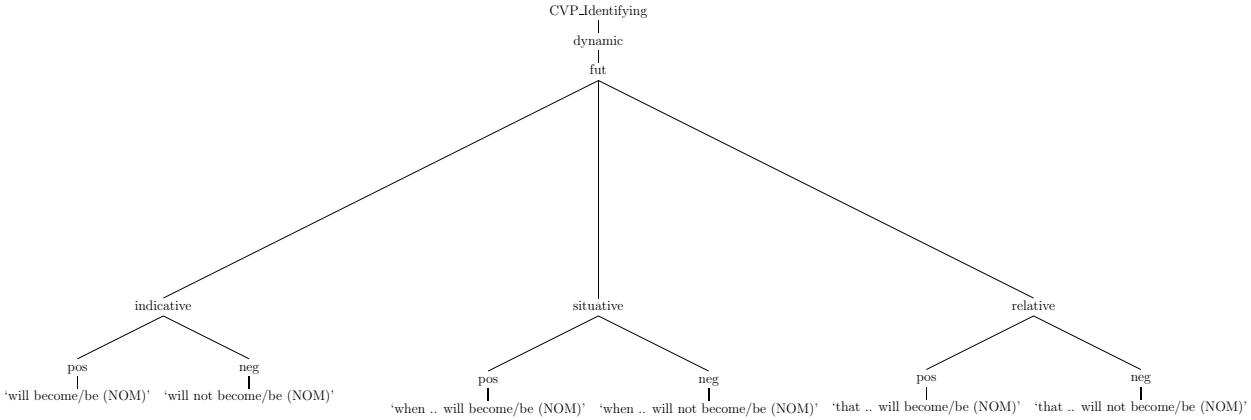


Figure 3.26: Identifying dynamic future tense (cf. Table 3.36)

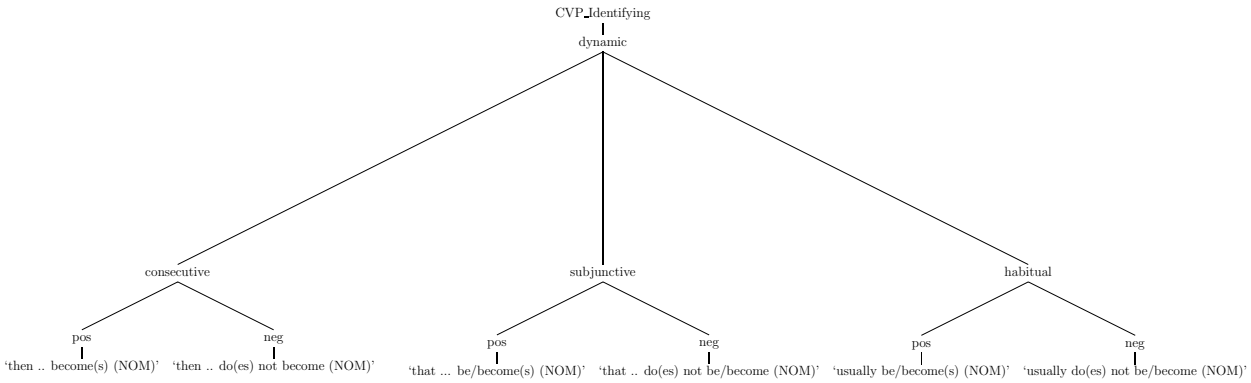


Figure 3.27: Identifying dynamic dependent clauses (cf. Tables 3.37 and 3.38)

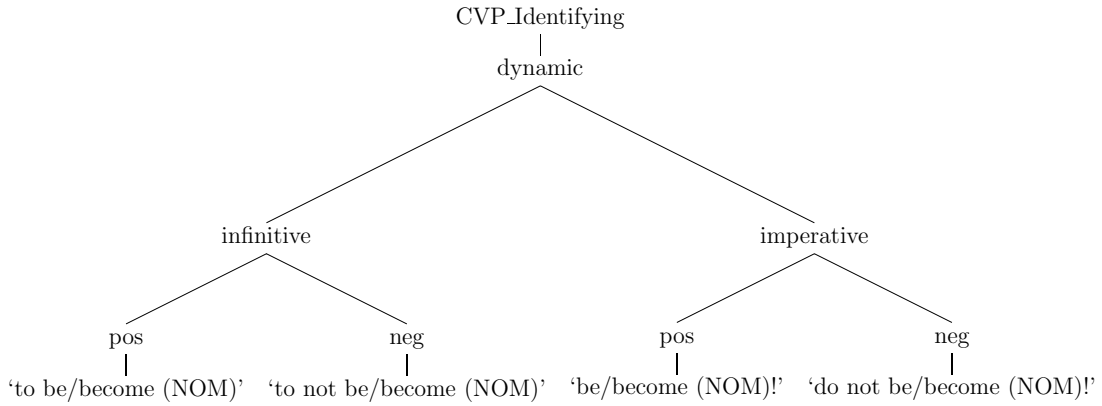


Figure 3.28: Identifying dynamic non-predicative clauses (cf. Table 3.39)

Table 3.38: The **dynamic** forms of the **identifying** copulative (COPDIDD) (**dependent constellations** part 2 of 2, Figure 3.27)

$^{\text{COPDIDD}}\text{V}$	
mood and actuality	**elements**
habitual pos	**CSPCSN**$_{\text{categ}}$ ***be***$_{\text{VCOP}}$
	examples: *(monna)* e_{CSNEUT} be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(a man) (usually) becomes (a teacher)' $ke_{\text{1CSPERS\_1sg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* 'I (usually) become (a teacher)'
habitual neg	**CSPCSN**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$ ***be***$_{\text{VCOP}}$
	(monna) e_{CSNEUT} $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* '(a man) (usually) does not become (a teacher)' $ke_{\text{1CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* 'I (usually) do not become (a teacher)'

Table 3.39: The **dynamic** forms of the **identifying** copulative (COPDI) (**non-predicative** constellations, Figure 3.28)

$^{\text{COPDI}}\text{V}$	
mood and actuality	**elements**
infinitive pos	***go***$_{\text{MORPH\_cp15}}$ ***ba***$_{\text{VCOP}}$
	example: $go_{\text{MORPH\_cp15}}$ ba_{VCOP} *(morutiši*$_{\text{N01}}$*)* 'to become a teacher'
infinitive neg	***go***$_{\text{MORPH\_cp15}}$ ***se***$_{\text{MORPH\_neg}}$ ***be***$_{\text{VCOP}}$
	$go_{\text{MORPH\_cp15}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši*$_{\text{N01}}$*)* 'not to / to not become a teacher'
imperative pos	***eba***$_{\text{VCOP}}$
	eba_{VCOP} *(morutiši*$_{\text{N01}}$*!)* 'become (a teacher!)'
imperative neg	**se**$_{\text{MORPH\_neg}}$ **be**$_{\text{VCOP}}$
	$se_{\text{MORPH\_neg}}$ be_{VCOP} *(morutiši!)* 'do not become (a teacher!)'

3.3.3 The descriptive copulative

The descriptive copulative differs from the identifying semantically in that it is used when describing properties of an entity (its subject). An important morphosyntactic property of these copulatives is that a subject concord appearing as a copula has to agree with its subject's category (in terms of its noun class).

It has not been necessary for the identifying copulative (cf. paragraph 3.3.2) to explicitly define its complement, as both predicative and non-predicative complements are possible. For the descriptive copulative, its complement could be expected to be solely adjectival, like, e.g. in the English clause 'he$_{\text{PRO}}$ is$_{\text{VCOP}}$ furious$_{\text{ADJ}}$'. However, while Lombard (1985, p. 196) generally states that the complement of the descriptive copulative is "non-verbal" or "non-predicative", Poulos and Louwrens (1994, p. 307) explicitly define that no adjectival complement appears in this copulative. According to them, adjectives instead appear in the identifying copulative, like in the example (62) from page 129, repeated here as (65) for the sake of convenience.

(65)

mošemane$_{\text{N01}}$	*ke*$_{\text{PART\_cop}}$	*yo*$_{\text{CDEM01}}$	*mošweu*$_{\text{ADJ}}$
boy	is	`dem-3rd-cl1`	white

'(a) boy is white'

Consequently, following Poulos and Louwrens would result in a definition of the descriptive copulative as not permitting certain constellations. On the other hand, grammar rules usually are not defined negatively, i.e. by excluding certain types of constituents and allowing all others as these authors suggest. Therefore, in order to be able to define the possible complements of the descriptive copulative, examples mentioned in the referenced literature were examined. Additionally, some cleaned parts of the Northern Sotho text collection, the *University of Pretoria Sepedi Corpus* (PSC), (cf. De Schryver and Prinsloo (2000)) were searched for subject concords appearing as copulas.

These examinations revealed an interesting result: most occurrences of such copulas mentioned in the referenced literature to illustrate this copulative appeared with nouns of class 14 as their complement. It is a general property of Northern Sotho, that some parts of speech may (additionally) express semantic concepts not typically associated with them and this is especially true for nouns of class 14, like, e.g. *bohlale*$_{\text{N14}}$ 'skill, intelligence, wis-

dom' which also means 'clever'[22], *botse*$_{\text{N14}}$ 'beauty', but also 'beautiful', *boima*$_{\text{N14}}$ 'weight', but also 'heavy', etc. The other sources that were examined exposed complements like locative nouns with an adverbial content, e.g. *kgauswi* 'near' in *marega*$_{\text{N06}}$ *a*$_{\text{VCOP\_06}}$ *kgauswi*$_{\text{NLOC}}$ 'the winter is near', while the word class 'adverb' (ADV) was not found at all complementing this copula. However, as the research on this issue was not extensive, the morphosyntactic rules for this copulative defined in this study should therefore be taken as preliminary. We define the rules in a way so that they permit all nominals (in a narrower sense of nouns and pronouns) to be present.

The only other detectable morphosyntactic difference to the identifying copulative (which is, at the same time, a similarity with main verbs) is that subject concords used as copulatives have to agree with their subject's category in all cases, i.e. the invariable copulative particle does not appear. This entails the definition that the copulas referring to 1$^{\text{st}}$ and 2$^{\text{nd}}$ persons (*ke*$_{\text{VCOP\_1sg}}$, *o*$_{\text{VCOP\_2sg}}$, *re*$_{\text{VCOP\_1pl}}$, and *le*$_{\text{VCOP\_2pl}}$) appear in both. Consequently, from an analysis perspective, a parser cannot distinguish between an identifying and descriptive copulative if the subject of the copulative is a 1$^{\text{st}}$ or 2$^{\text{nd}}$ person and the complement is a nominal.

In summary, these copulatives make use of the sets 1VCOP$_{\text{categ}}$ and 2VCOP$_{\text{categ}}$ and are – as a preliminary rule – complemented by nominals only. The neutral subject concord may only appear according to its original definition, i.e. if no anaphoric reference to a subject is available (cf. paragraph 2.4.2). Just like the identifying copulatives, the descriptive is illustrated by figures and tables showing all stative (Figures 3.29 and 3.30 and Tables 3.40 and 3.41) and dynamic forms (Figures 3.31 to 3.35 and Tables 3.42 to 3.45).

[22]In the dictionary De Schryver (2007), these secondary meanings appear as "nominal relative", or as adjectival (being preceded by a demonstrative), their appearance as complements of copulas is not mentioned.

3.3.3.1 The stative

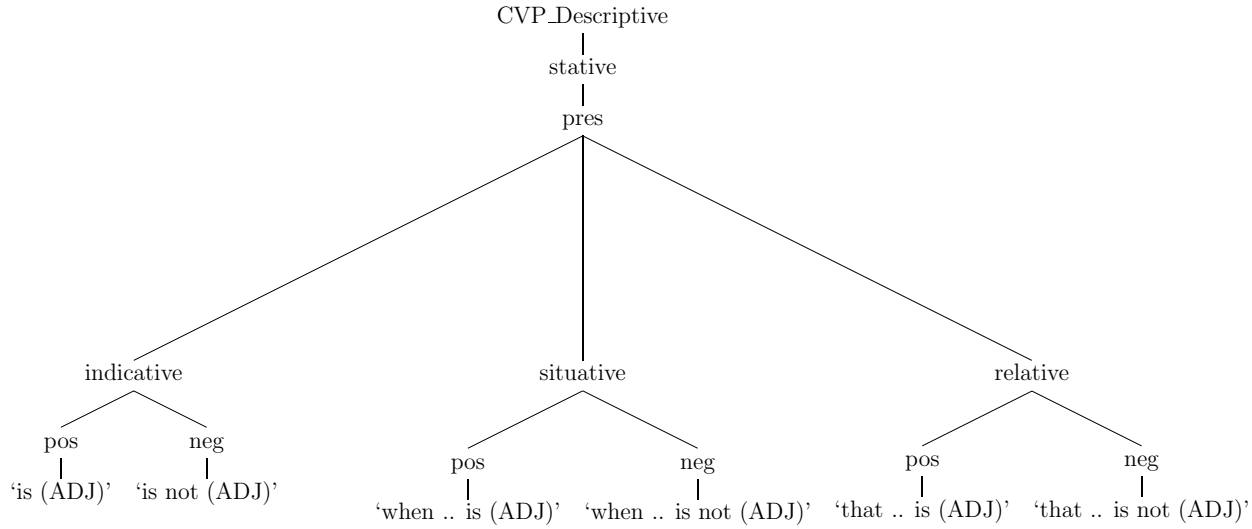


Figure 3.29: Descriptive stative present tense (cf. Table 3.40)

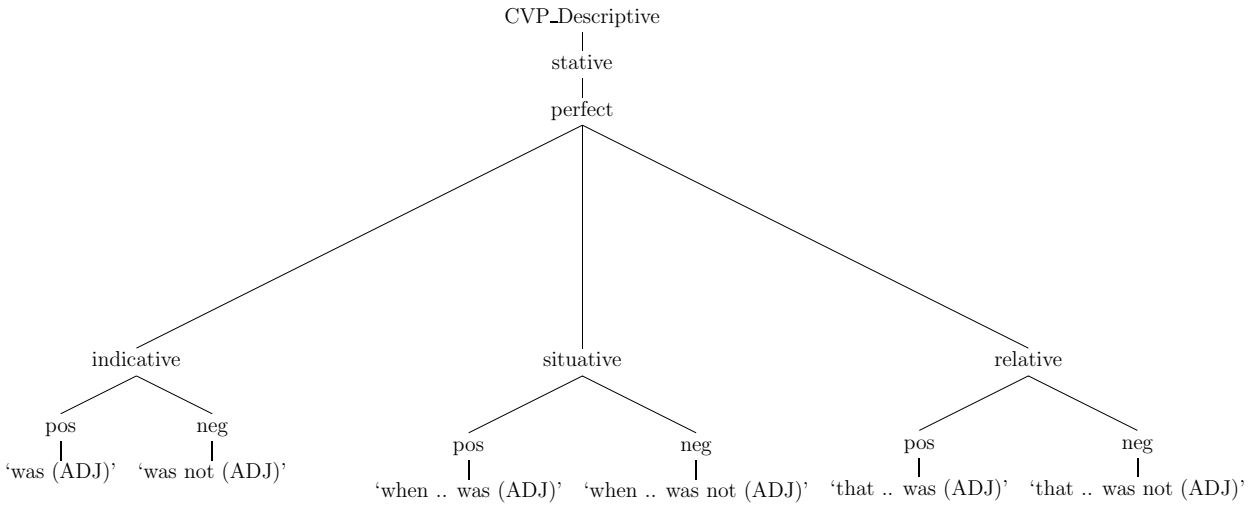


Figure 3.30: Descriptive stative perfect tense (cf. Table 3.41)

Table 3.40: The **stative** forms of the **descriptive** copulative (COPSDC) (**present tense**, Figure 3.29)

		$^{\text{COPSDC}}\text{V}$	
tense	**mood and actuality**	**elements**	**complement**
pres.	**ind pos**	**1VCOP**$_{\text{categ}}$	nominal
	examples:		
	(monna)	$o_{\text{VCOP\_01}}$ '(the man) is clever'	$bohlale_{\text{N14}}$
		$ke_{\text{VCOP\_1sg}}$ 'I am clever'	$bohlale_{\text{N14}}$
	ind neg	***ga***$_{\text{MORPH\_neg}}$ **2VCOP**$_{\text{categ}}$	nominal
	(monna)	$ga_{\text{MORPH\_neg}}$ $a_{\text{VCOP\_01}}$ '(the man) is not clever'	$bohlale_{\text{N14}}$
		$ga_{\text{MORPH\_neg}}$ $ke_{\text{VCOP\_1sg}}$ 'I am not clever'	$bohlale_{\text{N14}}$
	sit pos	**2CS**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$	nominal
	(ge monna)	a_{2CS01} le_{VCOP} '(as the man) is clever'	$bohlale_{\text{N14}}$
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ le_{VCOP} '(as) I am clever'	$bohlale_{\text{N14}}$
	sit neg	**CSPCSN**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$	nominal
	(ge monna)	a_{2CS01} $se_{\text{VCOP\_neg}}$ '(as the man) is not clever'	$bohlale_{\text{N14}}$
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $se_{\text{VCOP\_neg}}$ '(as) I am not clever'	$bohlale_{\text{N14}}$
	rel pos	**2CS**$_{\text{categ}}$ ***lego***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} $lego_{\text{VCOP}}$ '(the man who) is clever'	$bohlale_{\text{N14}}$
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $lego_{\text{VCOP}}$ '(I who) am clever'	$bohlale_{\text{N14}}$
	rel neg	**2CS**$_{\text{categ}}$ ***sego***$_{\text{VCOP\_neg}}$	nominal
	(monna yo)	a_{2CS01} $sego_{\text{VCOP\_neg}}$ '(the man who) is not clever'	$bohlale_{\text{N14}}$
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $sego_{\text{VCOP\_neg}}$ '(I who) am not clever'	$bohlale_{\text{N14}}$

Table 3.41: The **stative** forms of the **descriptive** copulative (COPSDC) (**perfect tense**, Figure 3.30)

		$^{\text{COPSDC}}\text{V}$	
tense	**mood and actuality**	**elements**	**complement**
perfect	**ind pos**	**1CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$	nominal
		examples:	
	(monna)	o_{1CS01} $be_{\text{V\_aux}}$ a_{2CS01} le_{VCOP}	$bohlale_{\text{N14}}$
		'(the man) was clever'	
		$ke_{\text{1CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ le_{VCOP}	$bohlale_{\text{N14}}$
		'I was clever'	
	ind neg	**1CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$	nominal
	(monna)	o_{1CS01} $be_{\text{V\_aux}}$ a_{2CS01} $se_{\text{VCOP\_neg}}$	$bohlale_{\text{N14}}$
		'(the man) was not clever'	
		$ke_{\text{1CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ se_{VCOP}	$bohlale_{\text{N14}}$
		'I was not clever'	
	sit pos	**2CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$	nominal
	(ge monna)	a_{2CS01} $be_{\text{V\_aux}}$ a_{2CS01} le_{VCOP}	$bohlale_{\text{N14}}$
		'(when the man) was clever'	
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ le_{VCOP}	$bohlale_{\text{N14}}$
		'(when) I was clever'	
	sit neg	**2CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$	nominal
	(ge monna)	a_{2CS01} $be_{\text{V\_aux}}$ a_{2CS01} $se_{\text{VCOP\_neg}}$	$bohlale_{\text{N14}}$
		'(when) (the man) was not clever	
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ se_{VCOP}	$bohlale_{\text{N14}}$
		'(when) I was not clever	
	rel pos	**2CS**$_{\text{categ}}$ ***bego***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***le***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} $bego_{\text{V\_aux}}$ a_{2CS01} le_{VCOP}	$bohlale_{\text{N14}}$
		'(the man who) was clever'	
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $bego_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ le_{VCOP}	$bohlale_{\text{N14}}$
		'(I who) was clever'	
	rel neg	**2CS**$_{\text{categ}}$ ***bego***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***se***$_{\text{VCOP\_neg}}$	nominal
	(monna yo)	a_{2CS01} $bego_{\text{V\_aux}}$ a_{2CS01} $se_{\text{VCOP\_neg}}$	$bohlale_{\text{N14}}$
		'(the man who) was not clever'	
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $bego_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ $se_{\text{VCOP\_neg}}$	$bohlale_{\text{N14}}$
		'(I who) was not clever'	

3.3.3.2 The dynamic

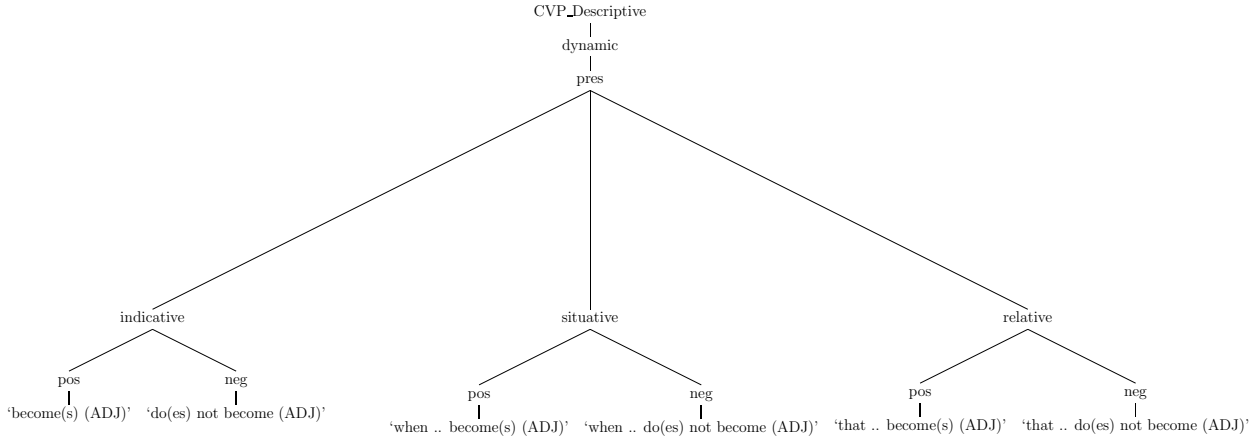


Figure 3.31: Descriptive dynamic present tense (cf. Table 3.42)

Note that the dynamic non-predicative copulatives of the descriptive are identical to the identifying forms, in spite of their complement being a nominal in all cases. As a result, Figure 3.35 is basically mirrored in Table 3.39 on page 141.

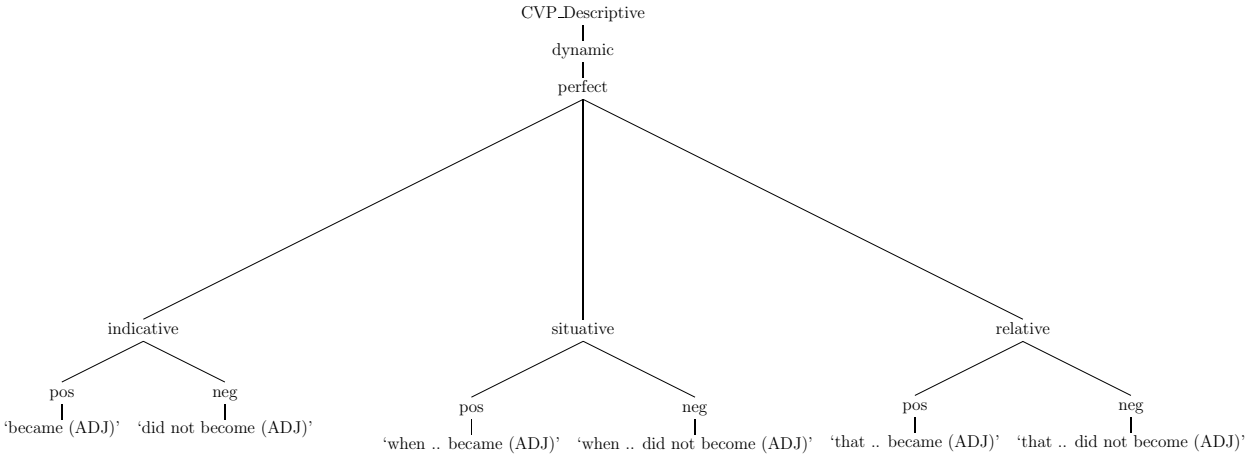


Figure 3.32: Descriptive dynamic perfect tense (cf. Table 3.43)

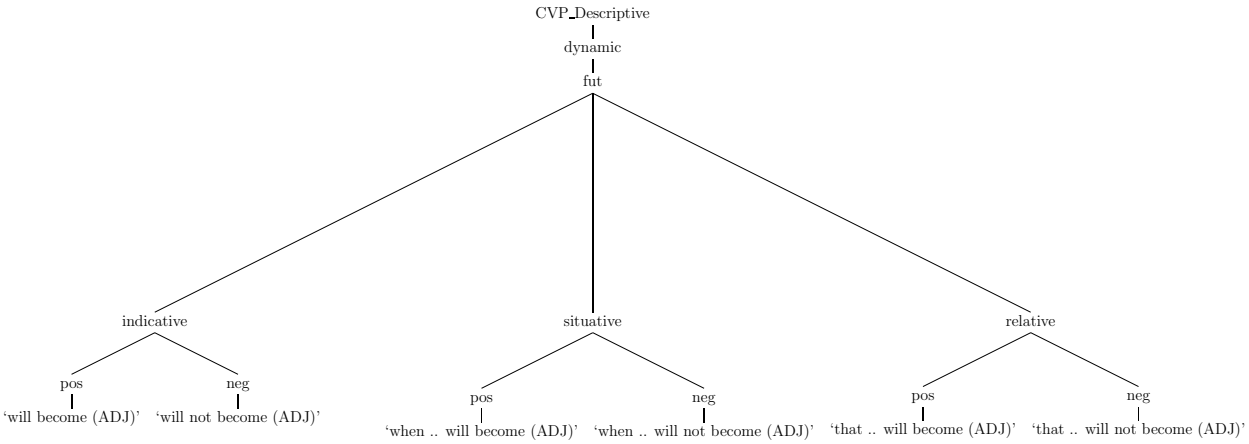


Figure 3.33: Descriptive dynamic future tense (cf. Table 3.44)

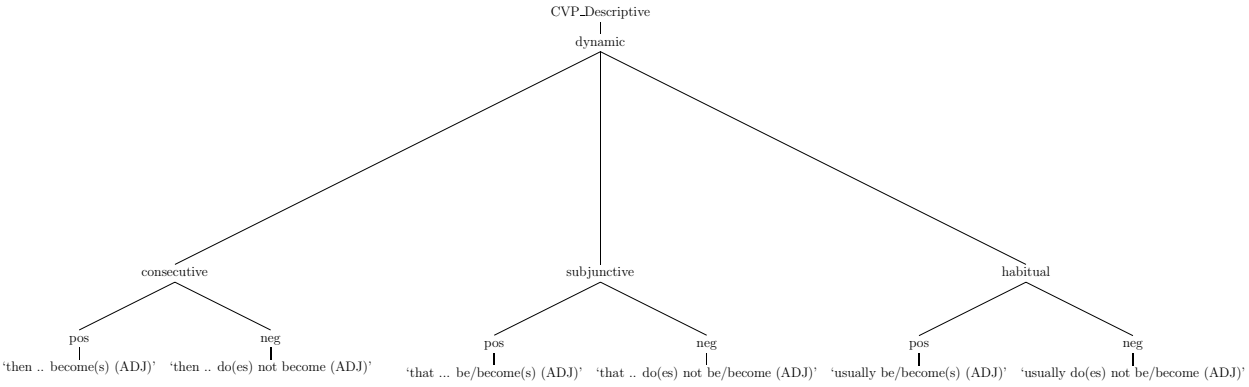


Figure 3.34: Identifying dynamic dependent clauses (cf. Tables 3.45)

Table 3.42: The **dynamic** forms of the **descriptive** copulative (COPDDC) (**present tense**, Figure 3.31)

COPDDC$_V$			
tense	**mood and actuality**	**elements**	**complement**
pres.	**ind pos**	**1CS**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$	nominal
		examples:	
	(monna)	o_{1CS01} ba_{VCOP} ‘(the man) becomes kind	$boleta_{\text{N14}}$
		$ke_{\text{1CSPERS\_1sg}}$ ba_{VCOP} ‘I become kind	$boleta_{\text{N14}}$
	ind neg	***ga***$_{\text{MORPH\_neg}}$ **2CS**$_{\text{categ}}$ ***be***$_{\text{VCOP}}$	nominal
	(monna)	$ga_{\text{MORPH\_neg}}$ a_{2CS01} be_{VCOP} ‘(the man) does not become kind	$boleta_{\text{N14}}$
		$ga_{\text{MORPH\_neg}}$ $ke_{\text{2CSPERS\_1sg}}$ be_{VCOP} ‘I do not become kind	$boleta_{\text{N14}}$
	sit pos	**2CS**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$	nominal
	(ge monna)	a_{2CS01} ba_{VCOP} ‘((as) the man) becomes kind	$boleta_{\text{N14}}$
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ ba_{VCOP} ‘(as) I become kind	$boleta_{\text{N14}}$
	sit neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***be***$_{\text{VCOP}}$	nominal
	(ge monna)	a_{2CS01} $sa_{\text{MORPH\_neg}}$ be_{VCOP} ‘(as the man) does not become kind	$boleta_{\text{N14}}$
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$ be_{VCOP} ‘(as) I do not become kind	$boleta_{\text{N14}}$
	rel pos	**2CS**$_{\text{categ}}$ ***bago***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} $bago_{\text{VCOP}}$ ‘(the man who) becomes kind	$boleta_{\text{N14}}$
	(nna yo)	$ke_{\text{2CSPERS\_sg}}$ $bago_{\text{VCOP}}$ ‘(I who) become kind	$boleta_{\text{N14}}$
	rel neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***bego***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} $sa_{\text{MORPH\_neg}}$ $bego_{\text{VCOP}}$ ‘(the man who) does not become kind	$boleta_{\text{N14}}$
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$ $bego_{\text{VCOP}}$ ‘(I who) do not become kind	$boleta_{\text{N14}}$

Table 3.43: The **dynamic** forms of the **descriptive** copulative (COPDDC) (**perfect tense**, Figure 3.32)

		$^{\text{COPDDC}}$V	
tense	**mood and actuality**	**elements**	**complement**
perfect	**ind pos**	**1CS**$_{\text{categ}}$ ***bile***$_{\text{VCOP}}$	nominal
	examples:		
	(monna)	o_{1CS01} $bile_{\text{VCOP}}$ '(the man) became kind'	$boleta_{\text{N14}}$
		$ke_{\text{1CSPERS\_1sg}}$ $bile_{\text{VCOP}}$ 'I became kind'	$boleta_{\text{N14}}$
	ind neg	***ga***$_{\text{MORPH\_neg}}$ ***se***$_{\text{MORPH\_neg}}$ **3CS**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$	nominal
	(monna)	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ a_{3CS01} ba_{VCOP} '(the man) did not become kind'	$boleta_{\text{N14}}$
		$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $ka_{\text{3CSPERS\_1sg}}$ ba_{VCOP} 'I did not become kind'	$boleta_{\text{N14}}$
	sit pos	**2CS**$_{\text{categ}}$ ***bile***$_{\text{VCOP}}$	nominal
	(ge monna)	a_{2CS01} $bile_{\text{VCOP}}$ '(when the man) became kind'	$boleta_{\text{N14}}$
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $bile_{\text{VCOP}}$ '(when) I became kind'	$boleta_{\text{N14}}$
	sit neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***ba***$_{\text{VCOP}}$	
	(ge monna)	a_{2CS01} $sa_{\text{MORPH\_neg}}$ ba_{VCOP} '(when the man) did not become kind'	$boleta_{\text{N14}}$
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$ ba_{VCOP} '(as) I did not become kind'	$boleta_{\text{N14}}$
	rel pos	**2CS**$_{\text{categ}}$ ***bilego***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} $bilego_{\text{VCOP}}$ '(the man who) became kind'	$boleta_{\text{N14}}$
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $bilego_{\text{VCOP}}$ '(I who) became kind'	$boleta_{\text{N14}}$
	rel neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$***bago***$_{\text{VCOP}}$	
	(monna yo)	a_{2CS01} $sa_{\text{MORPH\_neg}}$ $bago_{\text{VCOP}}$ '(the man who) did not become kind'	$boleta_{\text{N14}}$
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$ $bago_{\text{VCOP}}$ '(I who) did not become kind'	$boleta_{\text{N14}}$

Table 3.44: The **dynamic** forms of the **descriptive** copulative (COPDID) (**future tense**, Figure 3.33)

		$^{\text{COPDDC}}$V	
tense	**mood and actuality**	**tense marker and other elements**	**complement**
fut.	**ind pos**	**1CS**$_{\text{categ}}$ ***tlo/tla***$_{\text{MORPH\_fut}}$ ***ba***$_{\text{VCOP}}$	nominal
	examples:		
	(monna)	o_{1CS01} *tlo* ba_{VCOP}	$boleta_{\text{N14}}$
		'(the man) will become (kind)'	
		$ke_{\text{1CSPERS\_1sg}}$ *tlo* ba_{VCOP}	$boleta_{\text{N14}}$
		'I will become (kind)'	
	sit pos	**2CS**$_{\text{categ}}$ ***tlo/tla***$_{\text{MORPH\_fut}}$ ***ba***$_{\text{VCOP}}$	nominal
	(ge monna)	a_{2CS01} *tlo* ba_{VCOP}	$boleta_{\text{N14}}$
		'(when the man) will become (kind)'	
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ *tlo* ba_{VCOP}	$boleta_{\text{N14}}$
		'(when) I will become (kind)'	
	ind/sit neg	**2CS**$_{\text{categ}}$ ***ka***$_{\text{MORPH\_pot}}$ ***se***$_{\text{MORPH\_neg}}$***be***$_{\text{VCOP}}$	nominal
	((ge) monna)	a_{2CS01} $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP}	$boleta_{\text{N14}}$
		'((as) the man) will not become (kind)'	
	(ge)	*nka* ($ke_{\text{2CSPERS\_1sg}}$ + $ka_{\text{MORPH\_pot}}$) $se_{\text{MORPH\_neg}}$ be_{VCOP}	$boleta_{\text{N14}}$
		'(as) I will not become (kind)'	
	rel pos 1	**2CS**$_{\text{categ}}$ ***tlo/tla***$_{\text{MORPH\_fut}}$ ***bago***$_{\text{VCOP}}$	nominal
	rel pos 2	**2CS**$_{\text{categ}}$ ***tlogo/tlago***$_{\text{MORPH\_fut}}$ ***ba***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} *tlo* $bago_{\text{VCOP}}$	$boleta_{\text{N14}}$
		a_{2CS01} *tlogo* ba_{VCOP}	
		'(the man who) will become (kind)'	
	(nna yo)	$ke_{\text{2CSPERS\_sg}}$ *tla* $bago_{\text{VCOP}}$	$boleta_{\text{N14}}$
		$ke_{\text{2CSPERS\_sg}}$ *tlago* ba_{VCOP}	
		'(I who) become (kind)'	
	rel neg	**2CS**$_{\text{categ}}$ ***ka***$_{\text{MORPH\_pot}}$ ***se***$_{\text{MORPH\_neg}}$ ***bego***$_{\text{VCOP}}$	nominal
	(monna yo)	a_{2CS01} $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$ $bego_{\text{VCOP}}$	$boleta_{\text{N14}}$
		'(the man who) will not become (kind)'	
	(nna yo)	*nka* ($ke_{\text{1CSPERS\_1sg}}$ + $ka_{\text{MORPH\_pot}}$) $se_{\text{MORPH\_neg}}$ $bego_{\text{VCOP}}$	$boleta_{\text{N14}}$
		'(I who) will not become (kind)'	

Table 3.45: The **dynamic** forms of the **descriptive** copulative (COPDDCD) (**dependent constellations**, Figure 3.34)

	$^{\text{COPDDCD}}\text{V}$	
mood and actuality	**elements**	**complement**
consecutive pos	**3CS**$_{\text{categ}}$ ***ba***$_{\text{VCOP}}$	nominal
	examples: *(monna)* a_{3CS01} ba_{VCOP} *(bohlale)*$_{\text{N14}}$ '(then the man) becomes (clever)' $ka_{\text{3CSPERS\_1sg}}$ ba_{VCOP} *(bohlale)*$_{\text{N14}}$ '(then I) become (clever)'	
consecutive neg	**3CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$ ***be***$_{\text{VCOP}}$	nominal
	(monna) a_{3CS01} $se_{\text{MORPH\_neg}}$ be_{VCOP} *(bohlale)*$_{\text{N14}}$ '(then the man) does not become (clever)' $ka_{\text{3CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(bohlale)*$_{\text{N14}}$ '(then I) do not become (clever)'	
subjunctive pos habitual pos	**2CS**$_{\text{categ}}$ ***be***$_{\text{VCOP}}$	nominal
	((gore) monna) a_{2CS01} be_{VCOP} *(bohlale)*$_{\text{N14}}$ '(so that the man) becomes (clever)' '(the man) usually becomes (clever)' *(gore)* $ke_{\text{2CSPERS\_1sg}}$ be_{VCOP} *(bohlale)*$_{\text{N14}}$ '(so that) I become (clever)' 'I usually become (clever)'	
subjunctive neg habitual neg	**2CS**$_{\text{categ}}$ $se_{\text{MORPH\_neg}}$ ***be***$_{\text{VCOP}}$	nominal
	((gore) monna) a_{2CS01} $se_{\text{MORPH\_neg}}$ be_{VCOP} *(bohlale)*$_{\text{N14}}$ '(so that the man) does not become (clever)' '(the man) usually does not become (clever)' *(gore)* $ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ be_{VCOP} *(bohlale)*$_{\text{N14}}$ '(so that) I do not become (clever)' 'I usually do not become (clever)'	

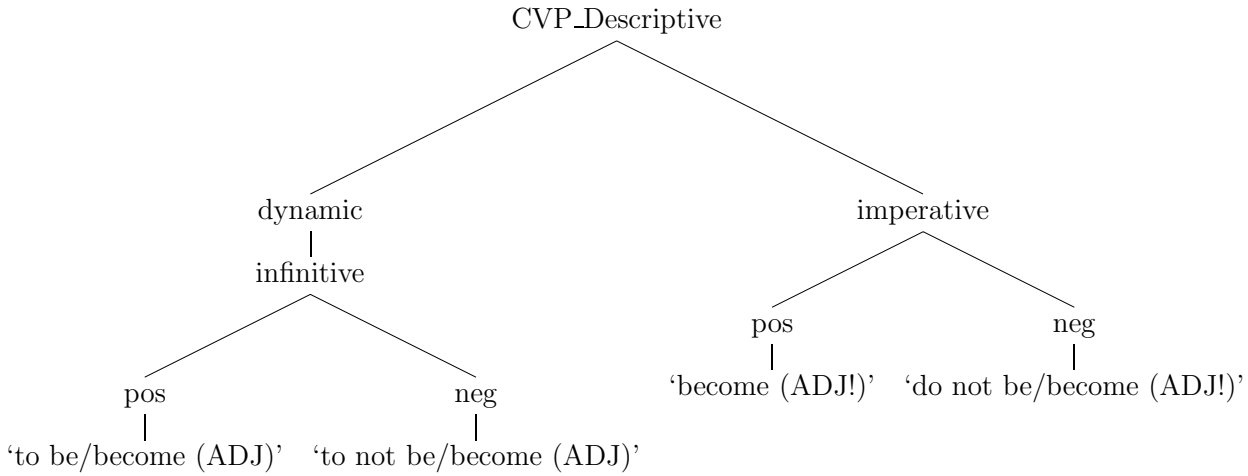


Figure 3.35: Identifying dynamic non-predicative clauses (cf. Table 3.39)

3.3.4 The associative copulative

Literally, this copulative is translated as 'be with', however, its usual meaning is 'to have'. Lombard (1985, p. 196) lists a few examples, of which some appear in 66 (a) to (d).

66(a) *mo*$_{\text{NLOC}}$ *seswantšong*$_{\text{N07\_loc}}$ *Mna.*$_{\text{ABBR}}$ *Ramokgopa*$_{\text{N01a}}$ *o*$_{\text{1CS01}}$ *na*$_{\text{VCOP}}$
on photograph Mr. Ramokgopa `subj-3rd-cl1` is
le$_{\text{PART\_con}}$ *mosadi*$_{\text{N01}}$ *wa*$_{\text{CPOSS01}}$ *gagwe*$_{\text{PROPOSS01}}$
`con` wife of his/hers
'on (the) photograph, Mr. Ramokgopa is with his wife'

(b) *bona*$_{\text{PROEMP02}}$ *ba*$_{\text{1CS02}}$ *na*$_{\text{VCOP}}$ *le*$_{\text{PART\_con}}$ *polase*$_{\text{N09}}$
`emp-3rd-cl02` `subj-3rd-cl2` is `con` farm
'these ones own (a) farm'

(c) *monna*$_{\text{N01}}$ *yo*$_{\text{CDEM01}}$ *a*$_{\text{2CS01}}$ *se*$_{\text{MORPH\_neg}}$ *nago*$_{\text{VCOP}}$ *thuthuthu*$_{\text{N09}}$
man `dem-3rd-cl1` `subj-3rd-cl1` `neg` that-is/have (a) motorbike
'(a) man who does not have a motorbike'

(d) *ga*$_{\text{MORPH\_neg}}$ *ke*$_{\text{1CSPERS\_1sg}}$ *na*$_{\text{VCOP}}$ *thuthuthu*$_{\text{N09}}$
`neg` `subj-1` is motorbike
'I don't have (a) motorbike'

(e) *monna*$_{\text{N01}}$ *ga*$_{\text{MORPH\_neg}}$ *a*$_{\text{2CS01}}$ *be*$_{\text{VCOP}}$ *le*$_{\text{PART\_con}}$ *tšhelete*$_{\text{N09}}$
man `neg` `subj-3rd-cl1` become `con` money
'(a) man does not become rich'

Examples 66 (a) and (b) illustrate the two meanings of the positive associative and may lead to the assumption, that the copula *na* itself does not indicate possession, as this aspect seems to be added by the connective particle *le* 'with'. However, this assumption is wrong as proved by examples (c) and (d) showing two regular negation constellations, where no connective particle appears. The connective particle however may appear in the negated dynamic forms, where it does not indicate possession, as in 66 (e).

These features constitute an interesting aspect of the associative copulative: as standalone items, the meaning of neither the copula *a* nor the connective particle *le* can be established; their context, i.e. their complements, must be taken into account. In the example 66 (a), *na le* is to be understood as 'to be with', while in 66 (b) they refer to ownership. In 66 (d), *na* alone has the same complement and can be understood the same way as *na le* in 66 (b), even though it is preceded here by a negation cluster.

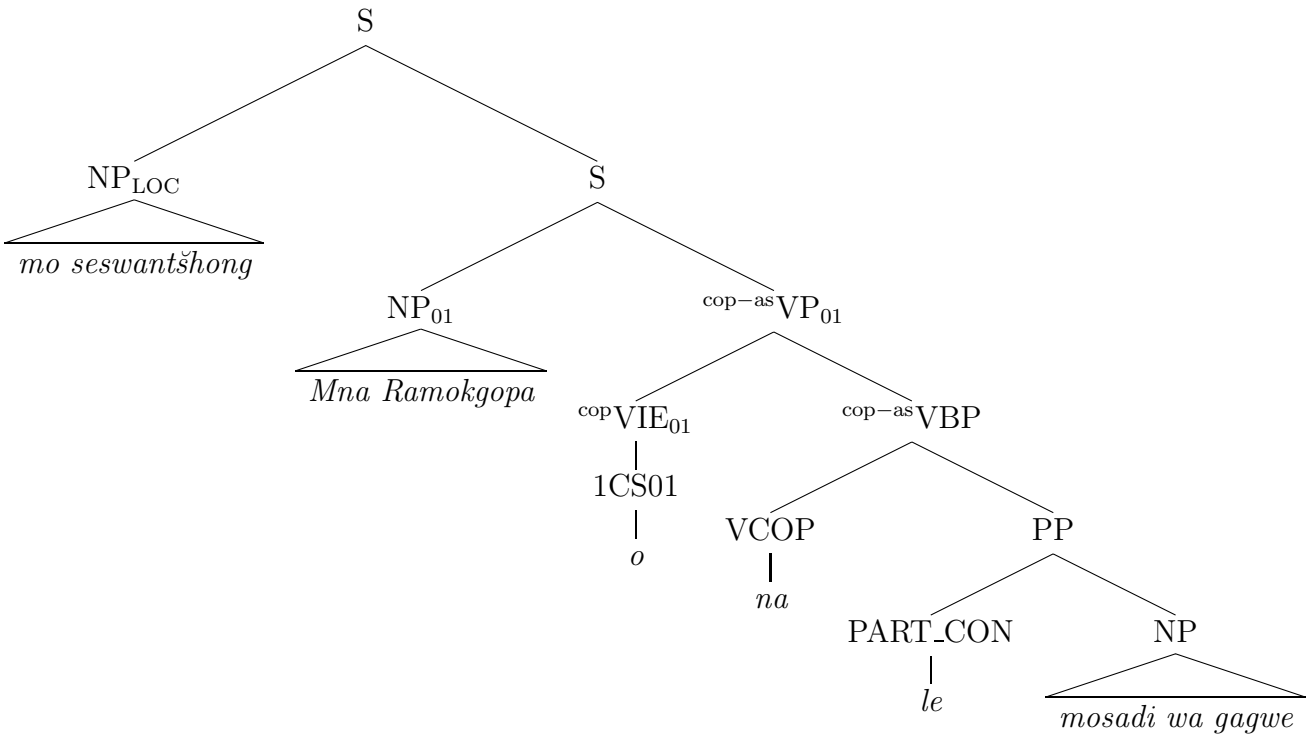


Figure 3.36: *mo seswantšhong Mna Ramokgopa o na le mosadi wa gagwe* 'on (the) photograph, Mr. Ramokgopa is with his wife'

It seems that copula and complement can be analysed similarly to a main verb and complemented while all other preceding elements may be seen as inflectional and thus constituting a separate element. In order to reflect the close relation between the copula and its complement, we have therefore opted for a morphosyntactic analysis similar to the successful strategy for main verbs in splitting the copulative into an ^{cop}VIE and a $^{cop-as}$VBP. Moreover, although the meanings of the copulas described in the previous paragraphs are not dependent on their complements, for the sake of consistency, we should like to suggest a similar strategy for their morphosyntactic analyses. Consequently, all verbal phrases of Northern Sotho will be analysed similarly. Figures 3.36 to 3.38 demonstrate such morphosyntactic analyses of examples 66 (a) to (c).

For each constellation, the associative copulative VBP, named $^{cop-as}$VBP is to be defined to include either a particle phrase headed by a connective particle (^{con}PP), or a nominal phrase (NP). The ^{cop}VIE is clearly marked as such, as it should not be confused with the other VIEs defined before (cf. paragraphs 3.2.1.1 and e.g. 3.2.5).

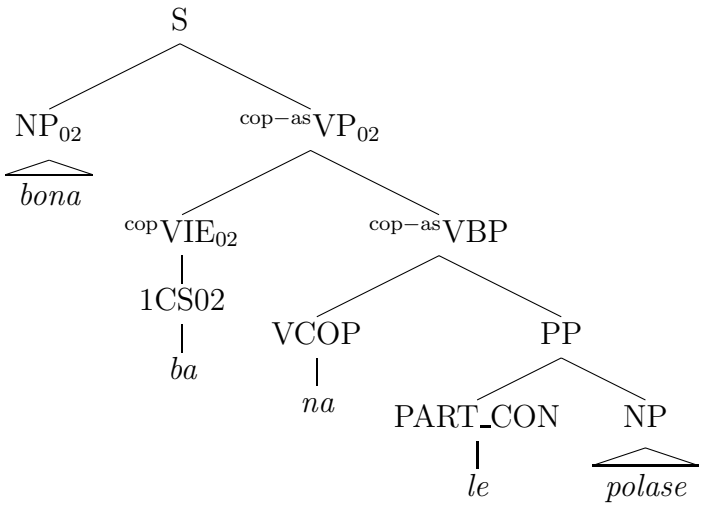


Figure 3.37: *bona ba na le polase* 'these ones own (a) farm'

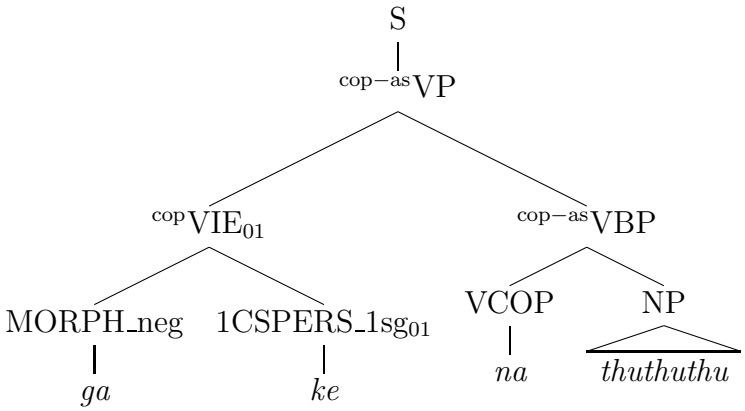


Figure 3.38: *ga ke na thuthuthu* 'I don't have (a) motorbike'

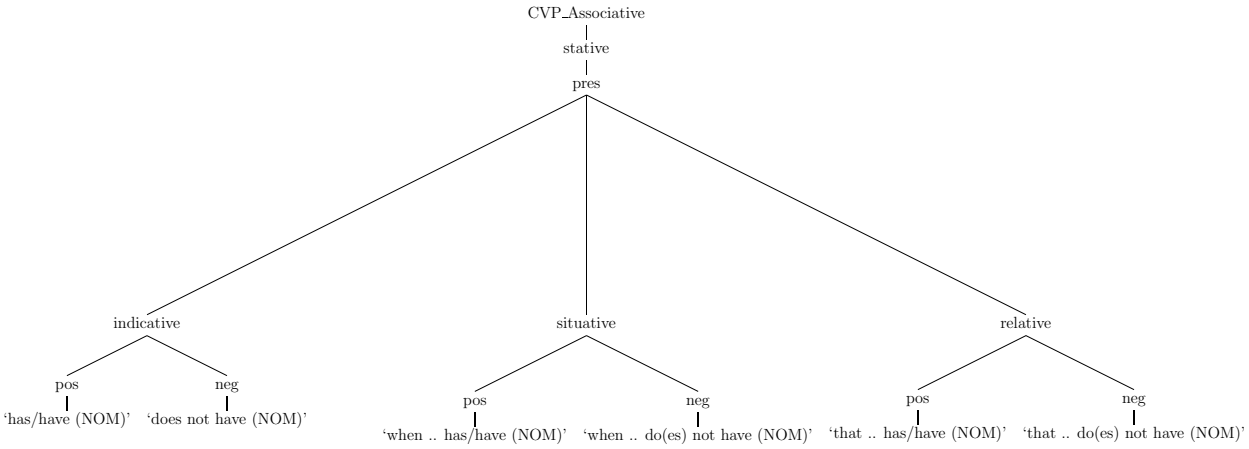


Figure 3.39: Associative stative present tense (cf. Table 3.46)

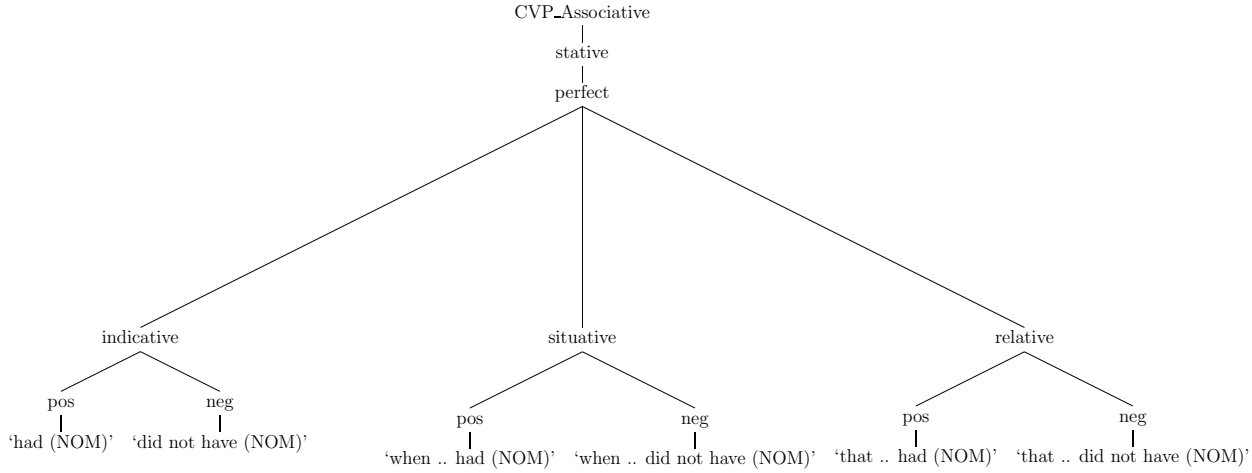


Figure 3.40: Associative stative perfect tense (cf. Table 3.47)

The following paragraphs illustrate the stative and the dynamic constellations of the associative copulative and the morphosyntactic rules to form them.

3.3.4.1 The stative

As Figures 3.39 and 3.40 show, this copulative usually indicates possession, though in some cases its meaning may be 'to be with'. Note that the copula *ena* may in some constellations replace *na*; this will be indicated by the appearance of *e(na)* in the appropriate rules.

Table 3.46: The **stative** forms of the **associative** copulative (COPSAS) (**present tense**, Figure 3.39)

		$^{\text{COPSAS}}$V		
tense	**mood and actuality**	**elements** $^{\text{cop}}$VIE	VCOP	complement
pres	**ind pos**	**1CS**$_{\text{categ}}$	***na***$_{\text{VCOP}}$	$^{\text{con}}$PP
	examples:			
	(monna)	o_{1CS01}	na_{VCOP}	*le tšhelete*
		‘(the man) has money’		
		$ke_{\text{1CSPERS\_1sg}}$	na_{VCOP}	*le tšhelete*
		‘I have money’		
	ind neg	***ga***$_{\text{MORPH\_neg}}$ **2CS**$_{\text{categ}}$	***na***$_{\text{VCOP}}$	NP
	(monna)	$ga_{\text{MORPH\_neg}}$ a_{2CS01}	na_{VCOP}	*tšhelete*
		‘(the man) does not have money’		
		$ga_{\text{MORPH\_neg}}$ $ke_{\text{2CSPERS\_1sg}}$	na_{VCOP}	*tšhelete*
		‘I do not have money’		
	sit pos	**2CS**$_{\text{categ}}$	***(e)na***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(ge monna)	a_{2CS01}	*(e)*na_{VCOP}	*le tšhelete*
		‘(when the man) has money’		
	(ge)	$ke_{\text{2CSPERS\_1sg}}$	*(e)*na_{VCOP}	*le tšhelete*
		‘(when) I have money’		
	sit neg	**2CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$	***na***$_{\text{VCOP}}$	NP
	(ge monna)	a_{2CS01} $se_{\text{MORPH\_neg}}$	na_{VCOP}	*tšhelete*
		‘(when the man) does not have money’		
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$	na_{VCOP}	*tšhelete*
		‘(when) I do not have money’		
	rel pos	**2CS**$_{\text{categ}}$	***nago***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01}	$nago_{\text{VCOP}}$	*le tšhelete*
		‘(the man who) has money’		
		$ke_{\text{2CSPERS\_1sg}}$	$nago_{\text{VCOP}}$	*le tšhelete*
		‘(I who) have money’		
	rel neg	**2CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$	***nago***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01} $se_{\text{MORPH\_neg}}$	$nago_{\text{VCOP}}$	*le tšhelete*
		‘(the man who) does not have money’		
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$	$nago_{\text{VCOP}}$	*le tšhelete*
		‘(I who) do not have money’		

Table 3.47: The **stative** forms of the **associative** copulative (COPSAS) (**perfect tense**, Figure 3.40)

		$^{\text{COPSAS}}$V		
tense	**mood / actuality**	**elements** $^{\text{cop}}$VIE	VCOP	complement
perfect	**ind pos**	**1CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$	***(e)na***$_{\text{VCOP}}$	$^{\text{con}}$PP
	examples:			
	(monna)	o_{1CS01} $be_{\text{V\_aux}}$ a_{2CS01} '(the man) had money'	$(e)na_{\text{VCOP}}$	*le tšhelete*
		$ke_{\text{1CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ '(I) had money'	$(e)na_{\text{VCOP}}$	*le tšhelete*
	ind neg	**1CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$	***(e)na***$_{\text{VCOP}}$	NP
	(monna)	o_{1CS01} $be_{\text{V\_aux}}$ a_{2CS01} $se_{\text{MORPH\_neg}}$ '(the man) did not have money'	$(e)na_{\text{VCOP}}$	*tšhelete*
		$ke_{\text{1CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ '(I) did not have money'	$(e)na_{\text{VCOP}}$	*tšhelete*
	sit pos	**2CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$	***(e)na***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(ge monna)	a_{2CS01} $be_{\text{V\_aux}}$ a_{2CS01} '(when the man) had money'	$(e)na_{\text{VCOP}}$	*le tšhelete*
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ '(when) I had money'	$(e)na_{\text{VCOP}}$	*le tšhelete*
	sit neg	**2CS**$_{\text{categ}}$ ***be***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$	***(e)na***$_{\text{VCOP}}$	NP
	(ge monna)	a_{2CS01} $be_{\text{V\_aux}}$ a_{2CS01} $se_{\text{MORPH\_neg}}$ $(e)na_{\text{VCOP}}$ '(when the man) did not have money'	*tšhelete*	
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $be_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ '(when) I did not have money'	$(e)na_{\text{VCOP}}$	*tšhelete*
	rel pos	**2CS**$_{\text{categ}}$ ***bego***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$	***na***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01} $bego_{\text{V\_aux}}$ a_{2CS01} '(the man who) had money'	na_{VCOP}	*le tšhelete*
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $bego_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ '(I who) had money'	na_{VCOP}	*le tšhelete*
	rel neg	**2CS**$_{\text{categ}}$ ***bego***$_{\text{V\_aux}}$ **2CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$	***na***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01} $bego_{\text{V\_aux}}$ a_{2CS01} $se_{\text{MORPH\_neg}}$ '(the man who) did not have money'	na_{VCOP}	*le tšhelete*
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $bego_{\text{V\_aux}}$ $ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$ '(I who) did not have money'	na_{VCOP}	*le tšhelete*

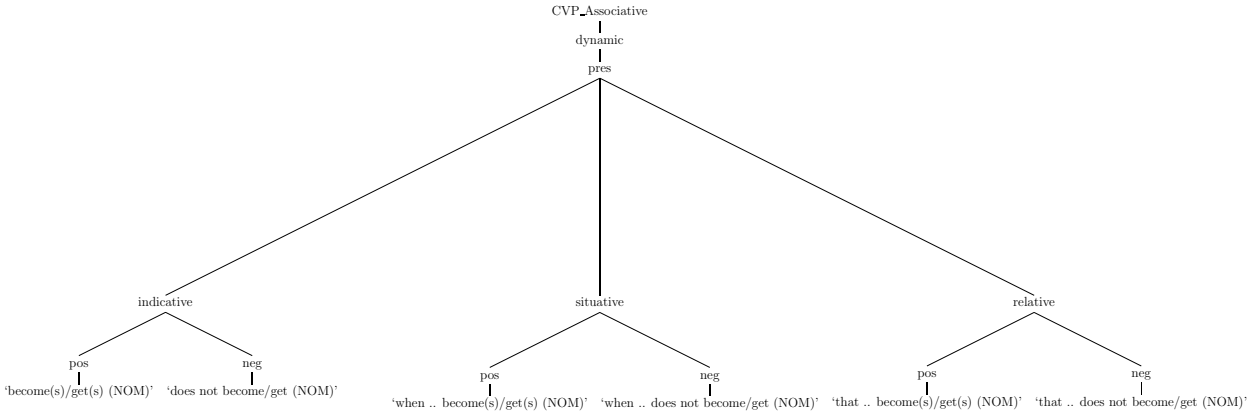


Figure 3.41: Associative dynamic present tense (cf. Table 3.48)

3.3.4.2 The dynamic

The dynamic associative shown in Figures 3.41 to 3.46 – like the dynamic forms of the identifying and descriptive copulative – contains an inchoative aspect. It makes use of *ba* 'become' and its variants *eba*, *be* used in negations, *bile* 'became', and ultimately its relative forms, *bago*, *bego*, *bilego* 'who becomes/became'. The use of the copula *be* is illustrated in the (negated) example 66 (e) of page 154 repeated here as 67 (a). This copulative can however also be translated as 'get' or 'acquire', like in 67 (b). All dynamic constellations of the associative have to be complemented by a connective particle phrase which reflects in the rules defined in Tables 3.48 to 3.52.

67 (a) *monna*$_{\text{N01}}$ *ga*$_{\text{MORPH\_neg}}$ *a*$_{\text{2CS01}}$ *be*$_{\text{VCOP}}$ *le*$_{\text{PART\_con}}$ *tšhelete*$_{\text{N09}}$
(the) man `neg` `subj-3rd-cl1` become `con` money
'(a) man does not become rich'

(b) *monna*$_{\text{N01}}$ *o*$_{\text{2CS01}}$ *ba*$_{\text{VCOP}}$ *le*$_{\text{PART\_con}}$ *mpša*$_{\text{N09}}$
(the) man `neg` `subj-3rd-cl1` become `con` dog
'(a) man gets a dog'

Table 3.51 (page 166) shows that the dependent clauses of the associative copulative only differ from those of the descriptive copulative (illustrated in Table 3.45 on page 152) in their complement, while it was nominal in the descriptive constellations, it must consist of a connective particle phrase in the associative, this underlies the necessity to analyse copula and complement on one level of description in order to determine the correct category.

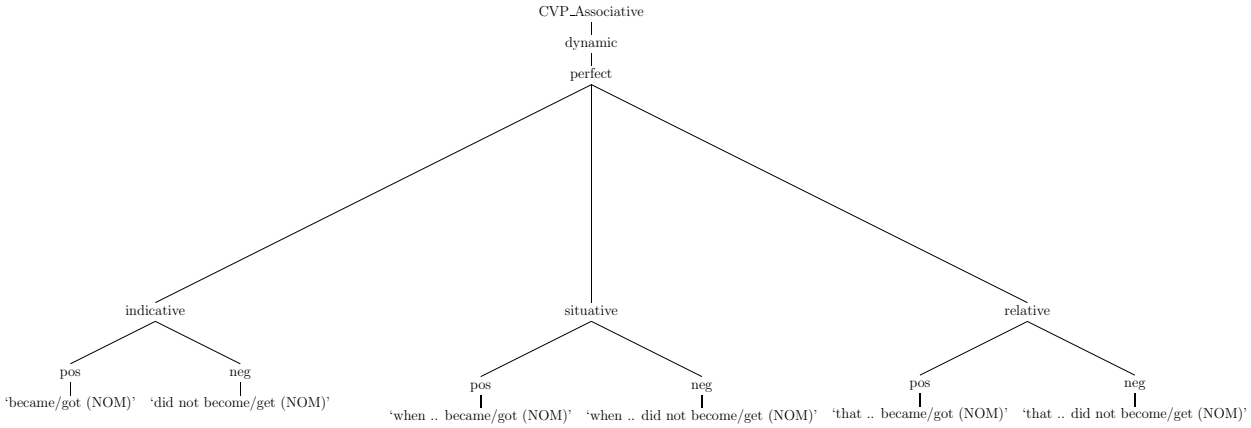


Figure 3.42: Associative dynamic perfect tense (cf. Table 3.49)

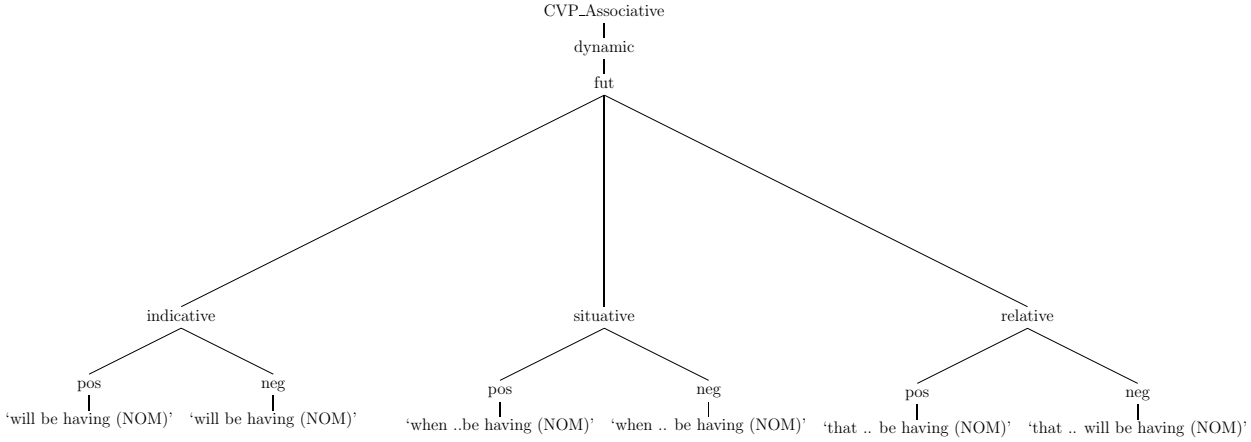


Figure 3.43: Associative dynamic future tense (cf. Table 3.50)

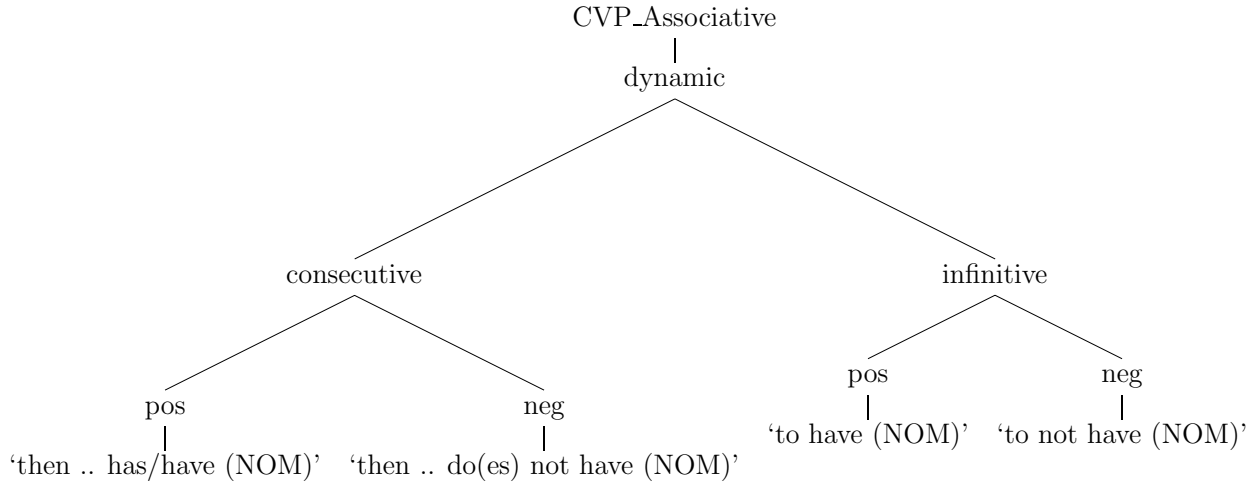


Figure 3.44: Associative dynamic dependent clauses 1 of 2 (cf. Table 3.51)

Table 3.48: The **dynamic** forms of the **associative** copulative (COPDAS) (**present tense**, Figure 3.41)

		$^{\text{COPDAS}}$V		
tense	**mood and actuality**	**elements** $^{\text{cop}}$VIE	VCOP	complement
pres	**ind pos**	**1CS**$_{\text{categ}}$	***ba***$_{\text{VCOP}}$	$^{\text{con}}$PP
	examples:			
	(monna)	o_{1CS01}	ba_{VCOP}	*le tšhelete*
		‘(the man) becomes rich’		
		$ke_{\text{1CSPERS\_1sg}}$	ba_{VCOP}	*le tšhelete*
		‘(I) become rich’		
	ind neg	***ga***$_{\text{MORPH\_neg}}$ **2CS**$_{\text{categ}}$	***be***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna)	$ga_{\text{MORPH\_neg}}$ a_{2CS01}	be_{VCOP}	*le tšhelete*
		‘(the man) does not becomes rich’		
		$ga_{\text{MORPH\_neg}}$ $ke_{\text{2CSPERS\_1sg}}$	be_{VCOP}	*le tšhelete*
		‘(I) do not become rich’		
	sit pos	**2CS**$_{\text{categ}}$	***eba***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(ge monna)	a_{2CS01}	eba_{VCOP}	*le tšhelete*
		‘(when the man) becomes rich’		
	(ge)	$ke_{\text{2CSPERS\_1sg}}$	eba_{VCOP}	*le tšhelete*
		‘(when) I become rich’		
	sit neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$	***be***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(ge monna)	a_{2CS01} $sa_{\text{MORPH\_neg}}$	be_{VCOP}	*le tšhelete*
		‘(when the man) does not become rich’		
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$	be_{VCOP}	*le tšhelete*
		‘(when) I do not become rich’		
	rel pos	**2CS**$_{\text{categ}}$	***bago***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01}	$bago_{\text{VCOP}}$	*le tšhelete*
		‘(the man who) becomes rich’		
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$	$bago_{\text{VCOP}}$	*le tšhelete*
		‘(I who) become rich’		
	rel neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$	***bego***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01} $sa_{\text{MORPH\_neg}}$	$bego_{\text{VCOP}}$	*le tšhelete*
		‘(the man who) does not become rich’		
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$	$bego_{\text{VCOP}}$	*le tšhelete*
		‘(I who) do not become rich’		

Table 3.49: The **dynamic** forms of the **associative** copulative (COPDAS) (**perfect tense**, Figure 3.42)

		$^{\text{COPDAS}}$V		
tense	**mood and actuality**	**elements** $^{\text{cop}}$VIE	VCOP	complement
perfect	**ind pos**	**1CS**$_{\text{categ}}$	***bile***$_{\text{VCOP}}$	$^{\text{con}}$PP
	examples:			
	(monna)	o_{1CS01}	$bile_{\text{VCOP}}$	*le tšhelete*
		‘(the man) became rich’		
		$ke_{\text{1CSPERS\_1sg}}$	$bile_{\text{VCOP}}$	*le tšhelete*
		‘(I) became rich’		
	ind neg	***ga***$_{\text{MORPH\_neg}}$ ***se***$_{\text{MORPH\_neg}}$ **3CS**$_{\text{categ}}$	***ba***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna)	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ a_{3CS01}	ba_{VCOP}	*le tšhelete*
		‘(the man) did not become rich’		
		$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $ka_{\text{3CSPERS\_1sg}}$	ba_{VCOP}	*le tšhelete*
		‘(I) did not became rich’		
	sit pos	**2CS**$_{\text{categ}}$	***bile***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(ge monna)	a_{2CS01}	$bile_{\text{VCOP}}$	*le tšhelete*
		‘(when the man) became rich’		
	(ge)	$ke_{\text{2CSPERS\_1sg}}$	$bile_{\text{VCOP}}$	*le tšhelete*
		‘(when) I became rich’		
	sit neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$	***ba***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(ge monna)	a_{2CS01} $sa_{\text{MORPH\_neg}}$	ba_{VCOP}	*le tšhelete*
		‘(when the man) did not become rich’		
	(ge)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$	ba_{VCOP}	*le tšhelete*
		‘(when) I did not become rich’		
	rel pos	**2CS**$_{\text{categ}}$	***bilego***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01}	$bilego_{\text{VCOP}}$	*le tšhelete*
		‘(the man who) became rich’		
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$	$bilego_{\text{VCOP}}$	*le tšhelete*
		‘(I who) became rich’		
	rel neg	**2CS**$_{\text{categ}}$ ***sa***$_{\text{MORPH\_neg}}$	***bago***$_{\text{VCOP}}$	$^{\text{con}}$PP
	(monna yo)	a_{2CS01} $sa_{\text{MORPH\_neg}}$	$bago_{\text{VCOP}}$	*le tšhelete*
		‘(the man who) did not become rich’		
	(nna yo)	$ke_{\text{2CSPERS\_1sg}}$ $sa_{\text{MORPH\_neg}}$	$bago_{\text{VCOP}}$	*le tšhelete*
		‘(I who) did not become rich’		

Table 3.50: The **dynamic** forms of the **associative** copulative (COPDAS) (**future tense**, Figure 3.43)

		$^{\text{COPDAS}}$V		
tense	**mood and actuality**	**elements** $^{\text{cop}}$VIE	VCOP	complement
fut	**ind pos**	**1CS$_{\text{categ}}$ *tlo/tla***	***ba*$_{\text{VCOP}}$**	$^{\text{con}}$PP
	examples:			
	(monna)	*o*$_{\text{1CS01}}$ *tlo/tla* ‘(the man) will get a dog’	*ba*$_{\text{VCOP}}$	*le mpša*
		ke$_{\text{1CSPERS\_1sg}}$ *tlo/tla* ‘I will get a dog’	*ba*$_{\text{VCOP}}$	*le mpša*
	sit pos	**2CS$_{\text{categ}}$ *tlo/tla***	***ba*$_{\text{VCOP}}$**	$^{\text{con}}$PP
	(ge monna)	*a*$_{\text{2CS01}}$ *tlo/tla* ‘(when the man) will get a dog’	*ba*$_{\text{VCOP}}$	*le mpša*
	(ge)	*ke*$_{\text{2CSPERS\_1sg}}$ *tlo/tla* ‘(when) I will get a dog’	*ba*$_{\text{VCOP}}$	*le mpša*
	ind/sit neg	**2CS$_{\text{categ}}$ *ka*$_{\text{MORPH\_pot}}$ *se*$_{\text{MORPH\_neg}}$**	***be*$_{\text{VCOP}}$**	$^{\text{con}}$PP
	((ge) monna)	*a*$_{\text{2CS01}}$ *ka*$_{\text{MORPH\_pot}}$ *se*$_{\text{MORPH\_neg}}$ ‘((when) the man) will not get a dog’	*be*$_{\text{VCOP}}$	*le mpša*
	(ge)	*nka* (=*ke*$_{\text{2CSPERS\_1sg}}$ *ka*$_{\text{MORPH\_pot}}$) *se*$_{\text{MORPH\_neg}}$ ‘(when) I will not get a dog’	*be*$_{\text{VCOP}}$	*le mpša*
	rel pos	**2CS$_{\text{categ}}$ *tlo/tla***	***bago*$_{\text{VCOP}}$**	$^{\text{con}}$PP
	(monna yo)	*a*$_{\text{2CS01}}$ *tlo/tla* ‘(the man who) will get a dog’	*bago*$_{\text{VCOP}}$	*le mpša*
	(nna yo)	*ke*$_{\text{2CSPERS\_1sg}}$ *tlo/tla* ‘(I who) will get a dog’	*bago*$_{\text{VCOP}}$	*le mpša*
	rel neg	**2CS$_{\text{categ}}$ *ka*$_{\text{MORPH\_pot}}$ *se*$_{\text{MORPH\_neg}}$**	***bego*$_{\text{VCOP}}$**	$^{\text{con}}$PP
	(monna yo)	*a*$_{\text{2CS01}}$ *ka*$_{\text{MORPH\_pot}}$ *se*$_{\text{MORPH\_neg}}$ ‘(the man who) will not get a dog’	*bego*$_{\text{VCOP}}$	*le mpša*
	(nna yo)	*nka* (=*ke*$_{\text{2CSPERS\_1sg}}$ *ka*$_{\text{MORPH\_pot}}$) *se*$_{\text{MORPH\_neg}}$ ‘(I who) will not get a dog’	*bego*$_{\text{VCOP}}$	*le mpša*

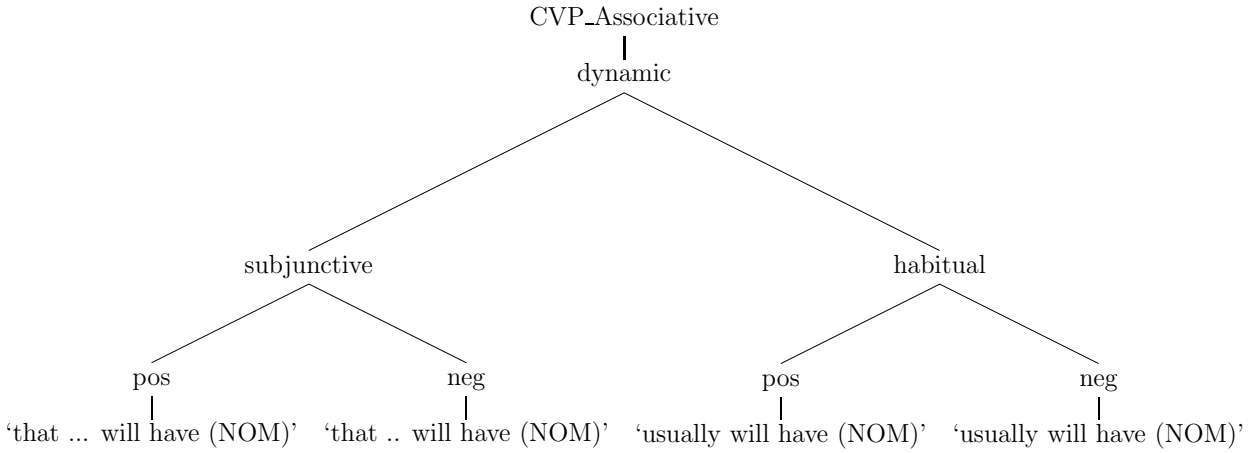


Figure 3.45: Associative dynamic dependent clauses 2 of 2 (cf. Tables 3.51)

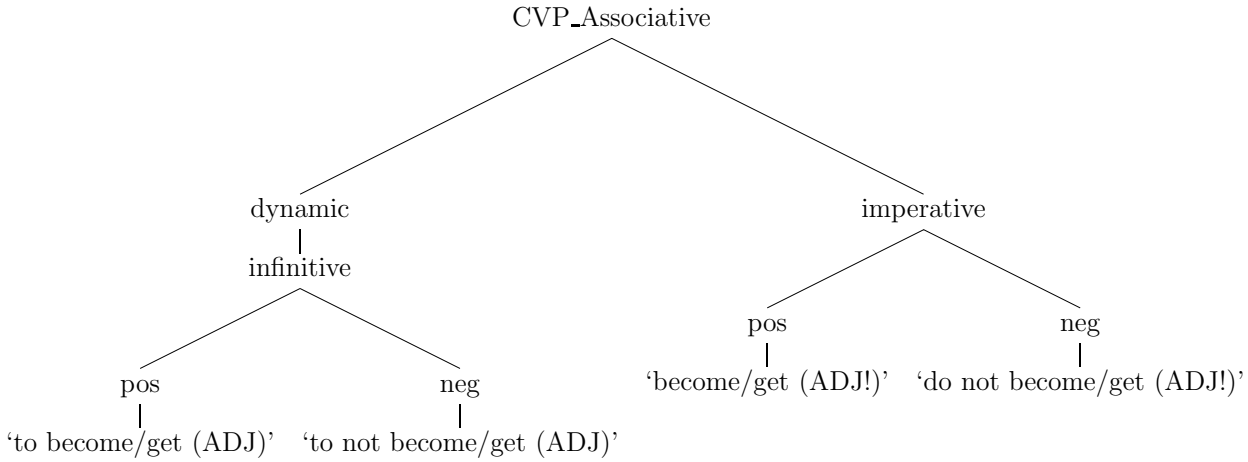


Figure 3.46: Associative dynamic non-predicative clauses (cf. Table 3.52)

Table 3.51: The **dynamic** forms of the **associative** copulative (COPDACD) (**dependent constellations**, Figures 3.44/3.45)

$^{\text{COPDACD}}$V			
mood and actuality	**elements** $^{\text{cop}}$**VIE**	VCOP	complement
consecutive pos	**3CS**$_{\text{categ}}$	***ba***$_{\text{VCOP}}$	$^{\text{con}}$PP
	examples:		
(monna)	a_{3CS01}	ba_{VCOP}	*le mpša*
	‘(then the man) gets a dog’		
	$ka_{\text{3CSPERS\_1sg}}$	ba_{VCOP}	*le mpša*
	‘(then) I get a dog’		
consecutive neg	**3CS**$_{\text{categ}}$ ***se***$_{\text{MORPH\_neg}}$	***be***$_{\text{VCOP}}$	$^{\text{con}}$PP
(monna)	a_{3CS01} $se_{\text{MORPH\_neg}}$	be_{VCOP}	*le mpša*
	‘(then the man) does not get a dog’		
	$ka_{\text{3CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$	be_{VCOP}	*le mpša*
	‘(then) I do not get a dog’		
subjunctive pos habitual pos	**2CS**$_{\text{categ}}$	***be***$_{\text{VCOP}}$	$^{\text{con}}$PP
((gore) monna)	a_{2CS01} be_{VCOP}	*le mpša*	
	‘(so that the man) gets a dog’		
	‘(the man) usually gets a dog’		
(gore)	$ke_{\text{2CSPERS\_1sg}}$ be_{VCOP}	*le mpša*	
	‘(so that) I get a dog’		
	‘I usually get (a dog)’		
subjunctive neg habitual neg	**2CS**$_{\text{categ}}$ $se_{\text{MORPH\_neg}}$	***be***$_{\text{VCOP}}$	$^{\text{con}}$PP
((gore) monna)	a_{2CS01} $se_{\text{MORPH\_neg}}$	be_{VCOP}	*le mpša*
	‘(so that the man) does not get a dog’		
	‘(the man) usually does not get a dog’		
(gore)	$ke_{\text{2CSPERS\_1sg}}$ $se_{\text{MORPH\_neg}}$	be_{VCOP}	*le mpša*
	‘(so that) I do not get a dog’		
	‘I usually do not get a dog’		

Table 3.52: The **dynamic** forms of the **associative** copulative (COPDA) (**non-predicative** constellations, Figure 3.46)

$^{\text{COPDA}}$V			
mood and actuality	**elements** $^{\text{cop}}$VIE	VCOP	complement
infinitive pos	***go***$_{\text{MORPH\_cp15}}$	***ba***$_{\text{VCOP}}$	$^{\text{con}}$PP
	example:		
	go$_{\text{MORPH\_cp15}}$	*ba*$_{\text{VCOP}}$	*le mpša*
	‘to acquire a dog’		
infinitive neg	***go***$_{\text{MORPH\_cp15}}$ ***se***$_{\text{MORPH\_neg}}$	***be***$_{\text{VCOP}}$	$^{\text{con}}$PP
	go$_{\text{MORPH\_cp15}}$ *se*$_{\text{MORPH\_neg}}$	*be*$_{\text{VCOP}}$	*le mpša*
	‘not to / to not acquire a dog’		
imperative pos		***eba***$_{\text{VCOP}}$	$^{\text{con}}$PP
		eba$_{\text{VCOP}}$	*le mpša*!
	‘get a dog!’		
imperative neg	**se**$_{\text{MORPH\_neg}}$	**be**$_{\text{VCOP}}$	$^{\text{con}}$PP
	se$_{\text{MORPH\_neg}}$	*be*$_{\text{VCOP}}$	*le mpša*!
	‘do not get a dog!’		

Concerning the non-predicative cases, i.e. the imperative and the infinitive, again only the complementary connective particle phrase (cf. 3.52) identifies the associative copulative as its constellations are identical to the ones shown in Table 3.39 for the identifying/descriptive cases.

3.3.5 The copulative constellations: A summary

As was shown in the previous paragraphs, the three groups, identifying, descriptive, and associative copulatives, may in turn each be divided into two categories, stative and dynamic/inchoative. Copulatives occur in moods, and as such there are a number of different forms, comparable to the constellations described for main verbs. The many forms often differ only slightly, for instance, containing subject concords of set 1 versus subject concords of set 2. Whenever the subject refers to the first or second person, the differences between identifying and descriptive copulative can often not be recognised at all. Moreover, the two subject concord sets used by the majority of cases only differ in the concords of class 1 (*o* in set 1 versus *a* in set 2). It might however be possible to distinguish the cases by taking their complements into account. As the semantic content of the complements are not included in the scope of this study, an attempt was made to at least narrow the cases by defining word classes that in the majority of cases appear as such complements, i.e. nouns for the descriptive copulative and nominal phrases or connective particle phrases for the associative. Lastly, just as for main verbs, phrase-introducing conjunctions may help to distinguish the indicative from the situative mood.

3.4 Auxiliary verbs

From a grammatical perspective, auxiliaries (cf. paragraph 2.7.3 on page 51) subcategorise verbal phrases as they require a full verb as their complement. According to the grammatical descriptions (e.g. (Lombard, 1985, p. 186 et seq.)) the subcategorised verb has to follow the auxiliary and begins with a subject concord of set 2 or set 3. However, like the subcategory verb, the auxiliary itself needs a subject concord (of set 1) to link it to the subject. This subject concord precedes the auxiliary, like in (68).

(68)

wena$_{\text{PROEMPPERS\_2sg}}$	***o***$_{\text{1CSPERS\_2sg}}$	***be***$_{\text{V\_aux}}$	*o*$_{\text{2CSPERS\_2sg}}$	*šomile*$_{\text{V\_itr}}$
`emp-2nd-sg`	`subj-2nd-sg`	`past`	`subj-2nd-sg`	work

'you have worked'

A simple phrase grammar rule to describe the auxiliary would thus be VP $\rightarrow$ $^{\text{AUX}}$VP $^{\text{complementary}}$VP. However, the auxiliary additionally requires the complementary VP to have the same category, i.e. both have to agree in their noun class (and with the subject noun they both refer to). Auxiliary constellations appear in the perfect, imperfect (as described in paragraph 2.7.3, these tenses are lexicalised) and future tense. The data shown in Table 3.53 is based on Poulos and Louwrens (1994, p. 247 et seq. and p. 256), though not all constellations described there will be found in this study, as no further proof of their existence in our corpora was found. The examples contained in this table are also taken from Louwrens (1991, p. 52).

Table 3.53: The auxiliary verbal phrase $^{\text{AUX}}$VP

	$^{\text{AUX}}$VP				
tense	**elements**		**complement**		
pres/perf.	**1CS**$_{\text{categ}}$	**V_aux**	**VP**$_{\text{categ}}$		
	example:				
	ba$_{\text{1CS02}}$	*setše*$_{\text{V\_aux}}$	(*ba*$_{\text{1CS02}}$	*šoma*	Tshwane)
	subj-cl02	already	(subj-cl02	work	Pretoria
	'they already (work in Pretoria.)'				
	example:				
	o$_{\text{1CS01}}$	*be*$_{\text{V\_aux}}$	(*a*	*sa*	*tsebe*)
	subj-cl01	perfect	(subj-cl01	neg	know
	'(s)he did (not know)'				
future	**1CS**$_{\text{categ}}$ ***tlo/tla***$_{\text{MORPH\_fut}}$	**V_aux**	**VP**$_{\text{categ}}$		
	example:				
	o$_{\text{1CS01}}$ *tlo*$_{\text{MORPH\_fut}}$	*tsama*$_{\text{V\_aux}}$	(*a*	*nwa*	*bjalwa*)
	subj- fut cl01	continually	(subj-cl01	drink	beer)
	'(s)he will continually (drink beer)'				
neg	***ga***$_{\text{MORPH\_neg}}$**1CS**$_{\text{categ}}$	**V_aux**	**VP**$_{\text{categ}}$		
	example:				
	ga$_{\text{MORPH\_neg}}$ *ke*$_{\text{CSPERS\_1sg}}$	*ešo*	*ka*	*reka*	*dipuku*
	neg subj-1-sg	yet	subj-1-sg	buy	books
	'I have not yet bought the books'				

3.5 Other verbal structures

3.5.1 The hortative constellation

Hortatives are described by Lombard (1985, pp. 155 to 156 and p. 171) as to express 'wishes and requests'. The hortative word class only consists of few forms, '*a*, *ake*, *anke* and *ga*, as described in paragraph 2.10.4. This particle can precede any subjunctive VP and is found in slot zero-3, as Table 3.54 demonstrates.

Table 3.54: The hortative constellation

description	**zero-3**	**zero-2 to zero**
	PART_hort	$^{\text{subjunctive}}$VP
Example	$a_{\text{PART\_hort}}$	$re_{\text{CSPERS\_2sg}}$ $reke_{\text{V\_tr}}$ $dipuku_{\text{N10}}$ 'let us buy books'

3.5.2 Potential forms

The potential expresses modal aspects of a following predicate in terms of its possibility. Poulos and Louwrens (1994) devote two sections of their book (paragraph 5.13, pp. 229 to 234, and paragraph 5.18.2.3 pp. 255 to 260) to the potential, describing a number of potential forms. Lombard (1985, p. 190) on the other hand describes only one constellation in one sentence: "The potential deficient verb form is $ka_{\text{MORPH\_pot}}$." He illustrates by example that this morpheme is inserted between the subject concord and the verb as in example (69), repeating example (23) for sake of convenience (cf. paragraph 2.9.4 on page 58).

(69) di_{CS10} ***ka***$_{\text{MORPH\_pot}}$ $fula_{\text{V\_itr}}$
`subj-3rd-cl10` **may** graze
'they may graze'

For the purpose of this study, we will briefly introduce the potential constellations according to Poulos and Louwrens (1994, pp. 229 to 234). However, their 'principal' is transliterated into Lombard's indicative and their participial into Lombard's situative. The potential is illustrated in the same way as we have described the previous main verbal constellations, dividing the verb into a preceding VIE and a VBP. Note that a specific positive future tense of the potential does not exist while the negated future tenses of the independent moods on the other hand, all make use of the negated potential morpheme, *ka se* 'will/shall/may not'

(cf. paragraphs 3.2.5.3, 3.2.6.3 and 3.2.7.3). Table 3.55 on page 173 shows a brief overview of some of the potential constellations. Other possible constellations, like, e.g. the forms described by Poulos and Louwrens (1994, pp. 255 to 260) might be added at a later stage.

3.6 Adverbial phrases (ADVP)

Adverbs may be added to a verbal phrase, as shown in the examples in (70). All such adverbs are adjuncts, i.e. grammatically not required. Instead they supply additional information about the predicate or modifying its meaning. Adverbs follow the basic verbal phrase. Note that nouns used as adverbs, like locative nouns or particle phrases may also extend all of the defined verbal phrases above.

70(a) *re*$_{\text{1CSPERS\_1sg}}$ *phela*$_{\text{V\_itr}}$ *gabotse*$_{\text{ADV}}$
`subj-1st-pl` live well
'we live well'

(b) *ba*$_{\text{1CS02}}$ *bofagane*$_{\text{V\_itr}}$ *molaong*$_{\text{N03\_loc}}$
`subj-3rd-cl2` married lawful
'they are lawfully married'

(c) *ba*$_{\text{1CS02}}$ *e*$_{\text{CO04/09}}$ *swere*$_{\text{V\_itr}}$ *ruri*$_{\text{N03\_loc}}$
`subj-3rd-cl2` `obj-3rd-cl4/9` holding(perfect) surely/certainly
'they were holding it firmly'

To describe adverbial phrases hence only few rules are necessary, e.g. ADVP $\rightarrow$ N_{categ} to allow nouns to appear as adverbs or VP $\rightarrow$ VP ADVP to extend the previously defined verbal phrases.

	^{pot}VP			
	^{pot}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
pot.pres.ind/sit.pos.	**2CS**$_{categ}$ ***ka*/subsMORPH_pot**		VBP	*–a*
	example: *(ge)* a_{2CS01} ka_{MORPH_pot} (when) `subj-3rd-cl1 pot` '(when) (s)he may speak'		$bolela_{V\_itr}$ speak	
pot.fut.pos	nonexistent			
pot.neg. 1 (ind.fut.neg) (sit.fut.neg)	**2CS**$_{categ}$ ***ka*/subsMORPH_pot *se***$_{MORPH\_neg}$		VBP	*–e*
	example: a_{2CS01} ka_{MORPH_pot} se_{MORPH_neg} `subj-3rd-cl1 pot neg` '(s)he might not speak'		$bolele_{V\_itr}$ speak	
pot.neg. 2	**2CS**$_{categ}$ ***ka*/subsMORPH_pot *se***$_{MORPH\_neg}$ ***ke***$_{V\_aux}$ **3CS**$_{categ}$		VBP	*–a*
	example: a_{2CS01} ka_{MORPH_pot} se_{MORPH_neg} ke_{V_aux} a_{3CS01} `subj-3rd-cl1 pot neg neg subj-3rd-cl1` '(s)he might not speak'		$bolela_{V\_itr}$ speak	

Table 3.55: The potential forms

3.7 Summary of the verbal phrases

In this chapter, an overview of forms of the VP of Northern Sotho was given from a computational perspective. A number of VPs representing moods according to Lombard (1985) have been described in terms of possible appearances and order of their elements, bound and free morphemes, based on the descriptions of several authors describing Northern Sotho grammar. Five types of verbal phrases have been defined according to their contents:

- Independent VPs consisting of a VBP, a basic verbal phrase, i.e. a verb stem, possibly complemented by one or more objects, these VPs have no noun class assigned because they appear without a subject;
- independent or dependent VPs with a noun class assigned, consisting of a VBP preceded by a verbal element, VIE, i.e. bound morphemes;
- copulative VPs consisting of multiword, fixed expressions, if their subject refers to the 3rd person, i.e. the noun classes, they have no noun class assigned;
- copulative VPs that always agree with their subject;
- auxiliary VPs, consisting of an auxiliary, with a noun class assigned, followed by a main verb or copulative VP with the same class assigned;
- hortative constellations requiring a subjunctive VP as their supplement;
- potential constellations that appear to share forms with the independent moods.

All these VPs can be followed by adjuncts which can either be adverbs or nominals that appear as adverbs[23].

VPs use nominal phrases as their arguments, i.e. as subject, objects, etc. The following chapter will give a brief overview of these phrases of Northern Sotho.

[23]Note that we will extent this definition in section 3.9 on particle phrases.

3.8 Constellations of the Noun Phrase (NP)

3.8.1 Introduction

The noun phrase of Northern Sotho shows a wide range of constellations, especially concerning its part of speech order, this section therefore only constitutes a basic attempt to describe all the possible forms a noun phrase might appear in. Our aim is rather to illustrate the most frequent constellations. Other cases will have to be added at a later stage.

The set of Northern Sotho word classes does not contain determiners. In the translations of examples in this study, determiners are therefore usually placed in brackets to demonstrate that for a correct translation from Northern Sotho to English, such determiners have to be inserted: *monna*$_{\text{N01}}$, '(a) man' or, if assumed that the entitiy *monna* has been introduced already in the discourse, '(the) man'. On the other hand, this non-existence of determiners implies that a noun phrase (NP) in Northern Sotho can consist solely of a noun (cf. example 71[24] (a)), while in other languages at least one qualifying element (determiner/quantifier or e.g. adjective) has to be present in most cases.

71(a) *nama*$_{\text{N09}}$
meat
'meat'

The pronouns of Northern Sotho (cf. paragraph 2.3), and the deictic demonstrative concords that have both a concordial and a pronominal character (cf. paragraph 2.4.5 and section 3.8.4.2), both co-occur with nouns and in this case show the character of (often deictic) determiners, as illustrated in 71 (b).

71(b) *nama*$_{\text{N09}}$ *ye*$_{\text{CDEM09}}$
meat `dem-3rd-cl9`
'this meat'

Some of these elements can occur either in front of the noun or follow it. As both pronoun and demonstrative concord have a pronominal character, they could also constitute a nominal phrase, as in 71 (c).

[24]Most examples in (71) are excerpts of examples taken from Ziervogel (1988, p. 124).

71(c) *ye*$_{\text{CDEM09}}$
`dem-3rd-cl9`
'this one'

The demonstrative concord can, together with an ADJ, form an adjective phrase (AP) (cf. paragraph 3.8.4.2) and such an AP modifies the noun semantically (cf. 71 (d)). The nominal character of the ADJ element of the adjective (cf. paragraph 2.5) allows to set an AP syntactically equal to an NP, as such it can also stand in place of an omitted noun to which it refers anaphorically as in 71 (e). However, the Northern Sotho word class ADJ only contains a few elements, the language therefore makes extensive use of possessives, cf. 71(f). Like the possessive of e.g. English, possessive noun phrases can be recursive, as in 71 (g).

71(d) *kgomo*$_{\text{N09}}$ *e*$_{\text{CDEM09}}$ *botse*$_{\text{ADJ09}}$
cow `dem-3rd-cl9` beautiful
'the beautiful cow'

(e) *e*$_{\text{CDEM09}}$ *botse*$_{\text{ADJ09}}$
`dem-3rd-cl9` beautiful
'the beautiful one'

(f) *meetse*$_{\text{N06}}$ *a*$_{\text{CPOSS06}}$ *boruthо*$_{\text{N14}}$
water of warmth
'warm water'

(g) *dikgomo*$_{\text{N10}}$ *tša*$_{\text{CPOSS10}}$ *mosadi*$_{\text{N01}}$ *wa*$_{\text{CPOSS01}}$ *kgoši*$_{\text{N09}}$
cattle of wife of king
'(the) king's wife's cattle'

To cater for the many possible ways of building a nominal phrase, we set up a slot system as in the definition of verbal phrases above. As different elements may be present, we describe the NP on three levels of which the first level contains the head of the phrase (slot zero, pos 0) and reserves two more positions for demonstratives which may occur in a preceding or a following position (slot zero, pos-1 and pos+1). The second level adds two more slots containing adjectival modifiers.

In terms of filling the fields, slot pos-0 is reserved solely for the word class 'noun' (cf. section 2.2 beginning on page 22), while other positions of slot zero can contain pronouns (section 2.3) and demonstrative concords (paragraph 2.4.5), but not nouns. In the case that slot zero, pos-0 is empty, i.e. when the noun is omitted, the rightmost pronoun/demonstrative

acquires the pronominal status (while the others remain in their status as determiners). Such a decision is however abritrary and might have to be reconsidered on further examination of text collections.

Adjectives and possessives may be then added on the second level. These usually follow the noun (and its pronouns) and are therefore placed in slots zero+1 and zero+2. However, if the first and the second level are both empty, one of these will acquire its role, cf. the slot definition in Table 3.56 and the illustrating trees in Figures 72 (a) to (f) where the head of the phrase assumed is written in bold face. If both of the righthandside slots, zero+1 and zero+2 are filled, i.e. if an adjectival phrase and a successive possessive nominal phrase appear while slot zero is empty, we consider the adjectival phrase as the head of the phrase (72). Again, this is a rather arbitrary decision, however based on the claim of some Northern Sotho linguistists (inter alia (Van Wyk et al., 1992, p. 73), see also paragraph 2.5 on page 45) that a word class 'adjective' does not exist in Northern Sotho: what is called adjective here, is rather an "adjectival noun".

In summary, we assume that nominal phrases of Norther Sotho on the first level are basically left-headed, except in the semantically contrastive cases, where a demonstrative concord precedes the noun. This consequently means that if an NP consists of an adjectival phrase followed by a possessive phrase, it would be headed by the adjectival (or adjectival noun) phrase. Note again, this decision is arbitrary and it might be necessary to revisit it in future. In summary we currently define the following hierarchy of heads of a noun phrase: noun ⟩ leftmost element of slot zero ⟩ slot zero+1 ⟩ slot zero+2.

Table 3.56: Slots describing the noun phrase

	slot zero			
pos-1	**pos-0**	**pos+1**	**zero+1**	**zero+2**
^{pro}NP	noun	^{pro}NP	AP	^{pos}NP

3.8.2 An overview of some nominal phrases

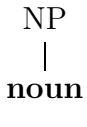
In the following tables we will fill the pre-defined slots of Table 3.56 with appropriate elements, beginning with the simplest noun phrase containing a noun only to its modified forms, containing pronouns, concords, and other phrases (which will be described in detail in paragraphs 3.8.3, 3.8.5 and 3.8.4.2). Table 3.57 shows the possible constellations filling

slot zero, i.e. the basic noun phrase. Table 3.58 then describes how this basic noun phrase may be extended with other elements, e.g. APs or $^{\text{pos}}$NPs.

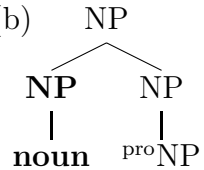
Table 3.57: The basic noun phrase

description		**slot zero**	
	pos-1	**pos-0**	**pos+1**
	$^{\text{pro}}$NP	noun	$^{\text{pro}}$NP
NP_{categ}		$\mathbf{N}_{\text{categ}}$	
example: NP_{10}		*dimpša*$_{\text{N10}}$ ‘dogs’	
NP_{categ}	$^{\text{pro}}\mathbf{NP}_{\text{categ}}$	$\mathbf{N}_{\text{categ}}$	
example: NP_{10}	*tše*$_{\text{CDEM10}}$ ‘these’	*dimpša*$_{\text{N10}}$ ‘dogs’	
	‘these (specific) dogs’		
NP_{categ}		$\mathbf{N}_{\text{categ}}$	$^{\text{pro}}\mathbf{NP}_{\text{categ}}$
example: NP_{10}		*dimpša*$_{\text{N10}}$ ‘dogs’	*tše*$_{\text{CDEM10}}$ ‘these’
		‘these dogs’	

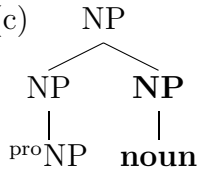
Figure 72 (a)



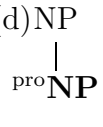
(b)



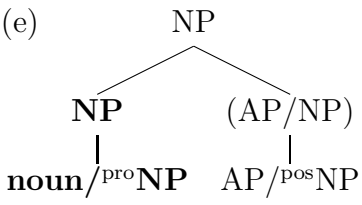
(c)



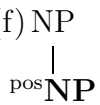
(d)



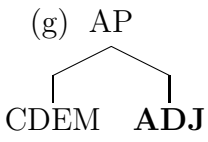
(e)



(f)



(g)



(h)

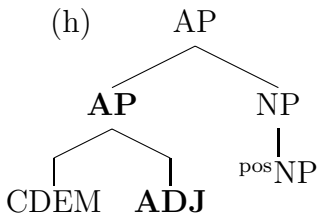


Table 3.58: The extended noun phrase

description	**slot zero**	**zero+1**	**zero+2**
NP_{categ}	$\mathbf{NP}_{\text{categ}}$	$\mathbf{AP}_{\text{categ}}$	
example:			
NP_{10}	*dimpša*$_{\text{N10}}$	*tše*$_{\text{CDEM10}}$ *mpe*$_{\text{ADJ10}}$	
	'dogs'	'bad'	
		'bad dogs'	
NP_{categ}	$\mathbf{NP}_{\text{categ}}$		$^{\text{pos}}\mathbf{NP}_{\text{categ}}$
example:			
NP_{10}	*dimpša*$_{\text{N10}}$		*tša*$_{\text{CPOSS10}}$ *gagwe*$_{\text{PROPOSS01}}$
	'dogs'		'of him'
		'his dogs'	
NP_{categ}	$\mathbf{NP}_{\text{categ}}$	$\mathbf{AP}_{\text{categ}}$	$^{\text{pos}}\mathbf{NP}_{\text{categ}}$
example:			
NP_{10}	*dimpša*$_{\text{N10}}$	*tše*$_{\text{CDEM10}}$ *mpe*$_{\text{ADJ10}}$	*tša*$_{\text{CPOSS10}}$ *gagwe*$_{\text{PROPOSS01}}$
	'dogs'	'bad'	'of him'
		'his bad dogs'	

3.8.3 The Pronominal Noun Phrase (^{pro}NP)

The ^{pro}NP can generally be described as containing emphatic (cf. paragraph 2.3.1) and/or quantitative (cf. paragraph 2.3.3) pronouns. If it appears together with the basic noun phrase, it is usually interpreted in the sense of an emphasis. Specifically the absolute/emphatic pronoun usually appears on the right side of the noun (slot +1), there it emphasises a contrast (cf. 73 (a) and (b)), while, when appearing on the left side (slot -1) it rather shows a specification, as in (73 (c)), compare Poulos and Louwrens (1994, p. 75 et seq.).

73(a) *mošemane*$_{N01}$ ***yena***$_{\text{PROEMP01}}$ *o*$_{\text{1CS01}}$ *rata*$_{\text{V\_tr}}$ *diapola*$_{\text{N10}}$
boy `emp-3rd-cl1` `subj-3rd-cl1` like apples
'as for (the) boy, he likes apples'

basetsana$_{\text{N02}}$ ***bona***$_{\text{PROEMP02}}$ *ba*$_{\text{1CS02}}$ *rata*$_{\text{V\_tr}}$ *dinamune*$_{\text{N10}}$
girls `emp-3rd-cl2` **`subj-3rd-cl2`** like oranges
'(but) as for (the) girls, they like oranges'

(b) *dikgomo*$_{\text{N10}}$ *di*$_{\text{1CS10}}$ *šetše*$_{\text{V\_aux}}$ *di*$_{\text{2CS10}}$ *gorogile*$_{\text{V\_itr}}$,
cattle `subj-3rd-cl10` already `subj-3rd-cl10` arrived,
'(the) cattle have already arrived,'

fela$_{\text{CONJ}}$ *poo*$_{\text{N09}}$ ***yona***$_{\text{PROEMP09}}$ *ga*$_{\text{MORPH\_neg}}$ *se*$_{\text{MORPH\_neg}}$ *ya*$_{\text{3CS09}}$
but bull `emp-3rd-cl9` `neg` `subj-3rd-cl09`
boa$_{\text{V\_itr}}$
return
'but the bull, on the contrary, has not returned'

(c) *lefasetere*$_{\text{N05}}$ *le*$_{\text{CDEM05}}$ *le*$_{\text{CS05}}$ *thubilwe*$_{\text{V\_tr}}$
window `dem-3rd-cl05` `subj-3rd-cl10` broken
'this window was broken'

ke$_{\text{PART\_agen}}$ ***yena***$_{\text{PROEMP01}}$ *Phetla*$_{\text{N01a}}$
by `emp-3rd-cl1` Phetla
'by Phetla and no one else'

Table 3.59 extends the definitions of Table 3.57 by adding the ^{pro}NP as another independent basic noun phrase.

Table 3.59: The basic noun phrase including ^{pro}NPs

description		**slot zero**	
	pos-1	**pos-0**	**pos+1**
	^{pro}NP	noun	^{pro}NP
NP_{categ}		$\mathbf{N}_{categ}$	
example: NP_{10}		*dimpša*$_{N10}$ ‘dogs’	
NP_{categ}	$^{pro}\mathbf{NP}_{categ}$	$\mathbf{N}_{categ}$	
example: NP_{10}	*tše*$_{CDEM10}$ `dem-3rd-cl10`	*dimpša*$_{N10}$ dogs ‘these (specific) dogs’	
NP_{categ}		$\mathbf{N}_{categ}$	$^{pro}\mathbf{NP}_{categ}$
example: NP_{10}		*dimpša*$_{N10}$ dogs ‘these dogs’	*tše*$_{CDEM10}$ `dem-3rd-cl10`
NP_{categ}			$^{pro}\mathbf{NP}_{categ}$
examples: NP_{10}			*tše*$_{CDEM10}$ `dem-3rd-cl10` ‘these’
NP_{01}			*yena*$_{PROEMP01}$ `emp-3rd-cl1` ‘the one’
NP_{02}			*bohle*$_{PROQUANT02}$ `quant-3rd-cl2` ‘all (of them)’

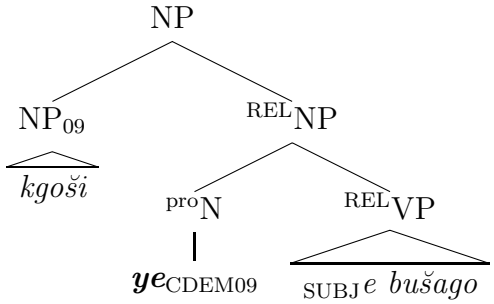
3.8.4 Nominal phrases headed by demonstrative concords

3.8.4.1 The relative Noun Phrase (^{rel}NP)

A demonstrative concord may appear in several pronominal functions. It may not only appear as the pronominal head of a noun phrase, as described in paragraph 3.8.3 (page 180), but also heading relative clauses (paragraph 3.2.7). For the sake of demonstration,

example 54 (a) is repeated as 71 (g) .

Figure 71(g)



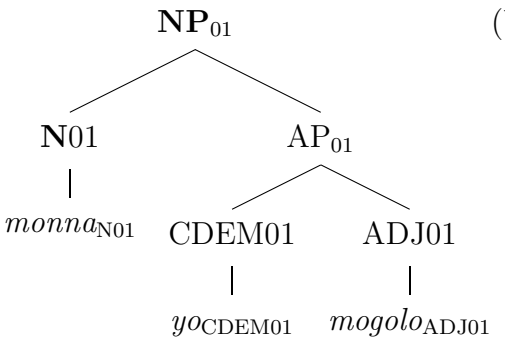
3.8.4.2 The Adjectival Phrase (AP)

As mentioned earlier (e.g. in paragraph 2.5), the Northern Sotho adjective is formed by a demonstrative concord (in its concordial function) preceding an ADJ, which has adjectival, but also nominal characteristics. Such properties of this constituent allow it to replace an NP altogether, cf. examples and Figures 74 (a) and (b). Consequently, all previously described NPs might also be represented by an AP, despite the descriptive copulative, which we explicitely defined as containing only nouns (and their respective pronouns) as complements (cf. paragraph 3.3.3).

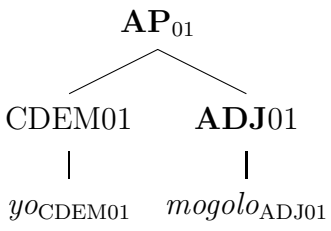
74(a) *monna*$_{N01}$ *yo*$_{\text{CDEM01}}$ *mogolo*$_{\text{ADJ}}$
man `dem-3rd-cl1` big
'(a) big man'

(b) *yo*$_{\text{CDEM01}}$ *mogolo*$_{\text{ADJ01}}$
`dem-3rd-cl1` big
'(a) big one'

74(a)



(b)



3.8.5 The Possessive Noun Phrase (^{pos}NP)

One general property of a phrase is that it may be recursive, as shown for the possessive NP in 71 (h) (demonstrating the recursion) with the sentence *dikgomo tša wa kgoši*, 'cattle of (the) chief'. As described in paragraph 2.3.2, such a constellation is preceded by a noun indicating possession (which may be omitted in certain discourses). The following possessive noun phrase ^{pos}NP then contains a possessive concord and another noun, the possessor. According to Van Wyk et al. (1992, p. 64), this possessor might be replaced by a possessive pronoun (PROPOSS$_{categ}$, cf. paragraph 2.3.2). Table 3.60 describes the possible constellations, the last element of each of the constellations can either be a noun, a possessive pronoun or another proNP.

Figure 71 (h)

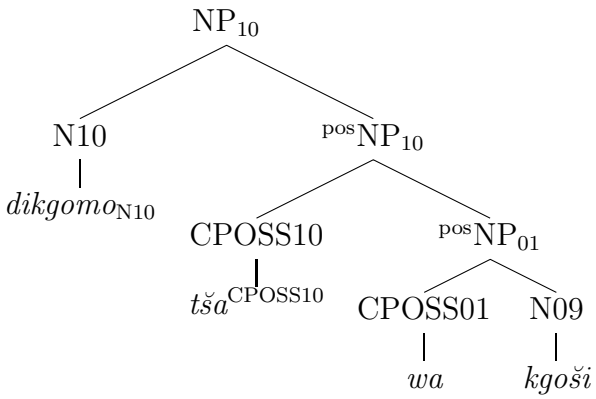


Table 3.60: The possessive noun phrase $^{\text{pos}}$NP

description	$^{\text{pro}}$**NP**		
$^{\text{pro}}\text{NP}_{\text{categ}}$		$\text{CPOSS}_{\text{categ}}$	N / PROPOSS / $^{\text{pro}}$NP
examples:			
$^{\text{pro}}\text{NP}_{10}$	(*dikgomo*$_{\text{N10}}$) (the cattle)	*tša*$_{\text{CPOSS10}}$	*mosadi*$_{\text{N01}}$ of (the) woman
	(*dikgomo*$_{\text{N10}}$) (the cattle)	*tša*$_{\text{CPOSS10}}$	*gagwe*$_{\text{PROPOSS01}}$ of him/her
	(*dikgomo*$_{\text{N10}}$) (the cattle)	*tša*$_{\text{CPOSS10}}$	*mosadi*$_{\text{N01}}$ *wa*$_{\text{CPOSS01}}$ *kgoši*$_{\text{N09}}$ of (the) wife of the king
		tša$_{\text{CPOSS10}}$	*mosadi*$_{\text{N01}}$ the ones of (the) woman
		tša$_{\text{CPOSS10}}$	*gagwe*$_{\text{PROPOSS01}}$ the ones of him/her
		tša$_{\text{CPOSS10}}$	*mosadi*$_{\text{N01}}$ *wa*$_{\text{CPOSS01}}$ *kgoši*$_{\text{N09}}$ the ones of (the) wife of the king

3.9 The Particle Phrase (PP)

Particle phrases are constellations consisting of a particle followed by its argument, usually a noun. These phrases appear as adjuncts to verbs, following the VP as illustrated in example (75) from (De Schryver, 2007, p. 81), and Figure 3.47, demonstrating its possible analysis.

(75) *monna*$_{N01}$ *o*$_{1CS01}$ *rema*$_{V\_tr}$ *mohlare*$_{N03}$ ***ka***$_{PART\_ins}$ ***selepe***$_{N07}$
man `subj-3rd-cl1` chop tree with axe
'(a) man is chopping the tree with (an) axe'

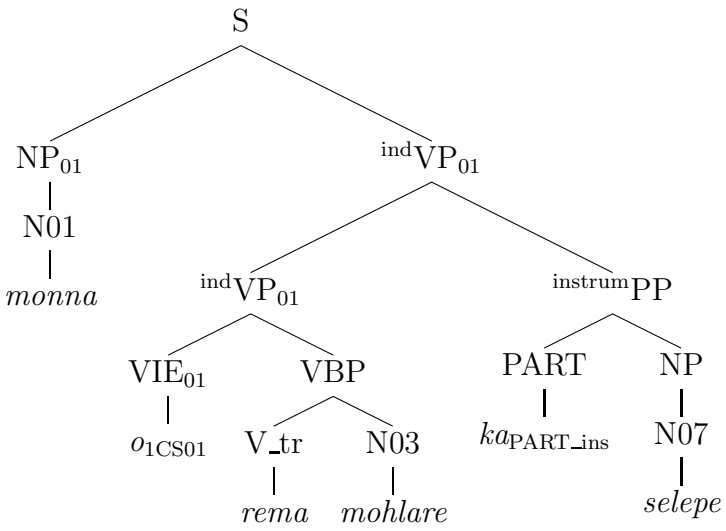


Figure 3.47: Example of a particle phrase (PP)

3.10 A sentence of Northern Sotho

3.10.1 The basic proposition

In section 3.2, we defined a sentence of Northern Sotho as possibly consisting of the verbal phrase alone, because the subject noun / nominal may be omitted in a known discourse, like in 76 (a). In other words, the VP of Northern Sotho alone in many cases already constitutes a sentence, hence this paragraph only adds to the previously described constellations.

Unless the subject noun / nominal is omitted, it usually appears in an initial position, cf. 76 (b), which constitutes the basic Northern Sotho propositional sentence.

76(a) O_{1CS01} $lapile_{\text{V\_itr}}$
`subj-3rd-cl1` (is tired)
'(s)he is tired'

(b) $Monna_{N01}$ o_{1CS01} $a_{\text{MORPH\_pres}}$ $boa_{\text{V\_itr}}$
man `subj-3rd-cl1` `pres` return
'(the) man returns'

Sentences may also be connected with conjunctions, like $gomme_{\text{CONJ}}$, as illustrated by 76(c). A possible analysis of such a constellation is shown in Figure 3.48.

76(c) $monna_{N01}$ o_{1CS01} $boa_{\text{V\_itr}}$ $gomme_{\text{CONJ}}$ o_{1CS01} $lapile_{\text{V\_itr}}$
man `subj-3rd-cl1` return and `subj-3rd-cl1` is tired
'(the) man returns and he is tired'

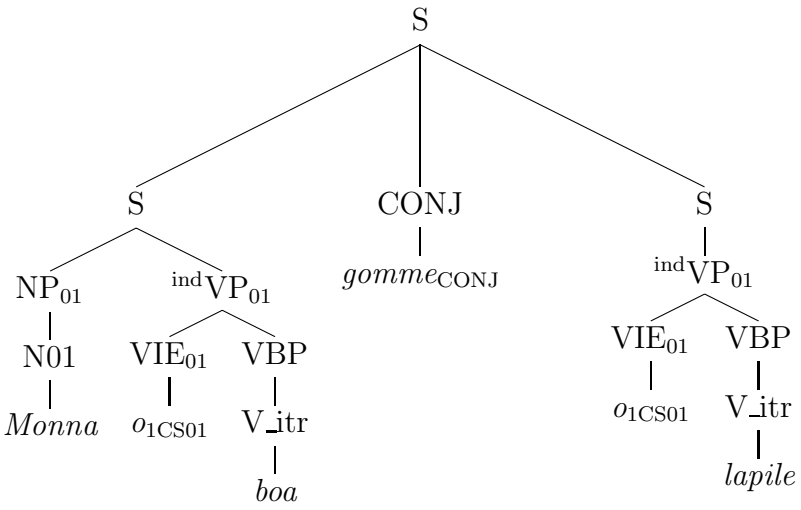


Figure 3.48: An example analysis of two sentences connected with a conjunction

3.10.2 The question

A sentence of Northern Sotho can easily be formulated as a question (interrogative or $^{\text{int}}$S) by adding one of the question particles, as described in paragraph 2.10.7, cf. example 76 (d), and Figure 3.49. The question mark is not mandatory, but usually appears.

76(d)	*na*$_{\text{PART\_que}}$	*monna*$_{N01}$	*o*$_{1\text{CS01}}$	*boile*$_{\text{V\_itr}}$*?*
	`que`	man	`subj-3rd-cl1`	return-`past`?
	'did (the) man return?'			

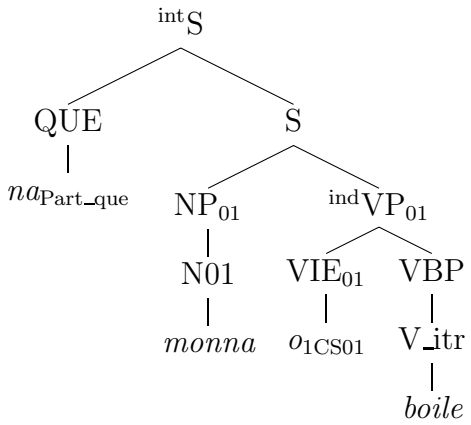


Figure 3.49: An example analysis of a sentence containing a question

Another way to formulate a question is to replace a noun with a question word (in English, usually the "wh question words" are used), as described in paragraph 2.10.9. Lombard (1985) does not describe such sentences explicitely (he classifies question words as adverbs), whereas Poulos and Louwrens (1994, p. 376) list a number of examples containing the question word *eng* 'what', cf. example (76), and Figure 3.50 respectively.

76(e)	*o*$_{1\text{CS01}}$	*bona*$_{\text{V\_tr}}$	*eng*$_{\text{QUE\_N09}}$*?*
	`subj-2nd-sg`	see	what?
	'what do you see?'		

3.11 A brief summary of our grammar fragment

We had defined our aim as describing a significant fragment of the grammar of Northern Sotho in this chapter. As the majority of possible constellations consist of or contain a verbal phrase VP, this phrase dominated our descriptions. We described main and copulative

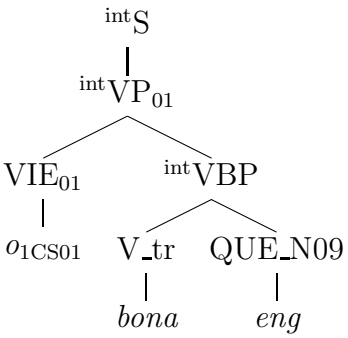


Figure 3.50: A Northern Sotho question making use of the "wh question word" *eng*

verbs in independent and dependent clauses by mainly following the definitions of Lombard (1985). As nouns and or other nominals forming noun phrases may appear as overt subjects and objects, but also as adjuncts of verbs, a substantial part of this chapter was dedicated to describing at least some of the possible constellations forming a NP. Other descriptions, such as section 3.9 concerning particle phrases, were kept brief as they resemble e.g. English prepositional phrases which are well-defined in literature.

Having defined a number of rules forming phrases and how they combine to form sentences, we will in the next chapters describe the possibilities and challenges of their implemention.

Chapter 4

Features of verbal phrases

4.1 Introduction

A significant sub-task in the preparation of parsing is that of gathering information which helps resolve ambiguities. It has been argued in paragraph 2.1 (on page 20) that distinguishing word classes according to their distribution supports the development of morphosyntactic rules leading to a lower number of resulting analyses than - for example - distinguishing them according to their morphological characteristics. Moreover, it was shown in paragraph 3.2.1.6 that labelling the transitivity of verbs - for example by defining labels of previously underspecified verb stem properties in greater detail - supports the identification of unmarked sentence borders. If lexical items are however intrinsically ambiguous , i.e. if they appear in a number of word classes, it will be necessary to find their distributional patterns on the basis of their context to support disambiguation on a lexical level. For example, Northern Sotho makes use of quite a number of highly ambiguous function words, of which e.g. *a* may either represent a subject concord of noun class 1, set 2 and 3 (2CS01, 3CS01) or of noun class 6 (all three sets: 1CS06, 2CS06, 3CS06), or a demonstrative concord (CDEM06), or a possessive concord (CPOSS06), both of noun class 6, or a hortative particle (PART_hort) etc. Even without taking the different subject concord sets into consideration, *a* already has a ninefold ambiguity, as Table 4.1 (taken from Faaß et al. (2009)) demonstrates. According to the definitions of this study, this token could be interpretated in 12 ways. In order to support identification of the correct POS of *a*, contextual partners must therefore be considered.

Any work on defining patterns in the co-occurrence of parts of speech could also assist in developing generalisations on the path to a more general linguistic modelling of Northern Sotho verbs than is undertaken in this study. This chapter could therefore form the basis

	Description	Example
1	subject concord of nominal cl. 1	*ge monna* ***a*** *fihla* conjunctive + noun cl. 1 + **subject concord cl. 1** + verb stem if/when + man + **subj-cl1** + arrive 'when the man arrives'
2	subject concord of nominal cl. 6	*masogana* ***a*** *thuša basadi* noun cl. 6 + **subject concord cl. 6** + verb stem + noun cl.2 young men + **subj-cl6** + help women 'the young men help the women'
3	object concord of	*moruti o* ***a*** *biditše* noun cl. 1 + subject concord cl. 1 + **object concord cl. 6** + verb stem teacher + subj-cl1 + **obj-cl6** + called 'the teacher called them'
4	possessive concord of nominal cl. 6	*maoto* ***a*** *gagwe* noun cl. 6 + **possessive concord cl. 6** + possessive pronoun cl. 1 feet + **of** + his 'his feet'
5	demonstrative concord of nominal cl. 6	*ba nyaka masogana* ***a*** subject concord cl. 2 + verb stem + noun cl. 6 + **demonstrative concord** they + look for + young men + **these** 'they are looking for these young men'
6	present tense morpheme	*morutiši o* ***a*** *bitša* noun cl. 1 + subject concord cl.1 + **present tense marker** + verb stem teacher + subj-cl1 + **pres** + call 'the teacher is calling'
7	past tense morpheme	*morutiši ga o* ***a*** *bitša masogana* noun cl. 1 + negation morpheme + subject concord cl.1 + **past tense marker** + verb stem + noun cl. 6 teacher + neg + subj-cl1 + **past** + call + young men 'the teacher did not call the young men'
8	hortative particle	***a*** *ba tsene* **hortative particle** + subject concord cl. 2 + verb stem let + subj-cl2 + come in 'let them come in'
9	interrogative particle	***a*** *o tseba Sepedi* **interrogative particle** + subject concord 2nd pers sg. + verb stem + noun cl. 7 **ques** + subj-2nd-pers-sg + know + Sepedi 'do you know Sepedi'

Table 4.1: The polysemy of *a* (taken from Faaß et al. (2009))

for future work, where other parsing strategies will come into use, concentrating on clusters of parts of speech found in text rather than using pre-defined VP-rules. In this chapter, Northern Sotho verbal constellations are summarised according to some of their features detectable from a computational perspective and we will attempt to provide a first set of generalisations for them. However first it is necessary to look at how a parser works and how the knowledge outlined above can be utilized.

4.2 Parsers: approaches to describe natural languages

Parsers, i.e. operational grammars, are capable of comparing a given text (usually a sentence) with pre-defined grammatical and lexical knowledge of a system, i.e. about the units and how they combine. "Rule-based" parsers contain a lexicon and a set of rules (e.g. context free grammars (CFG) as described in paragraph 1.4.4). Other parsers do not contain a separate rules-section, as the units of language in those systems (lexical elements and the constituents they form, i.e. phrases) are described using the same schemata. Such parsers are called "lexicon-based" (Hellwig, 1989, p. 422).

In principle, parsers match specific sets of properties allocated to pre-defined units of a language with flowing text provided as their input. During the first processing step, this flowing text is usually separated into units (tokens). As described in paragraph 1.4.2.1, such separation is no trivial task, however solvable with tokenization tools. The units identified in the sentence and the order in which they appear are compared to the system's knowledge on legal constellations. If there is a match, the parser will accept the sentence. Nowadays, most parsers will not only inform the user whether the sentence was accepted or not but also assign one or more 'well-formed' - i.e. correct - analyses (parses). Some parsers work shallowly, their analysis resulting in flat, non-recursive 'chunks' rather than phrases or sentences, like, e.g. noun phrases, these are often called shallow parsers or 'chunkers'. In a rule-based parser, grammatical knowledge is represented as constraints, and as long the constituent in question - chunk or sentence - conforms to these constraints, it is considered to be well-formed (Sag et al., 2003, p. 83).

The system's knowledge utilised by the parser often provides these constraints as sets of properties containing attribute-value pairs. Analyses result from certain principles, including the principle of uniqueness (cf. paragraph 5.1.2.2). For example, the English verb form 'sleeps' is often stored in parser lexicons with two attribute-value pairs related to the

expected subject, 'person'='third' and 'number'='singular'. Any noun that could possibly be the verb's subject (i.e. a noun that appears in a position where the grammar rule expects the subject to appear) will only be accepted as such by the parser if it is not stored with information contradicting these constraints. A phrase like *'pupils sleeps' will not be authorised because of such contradicting constraints of the units, as the noun 'pupils' is usually stored with the attribute-value pair 'number'='plural' and the uniqueness principle only allows for one value to appear with each attribute. In other words, the uniqueness principle proihibits units contributing to a structure (e.g. a phrase) from providing contradicting values for one attribute. Therefore, some constraint-based grammars are also called unification-based grammars.

If words or a sentence may be interpreted in several ways, they are ambiguous, lexically and/or structurally. Humans can disambiguate a number of such ambiguities using their world knowledge, so while humans consider the noun phrase 'mothers and children under 13' (cf. example (77)) to only have one reading, a parser will usually assign three analyses, as the mothers, the children or the mothers and the children may – from a syntactic perspective – all be under 13 ("pp-attachment", cf. e.g. Schütze (1995)).

(77)

bomme$_{\text{N02b}}$	*le*$_{\text{PART\_con}}$	*bana*$_{\text{N02}}$	*ba*$_{\text{CDEM02}}$	*ba*$_{\text{CS02}}$	*lego*$_{\text{VCOP}}$	*fase*$_{\text{NLOC}}$
mothers	`con`	children	`dem-3rd-cl2`	`subj-3rd-cl2`	who are	under
ga$_{\text{CPOSSLOC}}$	*mengwaga*$_{\text{N04}}$	*ye*$_{\text{CDEM04}}$	*13*$_{\text{NUM}}$			
of	years	`dem-3rd-cl4`	13			

'mothers and children under 13'

However, even where world knowledge, as in (77) is not necessary for the disambiguation of a sentence, tools analysing human language are often still confronted with lexical and structural ambiguities on all representation levels. Applying generalisations on features of constellations, i.e. signalling elements or signalling element clusters may help to reduce such ambiguities. Considering Northern Sotho, a typical example is the situative $^{\text{SIT}}$VP (cf. paragraph 3.2.6) which in a number of cases is identical on the surface with an indicative $^{\text{IND}}$VP (cf. paragraph 3.2.5). Such phrases are therefore ambiguous, however, as soon as the context is taken into account, their ambiguity can often be resolved; in the case of the situative, it is often preceded by the conjunction *ge* -'when' - which never occurs with an indicative. A feature catalogue of the Northern Sotho phrases, i.e. a **data category inventory** should support an effective reduction of ambiguities; some basics will be described in section 4.4.

For an overview of the many possible parsing strategies and algorithms, see e.g. (Hellwig, 1989). This section will solely describe the parsing algorithm (a right-corner parser) utilized in the example analysis shown in section 4.3, paragraphs 4.2.1 to 4.2.1.2 briefly explain the necessary terms.

4.2.1 Vertical parsing directions

There are top-down and bottom-up parsers. Following a top-down strategy, a set of all sentences that are possible according to the knowledge of the system (i.e. lexicon and rules) is generated first. The sentence in question, i.e. the sentence to be processed is then examined in order to find out if it is a member of the given set. A bottom-up strategy on the other hand begins with the surface sentence and combines the units of the sentence by utilizing the knowledge step by step until the highest possible node, the S-node is found (cf. paragraph 1.4.4 on page 13). If a system contains few rules and a small lexicon, a top-down strategy can be sufficient. However, in the case of Northern Sotho, where quite a few morphosyntactic rules have been defined (cf. chapter 3), we have chosen a bottom-up strategy to avoid over-generation[1].

4.2.1.1 Horizontal parsing directions

Depending on the language in question, some parsers process sentences from left to right, others from right to left. Here, it is not relevant how a human would read the sentence, rather how the number of (intermediate) illicit analyses can be kept low while processing the sentence. This issue would constitute a research question by itself for parsing Northern Sotho, exceeding the scope of this study. We have opted rather arbitrary to try a right-to-left analysis in section 4.3.

4.2.1.2 Right-corner parsing

Applications of parsing algorithms often combine both processing strategies - bottom-up and top-down - like, e.g. the left-corner parser which begins with the leftmost element of the sentence and, while constantly identifying lexicon entries fitting to the surface forms it meets, it “builds sentence structure in a left-to-right, bottom-up fashion, piecing together

[1]Over-generation describes a situation where illicit analyses are generated by a system. These have to be filtered out, either manually or by additional algorithms.

the left corner of a structural description first" (Petrick, 1989, p. 690). Our sample analysis proceeds similar to this parser, hence we call it 'right-corner parser'.

4.3 A sample analysis

For the sake of convenience, we repeat example (47) of page 105 (cf. Lombard (1985, p. 147)) as (78) here in order to illustrate the reduction of lexical and structural ambiguities step by step.

(78)

ge	*ba*	*bona*	*noga*	*ba*	*a*	*e*	*bolaya*
when	`subj-3rd-cl2`	see	$\text{snake}_{\text{N09}}$	`subj-3rd-cl2`	`pres`	`obj-3rd-cl9`	kill

'when they see a snake, they kill it'

The parser is assumed to know about the rules defined in chapter 3. The assumed system's lexicon contains the possible parts of speech for all of the contained tokens. Note that ':' in the following list indicates POS-ambiguity, e.g. *bona* may be an emphatic pronoun of class 2 (PROEMP02) or a possessive pronoun of the same class (PROPOSS02), or a transitive verb (V_tr, meaning '[to] see').

- *ge*
  CONJ
- *ba*
  1CS02:2CS02:3CS02:CDEM02:CO02:CPOSS02:V_AUX:VCOP
- *bona*
  PROEMP02:PROPOSS02:V_tr
- *noga*
  N09
- *ba*
  1CS02:2CS02:3CS02:CDEM02:CO02:CPOSS02:V_AUX:VCOP
- *a*
  1CS06:2CS06:3CS06:2CS01:3CS01:CDEM06:CO06:CPOSS06:
  MORPH_past:MORPH_pres:PART_hort:PART_que
- *e*
  1CS04:2CS04:1CS09:2CS09:CO04:CO09:CSNEUT
- *bolaya*
  V_tr

Our example analysis[2] begins with the rightmost element, which is $bolaya_{V\_tr}$ 'kill'. The parser first attempts to identify rules ending in this POS as hypothetical analyses. As none are found (no phrase consists of a transitive verb alone), the second element, *e*, is considered. In this study (cf. paragraph 3.2.1.1), a basic verbal phrase, VBP was defined as the smallest element of any verbal phrase containing a verbal stem and possibly its object(s). The token *e* which is highly ambiguous (1CS04:2CS04:1CS09:2CS09:CO04:CO09:CSNEUT) will therefore be assumed to be an object concord of either class 4 or class 9 (CO04:CO09), i.e the first (lexical) ambiguity is partially resolved, as shown in Figure 4.1.

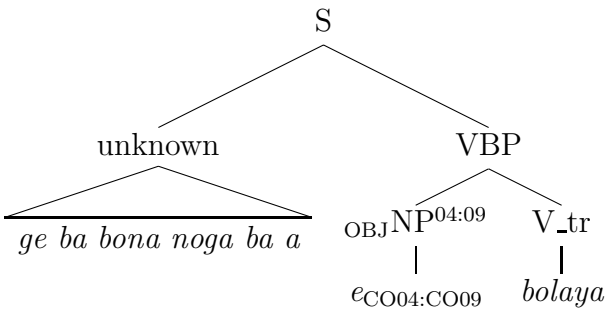


Figure 4.1: A partial analysis of (*ge ba bona noga ba a*) *e bolaya* '`obj-3rd-cl4:9` kill'

A number of possible analyses remain, as Table 4.2 shows. All verbal phrases that contain VBPs defined in chapter 3 containing a verb stem ending in *a* are considered.

The next element, *a*, is highly ambiguous; the set of possible parts of speech contains twelve elements. However, the particles may both be excluded from the set as PART_hort always precedes a subject concord, and we have already identified *e* as an object concord and PART_que ordinarily only appears as either the first or the last element of the sentence (the parser is informed about the total length, i.e. the number of tokens contained in the sentence). Due to the grammar rules describing NPs, the parser is aware that the possessives CPOSS01:CPOSS06 only occur in noun phrases where they precede nouns or possessive pronouns. If we look at the knowledge gained so far, a demonstrative concord CDEM06 could only appear here preceding a subject concord of class 6, which is not the case. This reading can therefore also be excluded from the set . The element *e* has already been identified as an object concord, therefore *a* cannot be CO06 (Northern Sotho only allows for one object concord to occur in the verbal phrase). Only seven possibilities

[2]Note that the following explanations are fairly abstract and thus simplified.

Table 4.2: The remaining set of possible analyses when considering the VBP ending in *a*

	$^{\text{IND}}$VP			
	VIE		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**zero +1**
pres.pos.long	$1\text{CS}_{\text{categ}}$	MORPH_pres	VBP^{p}	\$.
perf.neg. 1	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$		VBP	
perf.neg. 3	$ga_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$		VBP	
perf.neg. 4	$ga_{\text{MORPH\_neg}}$ $1\text{CS}_{\text{categ}}$ a		VBP	
fut.pos	$1\text{CS}_{\text{categ}}$	tlo/tla MORPH_fut	VBP	
	$^{\text{SIT}}$VP			
pres.pos.	$2\text{CS}_{\text{categ}}$		VBP	
perf.neg.1	$2\text{CS}_{\text{categ}}$ $se_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$		VBP	
perf.neg.2	$2\text{CS}_{\text{categ}}$ $se_{\text{MORPH\_neg}}$ $1\text{CS}_{\text{categ}}$		VBP	
perf.neg.3	$2\text{CS}_{\text{categ}}$ $sa_{\text{MORPH\_neg}}$		VBP	
fut.pos.	$2\text{CS}_{\text{categ}}$	tlo/tla MORPH_fut	VBP	
	$^{\text{REL}}$VP			
perf.neg.1	$2\text{CS}_{\text{categ}}$ $sego/seng$ MORPH_neg $3\text{CS}_{\text{categ}}$		VBP	
fut.pos.1	$2\text{CS}_{\text{categ}}$	$tlago/tlogo$ MORPH_fut	VBP	
fut.neg.1	$2\text{CS}_{\text{categ}}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	$tlago/tlogo$ MORPH_fut	VBP	

Table 4.3: The remaining set of possible analyses when considering *a*

	$^{\text{PRES}}$VP			
	VIE		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**zero +1**
pres.pos.long	$1CS_{categ}$	MORPH_pres	VBP^{p}	$.
perf.neg. 1	ga_{MORPH_neg} se_{MORPH_neg} $3CS_{categ}$		VBP	
perf.neg. 3	ga_{MORPH_neg} $3CS_{categ}$		VBP	
perf.neg. 4	ga_{MORPH_neg} $1CS_{categ}$ MORPH_past		VBP	
	$^{\text{SIT}}$VP			
pres.pos.	$2CS_{categ}$		VBP	
perf.neg.1	$2CS_{categ}$ se_{MORPH_neg} $3CS_{categ}$		VBP	
perf.neg.2	$2CS_{categ}$ se_{MORPH_neg} $1CS_{categ}$		VBP	
	$^{\text{REL}}$VP			
perf.neg.1	$2CS_{categ}$ *sego/seng* $MORPH\_neg$ $3CS_{categ}$		VBP	

therefore remain (2/3CS01:1/2/3CS06:MORPH_past/pres) as illustrated in Figure 4.2. By accepting that the constellation either contains a subject concord or a tense morpheme, all respective constellations can be excluded from the set of hypothetical analyses. The remaining set of possible analyses however still contains all predicative VPs (except for copulative constellations which can be excluded, because a main verb is contained), cf. Table 4.3 and Figure 4.2.

When examining the next preceding token, *ba*, some of its possible parts of speech may again be excluded on the basis of the information collected so far. In the given sentence, we can exclude the object concord reading, as *e* has already been successfully identified as an object concord which only appears alone in Northern Sotho verbs, secondly, these concords must appear directly in front of the verb stem. If *a* should however represent a tense morpheme, *ba* may be a subject concord of class 2 (1CS02:2CS02:3CS02) and as such it may function as the subject of the clause. The ambiguity remaining for the following element, *a*, can now be mostly resolved as, if preceded by a subject concord it can only

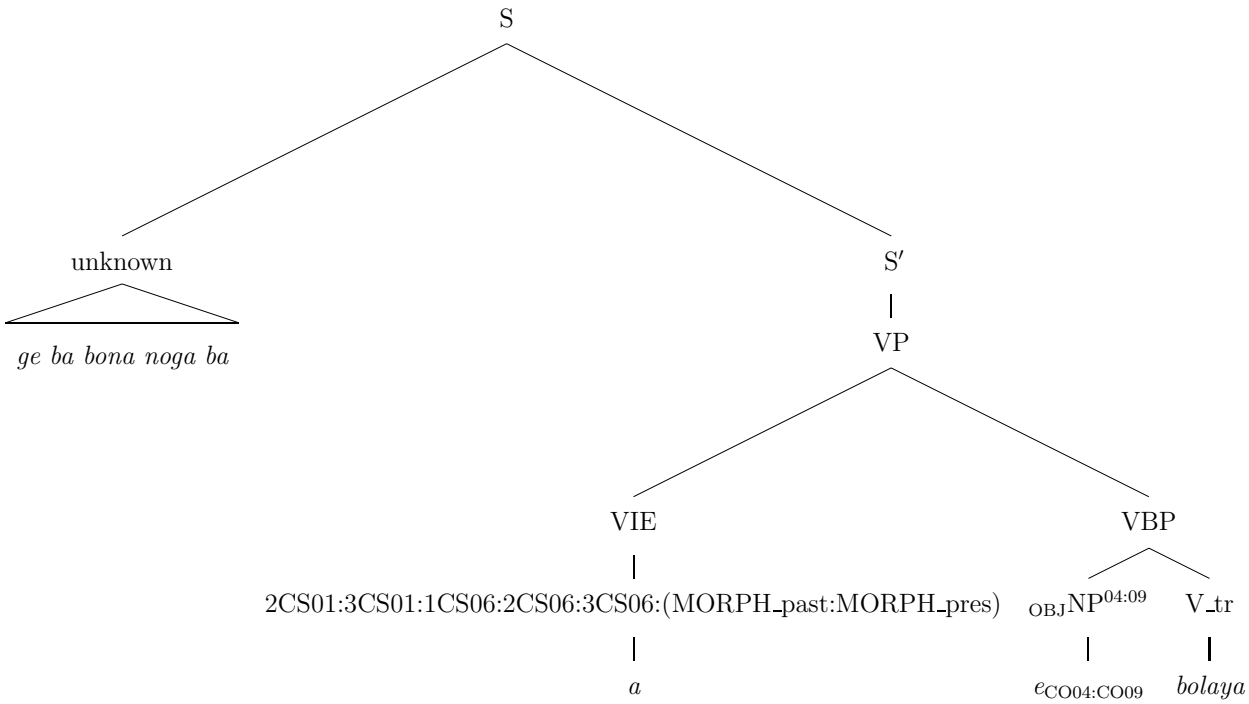


Figure 4.2: A partial analysis of (*ge ba bona noga ba*) *a e bolaya*'`subj-3rd(-cl1/-cl6)/pres/past obj-3rd-cl4:9` kill'

be either a present or past tense morpheme (another condition, namely that the sentence ends after the verb stem, is also fulfilled). The subject concord may in this case only be of the first set and the VP as a whole must be in the indicative mood (present tense, positive or perfect tense, negative), because the constellation contains a tense morpheme. Consider Table 4.4 and Figure 4.3.

Table 4.4: The remaining analyses when considering *ba* – indicative $^{\text{IND}}$VP

	VIE			**VBP**	
descr.	**zero-2**		**zero-1**	**zero**	**zero +1**
pres.pos.long		$1CS_{categ}$	MORPH_pres	VBP^{p}	$.
perf.neg. 4	ga_{MORPH_neg}	$1CS_{categ}$	MORPH_past	VBP	

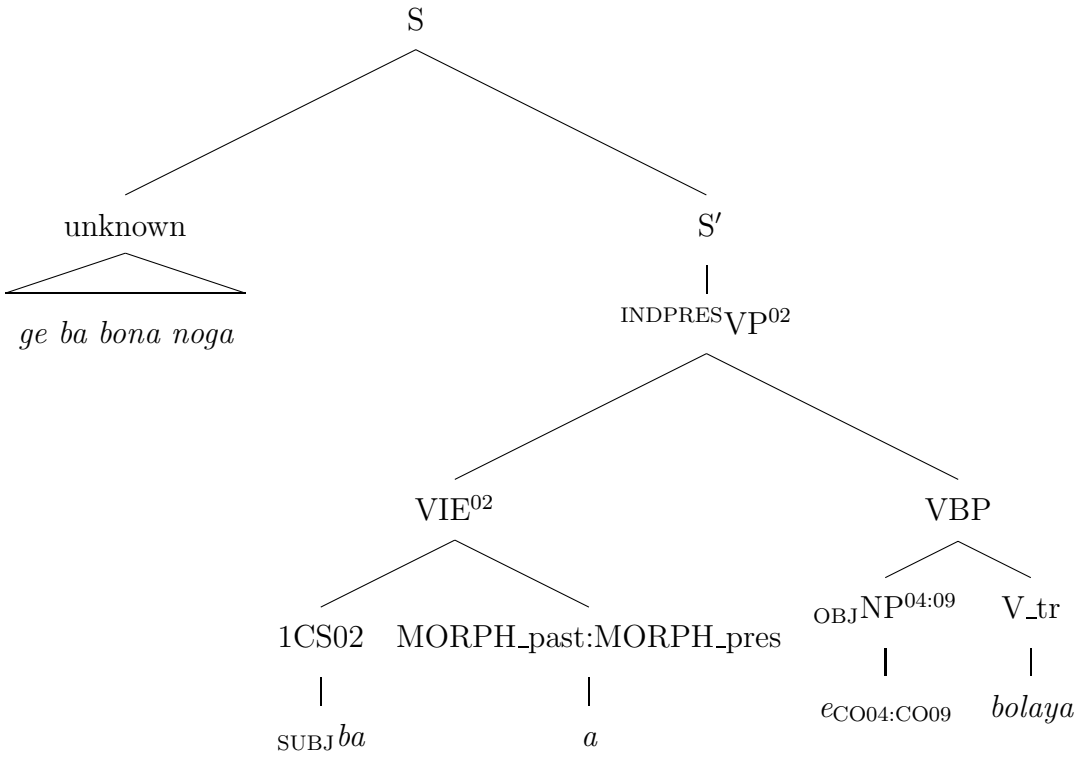


Figure 4.3: A first partial analysis of (*ge ba bona noga*) *ba a e bolaya* '`subj-3rd-cl2 pres obj-3rd-cl4:9` kill'

The token *ba*, however, could also be interpreted as an auxiliary verb[3] – if the following *a* is – as a subject concord – the first element of a VP, as demonstrated by the rules in Table 4.5. If the token should be considered to be a demonstrative concord, *ba* would be (part of) a $_{\text{subj}}$NP and would not be contained in the verbal phrase. In section 3.1 however, we described a general rule for subject-verb agreement in Northern Sotho stating that the subject concord and subject nominal must agree in their noun class. If *ba* represented a demonstrative (of class 2), it would not agree in its noun class with the supposed subject concord *a*, which is from either class 1 or 6. Therefore, the parser can rule out *ba* as CDEM02.

The next token under examination constitutes the unambiguous class 9 noun *noga* '(a) snake', as such it cannot be the subject of the assumed $^{\text{IND}}$VP, because the subject concord

[3]The token *ba* as an auxiliary is to be translated as 'furthermore' or as '(and) so', cf. De Schryver (2007)

Table 4.5: Possible analyses when considering *ba* – auxiliary

	$^{\text{AUX}}$VP		
tense	**elements**		**complement**
pres/perf.	$\mathbf{1CS}_{\text{categ}}$	**V_aux**	$\mathbf{VP}_{\text{categ}}$
future	$\mathbf{1CS}_{\text{categ}}$ $\boldsymbol{tlo/tla}_{\text{MORPH\_fut}}$	**V_aux**	$\mathbf{VP}_{\text{categ}}$
neg	$\boldsymbol{ga}_{\text{MORPH\_neg}}\mathbf{1CS}_{\text{categ}}$	**V_aux**	$\mathbf{VP}_{\text{categ}}$

it appears alongside is from class 2 (ba_{1CS02}). The fact that a noun appears, however, leads to resolving the remaining ambiguity of *a*. MORPH_pres is assigned, because the tag 'past tense morpheme' can only be assigned in the case of a negation morpheme (*ga*) preceding the subject concord *ba*. The noun may indeed also be a topicalized object (the ambiguity concerning the object concord $e_{\text{CO04:CO09}}$ is resolved in this case, cf. Figure 4.4). The second hypothesis entails that *noga* is part of a preceding sentence, not separated by punctuation from the current one under consideration, cf. Figure 4.5. A third and last hypothesis also needs considering, namely that *noga* may be the object of a preceding clause, which may be dependent on the main clause currently being analysed i.e. a situative, relative, consecutive, subjunctive or habitual (cf. Figure 4.6). In this case, it can also be assumed that the object concord of the main clause refers anaphorically to the object noun of the dependent clause, as both are of the same noun class.

In all of the described cases, however, we can already assume *(noga) ba a e bolaya* to constitute a full and complete $^{\text{ind}}\text{VP}^{02}$ of the present tense in the sense of '((a) snake) they kill it' that might agree with a subject nominal of noun class 2 still to appear. The other possibility is that the subject concord *ba* carries the subject function, here no subject nominal is expected anymore.

The verb/pronoun $bona_{\text{PROEMP02:PROPOSS02:V\_tr}}$ is the next item to be examined. Because it is followed by a noun, all three parts of speech categories are theoretically probable. (an NP may contain both, the POS-order PROEMP N and PROPOSS N, cf. section 3.8). However, the following noun is from class 9 and the pronominal word classes of *bona* are only found in

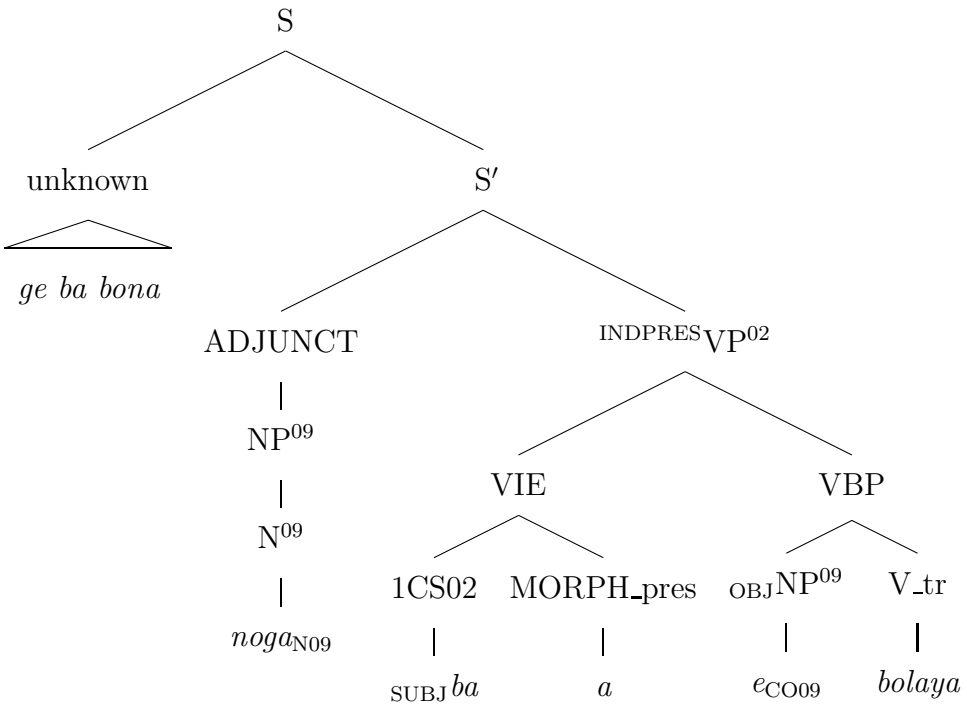


Figure 4.4: Hypothetical partial analysis 1 of (*ge ba bona*) *noga ba a e bolaya* 'snake `subj-3rd-cl2 pres obj-3rd-cl4:9` kill'

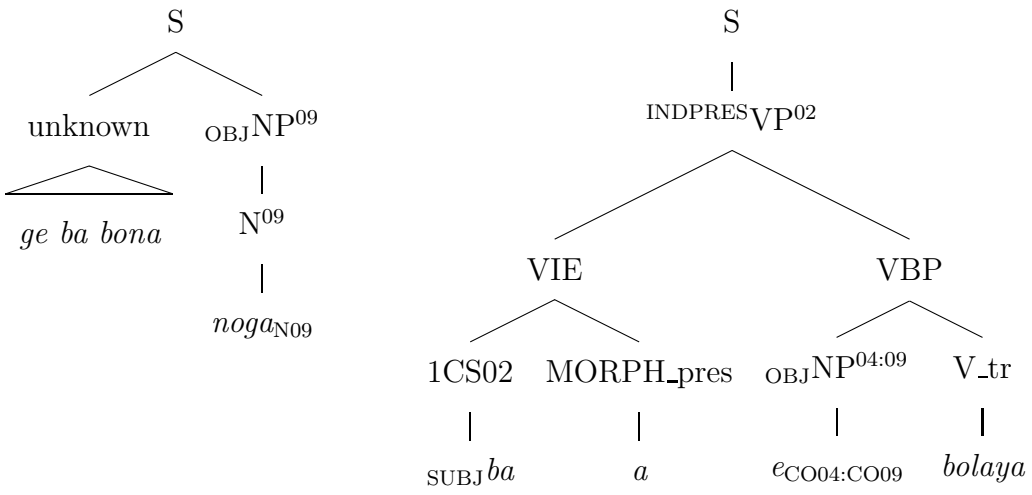


Figure 4.5: Hypothetical partial analysis 2 of (*ge ba bona*) *noga ba a e bolaya* 'snake `subj-3rd-cl2 pres obj-3rd-cl4:9` kill'

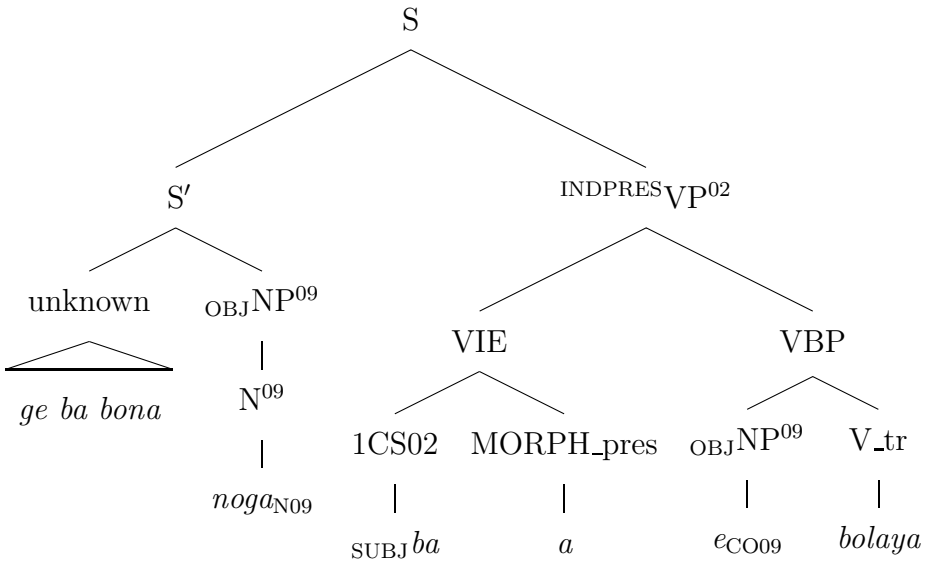


Figure 4.6: Hypothetical partial analysis 3 of (*ge ba bona*) *noga ba a e bolaya* 'snake `subj-3rd-cl2 pres obj-3rd-cl4:9` kill'

class 2. As a noun class agreement between all elements of an NP is mandatory, *bona* must therefore be determined as a transitive verb stem. The VBP-rules stated in paragraph 3.2.3 lead to the assumptions that either *noga* is the object of the transitive verb stem *bona* or that an object concord is to be expected to precede *bona*. Figures 4.7, 4.8, and 4.9 show the respective representations. For the analysis shown in Figure 4.7, note that the preceding element, *ba* is already assumed to be an object concord, as the constellation otherwise would not be legal (see further notes below).

When taking the next preceding token into account, the parser finds the ambiguous token *ba* again, which, as it precedes a transitive verb (which is followed by an object), can be identified by the parser as a subject concord (VCOP is ruled out as a main verb is present; a possessive or demonstrative cannot precede a predicative verb without a subject concord), therefore 1CS02:2CS02:3CS02 are assigned as possible annotations to *ba* and the hypothetical partial analysis shown in Figure 4.7 is abolished.

The structure defined so far is preceded by the last element found, the conjunction *ge*

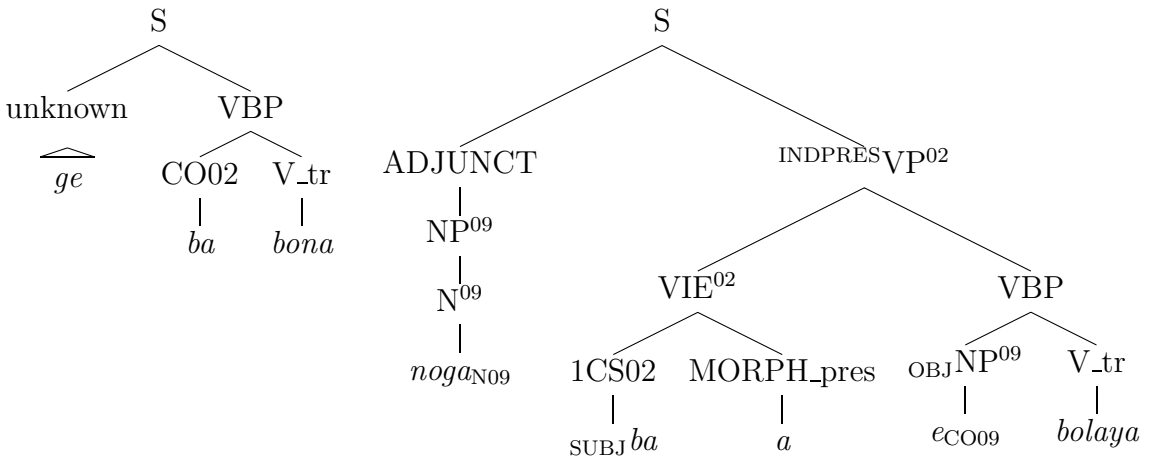


Figure 4.7: Hypothetical partial analysis 1 of (*ge*) *ba bona noga ba a e bolaya* 'obj-3rd-cl2 see (;) snake subj-3rd-cl2 pres obj-3rd-cl4:9 kill'

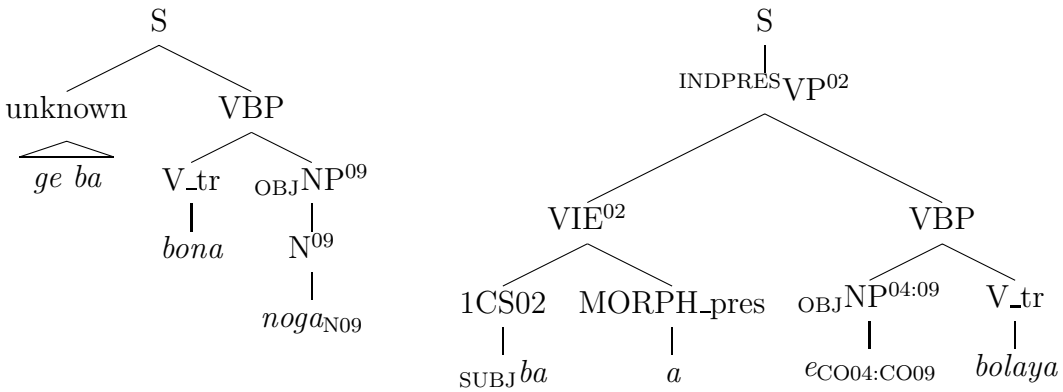


Figure 4.8: Hypothetical partial analysis 2 of (*ge ba*) *bona noga ba a e bolaya* 'see snake (;) subj-3rd-cl2 pres obj-3rd-cl4:9 kill'

'when' indicating that the hypothetical partial analysis 3 is correct and that a dependent clause can be identified (situative mood, cf. paragraph 3.2.6). Consequently, the element *ba* is identified as a subject concord of the second set, 2CS02. The object of the dependent clause, $noga_{\text{N09}}$ could probably be the antecedent for the object concord *e*. Therefore, as both are of the same class, it may be assumed that *e* anaphorically refers to this class 9 object. Figure 4.10 shows the resulting, unequivocal analysis.

This example demonstrated that utilising generalisations based on grammar rules reduces ambiguity significantly during analysis. Such generalisations – at a later stage – may also help to develop a linguistic sequence model of Northern Sotho sentences.

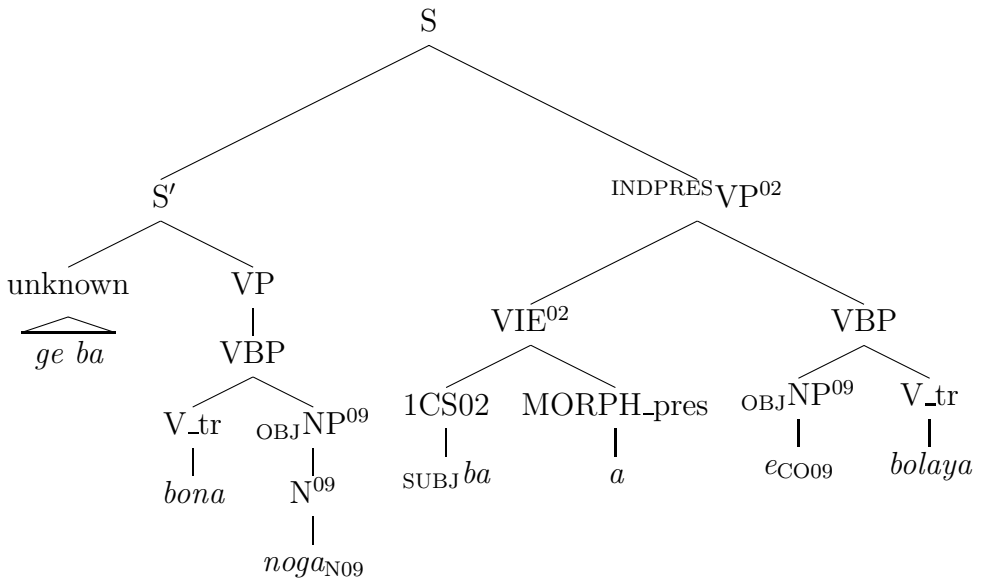


Figure 4.9: Hypothetical partial analysis 3 of (*ge ba*) *bona noga ba a e bolaya* 'see snake `subj-3rd-cl2 pres obj-3rd-cl4:9` kill'

In the following sections, the moods defined will be categorised by their features in order to find more such generalisations. The features considered are the selection of sets of subject concords; the utilization of tense morphemes; the negation constellations utilised by a mood and their respective combinations. Main verbs and copulative constellations are described separately because of the differences in their internal structures.

4.4 A basis for a data category inventory of main verbs

All main verb constellations as described in chapter 3[4], were examined for regular patterns in the distribution of features like verbal endings (abbreviated as Vend), negation morphemes or clusters thereof and subject concords. Table 4.6 illustrates the distribution of verbal endings. Tables 4.7 and 4.8 illustrate the selection of the three sets of subject concords (cf. Table 3.13), and the negated constellations of the described moods for the main verbs. Table 4.9 shows the distribution of morphemes indicating tenses.

[4]For the summaries, consider Tables 3.19 on page 104, 3.2.8 on page 119, and 3.2.11 on page 125.

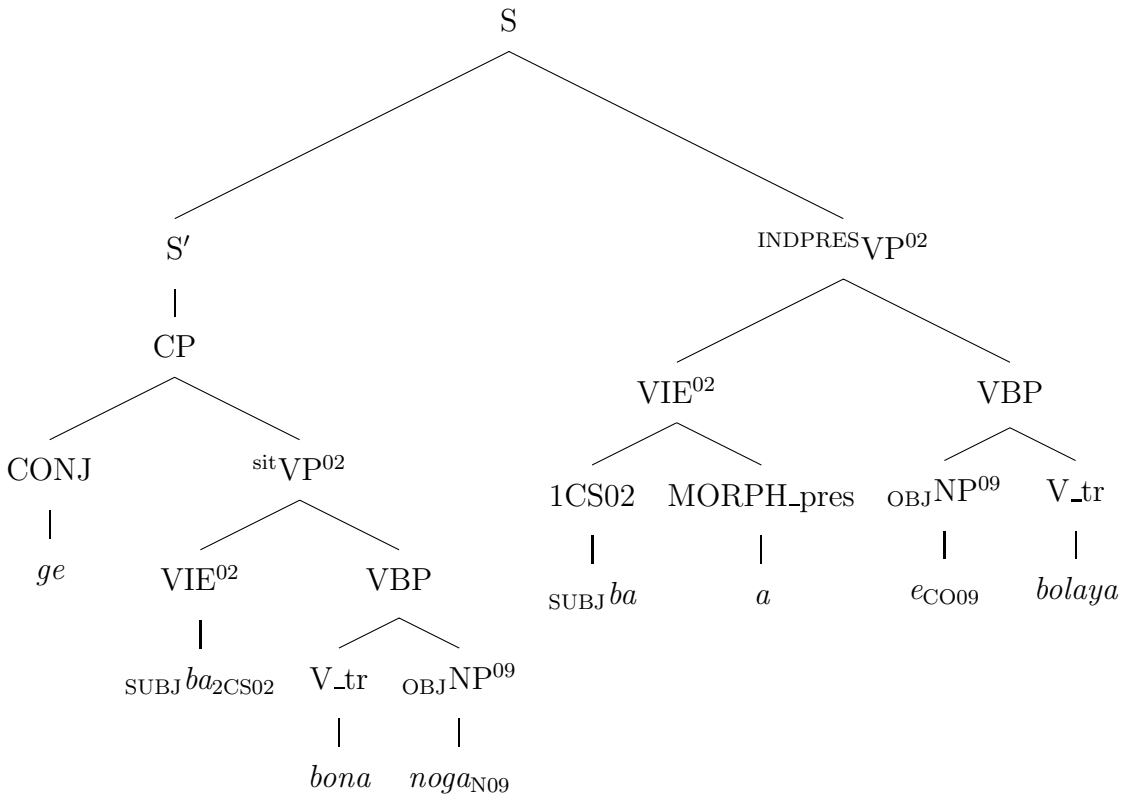


Figure 4.10: Resulting analysis of *ge ba bona noga ba a e bolaya* 'when *subj-3rd-cl2* see snake `subj-3rd-cl2 pres obj-3rd-cl9` kill'

From a right-to-left perspective[5], the Vend *–e* usually appears in negated predicative moods (with the exceptions of the subjunctive/habitual where it occurs in the positive case and of some non-predicative imperative constellations), however, there are a number of negated moods where Vend *–a* is used. Table 4.8 shows (from a left-to-right perspective) that each mood uses its own negation morpheme(s). The morpheme $ga_{\text{MORPH\_neg}}$ exclusively appears in indicative constellations, while *se* is used by the non-predicative (infinitive and imperative) and predicative dependent (consecutive, subjunctive and habitual) moods. However, this morpheme also appears in two alternatives of the negated perfect tense situatives. The negation $ka_{\text{MORPH\_pot}}$ $\text{se}_{\text{MORPH\_neg}}$ is contained in all negated future tense moods indicating that the future tense of Northern Sotho might be closely related to the potential, i.e. that it is of a more aspectual nature than the other tenses[6]. Interestingly, the negation *ka se* usually inhibits the appearance of a future tense morpheme with the one exception of the 2nd form of the relative future tense mood, where both occur. The past-tense morpheme in Table 4.9 only appears in one of the several possible negated perfect tense indicatives. Because neither non-predicative nor the dependent moods appear with tense morphemes, those moods are not mentioned in this table.
Combining the given information on 'verbal ending' with the attribute 'negation morphemes' and their clusters, i.e. simultaneously considering these two properties of VPs from right-to-left and from left-to-right may hint at a logic in their distribution, as $ga_{\text{MORPH\_neg}}$ only appears with Vend *–a* and $ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ (and also $se_{\text{MORPH\_neg}}$ when appearing alone) exclusively with Vend *–e*. As said above, the negated future tense moods make use of the negated potential, i.e. the morpheme cluster $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$. The negated future tense forms of indicative and situative appear with Vend *–e*, while the negated future tense of the relative appears with Vend *–a*. The negation $sa_{\text{MORPH\_neg}}$ appears with both verbal endings, it could however be relevant that the combination $sa_{\text{MORPH\_neg}}$+Vend *–a* only appears in the perfect tense while $sa_{\text{MORPH\_neg}}$+Vend *–e* seems to represent present tense forms.

Generalisations on the basis of the subject concords that might appear on the leftmost position of VP, are not easy to determine, as their appearances in the moods do not seem to follow rules. However, the selection of subject concords seems to be solely dependent

[5]As a preferable direction for the horizontal parsing of Northern Sotho text has not yet become evident, the possible directions are explicitly described for each generalisation mentioned in this section.

[6]Note again (cf. paragraph 3.5.2) that the potential itself does not appear in a future tense form.

Table 4.6: Verbal endings of Northern Sotho moods

mood	**verbal ending**					
	−a	***−e***	***−ile***	***−a*** +'rel.'	***−e*** +'rel.'	***−ile+−a*** +'rel.'
imperative.pos	√					
imperative.neg		√				
infinitive.pos	√					
infinitive.neg		√				
ind.pres.pos.long	√					
ind.pres.pos.short	√					
ind.pres.neg		√				
ind.perf.pos			√			
ind.perf.neg.1	√					
ind.perf.neg.2		√				
ind.perf.neg.3	√					
ind.perf.neg.4	√					
ind.fut.pos	√					
ind.fut.neg		√				
sit.pres.pos	√					
sit.pres.neg		√				
sit.perf.pos			√			
sit.perf.neg1	√					
sit.perf.neg2	√					
sit.perf.neg3	√					
sit.fut.pos	√					
sit.fut.neg		√				
rel.pres.pos				√		
rel.pres.neg					√	
rel.perf.pos						√
rel.perf.neg1	√					
rel.perf.neg2		√				
rel.fut.pos1	√					
rel.fut.pos2				√		
rel.fut.neg1	√					
rel.fut.neg2				√		
rel.fut.neg3				√		
consecutive.pos	√					
consecutive.neg		√				
subj./habit.pos		√				
subj./habit.neg		√				

on the moods in which they appear. Set 1, for example (using *o* for class 1 subjects), dominates the indicative (positive) forms, while set 2 (using *a* for class 1 subjects) appears in all moods, either indicating the negated independent forms or the modifying constellations. Concerning Table 4.7, note that some constellations make use of several subject concords; an auxiliary for example is always surrounded by subject concords. For the sake of convenience, here is example (68) again as (79) to demonstrate this issue. In Table 4.7, the position of the subject concord in question appears as an ordinal number.

(79) *wena*$_{\text{PROEMPPERS\_2sg}}$ ***o***$_{\text{1CSPERS\_2sg}}$ *be*$_{\text{V\_aux}}$ ***o***$_{\text{2CSPERS\_2sg}}$ *šomile*$_{\text{V\_itr}}$
`emp-2nd-sg` `subj-2nd-sg` `past` `subj-2nd-sg` work
'you have worked'

In terms of word order, the negation morpheme *ga* and the cluster *ga se* are the only negation constellations that may precede the subject concord while in all other moods it is the subject concord that introduces the verbal phrase. The non-predicative moods do not entail subject concords.

Still, some exceptions (like the subject concord of set one appearing in an auxiliary constellation in the negated perfect tense of the situative (cf. paragraph 3.2.6)) inhibit the formulation of rules solely based on the first or last element(s) of the VIE. So far, one cannot find a relation between the distribution of subject concords and negations or verbal endings, neither.

Table 4.7: Feature selection of Northern Sotho moods: subject concords in slot zero-2

mood	**subject concords**		
	1CS$_{\text{categ}}$	**2CS$_{\text{categ}}$**	**3CS$_{\text{categ}}$**
ind.pres.pos.long	√		
ind.pres.pos.short	√		
ind.pres.neg		√	
ind.perf.pos	√		
ind.perf.neg.1			√
ind.perf.neg.2		√	
ind.perf.neg.3			√
ind.perf.neg.4	√		
ind.fut.pos	√		
ind.fut.neg		√	
sit.pres.pos		√	
sit.pres.neg		√	
sit.perf.pos		√	
sit.perf.neg.1		√(1.)	√(2.)
sit.perf.neg.2	√(2.)	√(1.)	
sit.perf.neg.3		√	
sit.fut.pos		√	
sit.fut.neg		√	
rel.pres.pos		√	
rel.pres.neg		√	
rel.perf.pos		√	
rel.perf.neg.1		√(1.)	√(2.)
rel.perf.neg.2		√(2x)	
rel.fut.pos.1		√	
rel.fut.pos.2		√	
rel.fut.neg.1		√	
rel.fut.neg.2		√	
rel.fut.neg.3		√	
consecutive.pos			√
consecutive.neg			√
subj./habit.pos		√	
subj./habit.neg		√	

Table 4.8: Feature selection of Northern Sotho moods: negation morphemes in slot zero-2

mood	**negation morphemes MORPH_neg**							
	ga	***se***	***ga se***		***ka se***		***sa***	***sego/ seng***
imperative.neg		√						
infinitive.neg		√						
ind.pres.neg	√							
ind.perf.neg.1			√					
ind.perf.neg.2			√					
ind.perf.neg.3	√							
ind.perf.neg.4	√							
ind.fut.neg					√			
sit.pres.neg							√	
sit.perf.neg.1		√						
sit.perf.neg.2		√						
sit.perf.neg.3							√	
sit.fut.neg					√			
rel.pres.neg							√	
rel.perf.neg.1								√
rel.perf.neg.2								√
rel.fut.neg.1					√			
rel.fut.neg.2					√			
rel.fut.neg.3					√			
consecutive.neg		√						
subj./habit.neg		√						

Table 4.9: Feature selection of Northern Sotho moods: slot zero-1

mood	**tense morphemes MORPH_**			
	_pres	**'past'**	**_fut**	**_fut**
	a	***a***	***tlo/tla***	***tlogo/tlago***
ind.pres.pos.long	√			
ind.pres.pos.short				
ind.pres.neg				
ind.perf.pos				
ind.perf.neg.1				
ind.perf.neg.2				
ind.perf.neg.3				
ind.perf.neg.4		√		
ind.fut.pos			√	
ind.fut.neg				
sit.pres.pos				
sit.pres.neg				
sit.perf.pos				
sit.perf.neg.1				
sit.perf.neg.2				
sit.perf.neg.3				
sit.fut.pos			√	
sit.fut.neg				
rel.pres.pos				
rel.pres.neg				
rel.perf.pos				
rel.perf.neg.1				
rel.perf.neg.2				
rel.fut.pos.1				√
rel.fut.pos.2			√	
rel.fut.neg.1				√
rel.fut.neg.2			√	
rel.fut.neg.3				

4.5 A basis for a data category inventory of copulas

The data categories of the copulative constellations as described in section 3.3 can be sorted according to the copula heading them. A second way to gain an overview of the data categories is to sort the cases according to the constellations preceding the copulas. In this section, both illustrations are contained.

4.5.1 A right-to-left perspective

A number of regularities appear when the occurrences of the different copulas appearing on the right hand side of the copulative constellations have been examined. For example, the copula *le* only appears in stative copulatives, and is always preceded by (a combination of) subject concords and – in the perfect tense – also auxiliaries. Only descriptive constellations contain class-specific concords, while identifying constellations are preceded by CSPCSN (either CSNEUT, i.e. the neutral, or CSPERS, i.e. the personal subject concords[7]). Tables 4.10 (containing *ba, eba, bile*), 4.11 (containing *be*), 4.12 (containing *bago, bego, bilego*), 4.13 (containing *le, lego*), 4.14 (containing the subject concords appearing as copulas $1VCOP_{categ}$ (set 1) or $2VCOP_{categ}$ (set 2)), 4.15 (containing *na, (e)na, nago*), and 4.16 (containing *se, sego/seng*) show summaries of the moods together with the features of the respective copulas, i.e. the element clusters preceding them.

A number of constellations are inherently ambiguous, however, as indicated in the section on copulatives (for example in paragraph 3.3.3), the complement of the copulative may indicate the correct analysis. Moreover, some of the constellations are introduced by a specific conjunction or by demonstrative concords. Such and other phrase-external elements also support the disambiguation process.

[7]The personal subject concords and the neutral subject concord were grouped as 'CSPCSN' in 3.31 on page 131, and we will use this and the other abbreviations again in the following tables.

Table 4.10: Distribution of *ba, eba, bile*

VCOP	**preceding elements**	**copulative**				
ba	CSPCSN	identifying	dynamic	indicative situative	pres.	pos.
	CSPCSN *sa*$_{\text{MORPH\_neg}}$	identifying	dynamic	situative	perf.	neg.
	CSPCSN *tlo/tla*$_{\text{MORPH\_fut}}$	identifying	dynamic	indicative situative	fut.	pos.
	CSPCSN *tlogo/tlago*$_{\text{MORPH\_fut}}$	identifying	dynamic	relative	fut.	pos.2
	1CS$_{\text{categ}}$	descriptive associative	dynamic	indicative	pres.	pos.
	1CS$_{\text{categ}}$ *tlo/tla*$_{\text{MORPH\_fut}}$	descriptive associative	dynamic	indicative	fut.	pos.
	2CS$_{\text{categ}}$	descriptive	dynamic	situative	pres.	pos.
	2CS$_{\text{categ}}$ *sa*$_{\text{MORPH\_neg}}$	descriptive associative	dynamic	situative	perf.	neg.
	2CS$_{\text{categ}}$ *tlo/tla*$_{\text{MORPH\_fut}}$	descriptive associative	dynamic	situative	fut.	pos.
	2CS$_{\text{categ}}$ *tlogo/tlago*$_{\text{MORPH\_fut}}$	descriptive	dynamic	relative	fut.	pos.2
	3CS$_{\text{categ}}$	identifying descriptive	dynamic	consecutive	–	pos.
	go$_{\text{MORPH\_cp15}}$			infinitive	–	pos.
eba	2CS$_{\text{categ}}$	associative	dynamic	situative	pres.	pos.
	–			imperative	–	pos.
bile	CSPCSN	identifying	dynamic	indicative situative	perf.	pos.
	1CS$_{\text{categ}}$	descriptive associative	dynamic	indicative	perf.	pos.
	2CS$_{\text{categ}}$	descriptive associative	dynamic	situative	perf.	pos.

Table 4.11: Distribution of *be*

VCOP	preceding elements	copulative				
be	CSPCSN	identifying	dynamic	subjunctive habitual	–	pos.
	CSPCSN $se_{\text{MORPH\_neg}}$	identifying	dynamic	subjunctive habitual	–	neg.
	CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	situative	pres.	neg.
	$ga_{\text{MORPH\_neg}}$ CSPCSN	identifying	dynamic	indicative	pres.	neg.
	CSPCSN $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	identifying	dynamic	indicative situative	fut.	neg.
	$2CS_{\text{categ}}$	descriptive	dynamic	subjunctive habitual	–	pos.
	$2CS_{\text{categ}}$ $se_{\text{MORPH\_neg}}$	descriptive	dynamic	subjunctive habitual	–	neg.
	$2CS_{\text{categ}}$ $sa_{\text{MORPH\_neg}}$	descriptive associative	dynamic	situative	pres	neg.
	$ga_{\text{MORPH\_neg}}$ $2CS_{\text{categ}}$	descriptive associative	dynamic	indicative	pres.	neg.
	$2CS_{\text{categ}}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	descriptive associative	dynamic	indicative situative indicative situative	fut.	neg.
	$go_{\text{MORPH\_cp15}}$ $se_{\text{MORPH\_neg}}$			infinitive	–	neg.
	$se_{\text{MORPH\_neg}}$			imperative	–	neg.

Table 4.12: Distribution of *bago, bego, bilego*

VCOP	preceding elements	copulative				
bago	CSPCSN	identifying	dynamic	relative	pres.	pos.
	CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	relative	perf.	neg.
	CSPCSN $tlo/tla_{\text{MORPH\_fut}}$	identifying	dynamic	relative	fut.	pos.
	$2\text{CS}_{\text{categ}}$	descriptive associative	dynamic	relative	pres.	pos.
	$2\text{CS}_{\text{categ}}$ $sa_{\text{MORPH\_neg}}$	descriptive associative	dynamic	relative	perf.	neg.
	$2\text{CS}_{\text{categ}}$ $tlo/tla_{\text{MORPH\_fut}}$	descriptive associative	dynamic	relative	fut.	pos.
bego	CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	relative	pres.	neg.
	CSPCSN $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	identifying	dynamic	relative	fut.	neg.
	$2\text{CS}_{\text{categ}}$ $sa_{\text{MORPH\_neg}}$	descriptive associative	dynamic	relative	pres.	neg.
	$2\text{CS}_{\text{categ}}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	descriptive associative	dynamic	relative	fut.	neg.
bilego	CSPCSN	identifying	dynamic	relative	perf.	pos.
	$2\text{CS}_{\text{categ}}$	descriptive associative	dynamic	relative	perf.	pos.

Table 4.13: Distribution of *le, lego*

VCOP	preceding elements	copulative				
le	CSPCSN	identifying	stative	situative	pres.	pos.
	CSPCSN $be_{\text{V\_aux}}$ CSPCSN	identifying	stative	indicative situative	perf.	pos.
	CSPCSN $bego_{\text{V\_aux}}$ CSPCSN	identifying	stative	relative	perf.	pos.
	$1\text{CS}_{\text{categ}}$ $be_{\text{V\_aux}}$ $2\text{CS}_{\text{categ}}$	descriptive	stative	indicative	perf.	pos.
	$2\text{CS}_{\text{categ}}$	descriptive	stative	situative	pres.	pos.
	$2\text{CS}_{\text{categ}}$ $be_{\text{V\_aux}}$ $2\text{CS}_{\text{categ}}$	descriptive	stative	situative	perf.	pos.
	$2\text{CS}_{\text{categ}}$ $bego_{\text{V\_aux}}$ $2\text{CS}_{\text{categ}}$	descriptive	stative	relative	perf.	pos.
lego	CSPCSN	identifying	stative	relative	pres.	pos.
	$2\text{CS}_{\text{categ}}$	descriptive	stative	relative	pres.	pos.

Table 4.14: Distribution of *na, (e)na, nago*

VCOP	**preceding elements**	**copulative**				
na	$1CS_{categ}$	associative	stative	indicative	pres.	pos.
	$2CS_{categ}$ se_{MORPH_neg}	associative	stative	situative	pres.	pos.
	ga_{MORPH_neg} $2CS_{categ}$	associative	stative	indicative	pres.	neg.
	$2CS_{categ}$ $bego_{V\_aux}$ $2CS_{categ}$	associative	stative	relative	perf.	pos.
	$2CS_{categ}$ $bego_{V\_aux}$ $2CS_{categ}$ se_{MORPH_neg}	associative	stative	relative	perf.	neg.
(e)na	$1CS_{categ}$ be_{V_aux} $2CS_{categ}$	associative	stative	indicative	perf.	pos.
	$1CS_{categ}$ be_{V_aux} $2CS_{categ}$ se_{MORPH_neg}	associative	stative	indicative	perf.	neg.
	$2CS_{categ}$	associative	stative	situative	pres.	pos.
	$2CS_{categ}$ be_{V_aux} $2CS_{categ}$	associative	stative	situative	perf.	pos.
	$2CS_{categ}$ be_{V_aux} $2CS_{categ}$ se_{MORPH_neg}	associative	stative	situative	perf.	neg.
nago	$2CS_{categ}$	associative	stative	relative	pres.	pos.
	$2CS_{categ}$ se_{MORPH_neg}	associative	stative	relative	pres.	neg.

Table 4.15: Distribution of $VCOP_{categ}$

VCOP	**preceding elements**	**copulative**				
1VCOP$_{categ}$	–	descriptive	stative	indicative	pres.	pos.
2VCOP$_{categ}$	ga_{MORPH_neg}	descriptive	stative	indicative	pres.	neg.
VCOP_pers	–	identifying	stative	indicative	pres.	pos.
VCOP_pers	ga_{MORPH_neg}	identifying	stative	indicative	pres.	neg.

Table 4.16: Distribution of *se, sego/seng*

VCOP	**preceding elements**	**copulative**				
se$_{\text{VCOP\_neg}}$	CSPCSN	identifying descriptive	stative	situative	pres.	neg.
	CSPCSN *be*$_{\text{V\_aux}}$ CSPCSN	identifying	stative	indicative situative	perf.	neg.
	CSPCSN *bego*$_{\text{V\_aux}}$ CSPCSN	identifying	stative	relative	perf.	neg.
	1CS$_{\text{categ}}$ *be*$_{\text{V\_aux}}$ 2CS$_{\text{categ}}$	descriptive	stative	indicative	perf.	neg.
	2CS$_{\text{categ}}$ *be*$_{\text{V\_aux}}$ 2CS$_{\text{categ}}$	descriptive	stative	situative	perf.	neg.
	2CS$_{\text{categ}}$ *bego*$_{\text{V\_aux}}$ 2CS$_{\text{categ}}$	descriptive	stative	relative	perf.	neg.
	ga$_{\text{MORPH\_neg}}$	identifying	stative	pres.	neg.	
sego/seng	CSPCSN	identifying	stative	relative	pres.	neg.
VCOP_neg	2CS$_{\text{categ}}$	descriptive	stative	relative	pres.	neg.

Table 4.17: Distribution of constellations solely containing copulas

preceding elements	**copulative**					**VCOP**
–	identifying	stative	indicative	pres.	pos.	VCOP_pers
–	descriptive	stative	indicative	pres.	pos.	$1VCOP_{categ}$
–			imperative	–	pos.	*eba*

4.5.2 A left-to-right perspective

Sorting the copulatives according to the element constellations preceding them (as in Tables 4.17 to 4.29) shows that there are a number of generalisations possible on the basis of the left-hand side of the copula; an observation that could support a left-to-right parsing strategy.

We begin with Table 4.17, illustrating constellations that solely contain copulas. Tokens named as VCOP are actually subject concords and their identification as copulas is only possible when taking the context on their righthand side into account: if no verb stem appears, the subject concord may be assumed to be a copulative.

The subject concord group CSPCSN (consisting of the set of all $CSPERS_{categ}$ and of CSNEUT) never appears as an element of the associative copulatives (cf. Table 4.18) while the combination CSPCSN and future tense morpheme moreover never appears in an indicative (cf. Table 4.19). On the other hand, if CSPCSN is followed by either the negation morpheme *se* or *sa* or the cluster *ka se* (as in Tables 4.20 and 4.21), the constellation is clearly an identifying dynamic copulative (but no indicative).

Table 4.22 shows that CSPCSN constellations containing auxiliaries always appear in the perfect tense of an identifying stative copulative. These constellations appear with the copula *le* in the positive and are all negated by the negated copula se_{MORPH_neg}. All VIEs beginning with a subject concord of the first set are indeed indicatives, however, they do not present identifying copulatives, as shown in Table 4.23. The latter is also true for the VIEs beginning with a subject concord of the second set, however, all modifying and the dependent moods are found in Table 4.24.

Table 4.18: Distribution of CSPERS and CSNEUT (CSPCSN)

preceding elements	**copulative**					**VCOP**
CSPCSN	identifying	stative	situative	pres.	pos.	*le*
CSPCSN	identifying	stative	situative	pres.	neg.	$se_{\text{VCOP\_neg}}$
CSPCSN	identifying	stative	relative	pres.	pos.	*lego*
CSPCSN	identifying	stative	relative	pres.	neg.	$sego/seng_{\text{VCOP\_neg}}$
CSPCSN	identifying	dynamic	indicative	pres.	pos.	*ba*
CSPCSN	identifying	dynamic	indicative	perf.	pos.	*bile*
CSPCSN	identifying	dynamic	situative	pres.	pos.	*ba*
CSPCSN	identifying	dynamic	situative	perf.	pos.	*bile*
CSPCSN	identifying	dynamic	relative	pres.	pos.	*bago*
CSPCSN	identifying	dynamic	relative	perf.	pos.	*bilego*
CSPCSN	identifying	dynamic	subjunctive	–	pos.	*be*
CSPCSN	identifying	dynamic	habitual	–	pos.	*be*
CSPCSN	descriptive	stative	situative	pres.	neg.	$se_{\text{VCOP\_neg}}$

Table 4.19: Distribution of CSPCSN and future tense morphemes

preceding elements	**copulative**					**VCOP**
CSPCSN $tlo/tla_{\text{MORPH\_fut}}$	identifying	dynamic	indicative	fut.	pos.	*ba*
CSPCSN $tlo/tla_{\text{MORPH\_fut}}$	identifying	dynamic	situative	fut.	pos.	*ba*
CSPCSN $tlo/tla_{\text{MORPH\_fut}}$	identifying	dynamic	relative	fut.	pos.	*bago*
CSPCSN $tlogo/tlago_{\text{MORPH\_fut}}$	identifying	dynamic	relative	fut.	pos.2	*ba*

Table 4.20: Distribution of CSPCSN followed by a negation morpheme

preceding elements	**copulative**					**VCOP**
CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	situative	pres.	neg.	*be*
CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	situative	perf.	neg.	*ba*
CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	relative	pres.	neg.	*bego*
CSPCSN $sa_{\text{MORPH\_neg}}$	identifying	dynamic	relative	perf.	neg.	*bago*
CSPCSN $se_{\text{MORPH\_neg}}$	identifying	dynamic	subjunctive	–	neg.	*be*
CSPCSN $se_{\text{MORPH\_neg}}$	identifying	dynamic	habitual	–	neg.	*be*

Table 4.21: Distribution of CSPCSN followed by *ka se*

preceding elements	**copulative**					**VCOP**
CSPCSN $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	identifying	dynamic	indicative	fut.	neg.	*be*
CSPCSN $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	identifying	dynamic	situative	fut.	neg.	*be*
CSPCSN $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$	identifying	dynamic	relative	fut.	neg.	*bego*

bottomline

Table 4.22: Distribution of CSPCSN followed by auxiliary constellations

preceding elements	**copulative**					**VCOP**
CSPCSN $be_{\text{V\_aux}}$ CSPCSN	identifying	stative	indicative	perf.	pos.	*le*
CSPCSN $be_{\text{V\_aux}}$ CSPCSN	identifying	stative	indicative	perf.	neg.	$se_{\text{VCOP\_neg}}$
CSPCSN $be_{\text{V\_aux}}$ CSPCSN	identifying	stative	situative	perf.	pos.	*le*
CSPCSN $be_{\text{V\_aux}}$ CSPCSN	identifying	stative	situative	perf.	neg.	$se_{\text{VCOP\_neg}}$
CSPCSN $bego_{\text{V\_aux}}$ CSPCSN	identifying	stative	relative	perf.	pos.	*le*
CSPCSN $bego_{\text{V\_aux}}$ CSPCSN	identifying	stative	relative	perf.	neg.	$se_{\text{VCOP\_neg}}$

Table 4.23: Distribution of constellations beginning with $1CS_{\text{categ}}$

preceding elements	**copulative**					**VCOP**
$1CS_{\text{categ}}$	descriptive	dynamic	indicative	pres.	pos.	*ba*
$1CS_{\text{categ}}$	descriptive	dynamic	indicative	perf.	pos.	*bile*
$1CS_{\text{categ}}$	associative	dynamic	indicative	pres.	pos.	*ba*
$1CS_{\text{categ}}$	associative	dynamic	indicative	perf.	pos.	*bile*
$1CS_{\text{categ}}$	associative	stative	indicative	pres.	pos.	*na*
$1CS_{\text{categ}}$ $tlo/tla_{\text{MORPH\_fut}}$	descriptive	dynamic	indicative	fut.	pos.	*ba*
$1CS_{\text{categ}}$ $tlo/tla_{\text{MORPH\_fut}}$	associative	dynamic	indicative	fut.	pos.	*ba*
$1CS_{\text{categ}}$ $be_{\text{V\_aux}}$ $2CS_{\text{categ}}$	descriptive	stative	indicative	perf.	pos.	*le*
$1CS_{\text{categ}}$ $be_{\text{V\_aux}}$ $2CS_{\text{categ}}$	descriptive	stative	indicative	perf.	neg	$se_{\text{VCOP\_neg}}$
$1CS_{\text{categ}}$ $be_{\text{V\_aux}}$ $2CS_{\text{categ}}$	associative	stative	indicative	perf.	pos.	*(e)na*
$1CS_{\text{categ}}$ $be_{\text{V\_aux}}$ $2CS_{\text{categ}}$ $se_{\text{MORPH\_neg}}$	associative	stative	indicative	perf.	neg.	*(e)na*

Table 4.24: Distribution of $2CS_{\text{categ}}$

preceding elements	**copulative**					**VCOP**
$2CS_{\text{categ}}$	descriptive	stative	situative	pres.	pos.	*le*
$2CS_{\text{categ}}$	descriptive	stative	relative	pres.	pos.	*lego*
$2CS_{\text{categ}}$	descriptive	stative	relative	pres.	neg	$sego/seng_{\text{VCOP\_neg}}$
$2CS_{\text{categ}}$	descriptive	dynamic	situative	pres.	pos.	*ba*
$2CS_{\text{categ}}$	descriptive	dynamic	situative	perf.	pos.	*bile*
$2CS_{\text{categ}}$	descriptive	dynamic	relative	pres.	pos.	*bago*
$2CS_{\text{categ}}$	descriptive	dynamic	relative	perf.	pos.	*bilego*
$2CS_{\text{categ}}$	descriptive	dynamic	subjunctive	–	pos.	*be*
$2CS_{\text{categ}}$	descriptive	dynamic	habitual	–	pos.	*be*
$2CS_{\text{categ}}$	associative	stative	situative	pres.	pos.	*(e)na*
$2CS_{\text{categ}}$	associative	stative	relative	pres.	pos.	*nago*
$2CS_{\text{categ}}$	associative	dynamic	situative	pres.	pos.	*eba*
$2CS_{\text{categ}}$	associative	dynamic	situative	perf.	pos.	*bile*
$2CS_{\text{categ}}$	associative	dynamic	relative	pres.	pos.	*bago*
$2CS_{\text{categ}}$	associative	dynamic	relative	perf.	pos.	*bilego*

Table 4.25: Distribution of $2CS_{categ}$ followed by future tense morphemes

preceding elements		**copulative**					**VCOP**
$2CS_{categ}$ *tlo/tla*$_{MORPH\_fut}$	descriptive	dynamic	situative	fut.	pos.	*ba*
$2CS_{categ}$ *tlo/tla*$_{MORPH\_fut}$	descriptive	dynamic	relative	fut.	pos.	*bago*
$2CS_{categ}$ *tlogo/tlago*$_{MORPH\_fut}$	descriptive	dynamic	relative	fut.	pos.2	*ba*
$2CS_{categ}$ *tlo/tla*$_{MORPH\_fut}$	associative	dynamic	situative	fut.	pos.	*ba*
$2CS_{categ}$ *tlo/tla*$_{MORPH\_fut}$	associative	dynamic	relative	fut.	pos.	*bago*

As a counterpart to Table 4.19 (containing identifying copulatives), future tense morphemes following a subject concord of the second set indicate a descriptive or associative copulative, as illustrated in Table 4.25. In both cases (CSPCSN and $2CS_{categ}$), the copula *ba*$_{VCOP}$ appears in the indicative/situative, its relative form *bago* in the relative moods. The same stands for the copulatives containing a subject concord of the second set followed by the negation morphemes *sa*, *se*, or the cluster *ka se*, as in Table 4.26. Table 4.22, where the VIE was introduced by CSPCSN (followed by an auxiliary constellation), only contained identifying stative copulatives. Table 4.27 shows that such copulatives introduced by $2CS_{categ}$ indicate the descriptive and associative cases. There are only two cases beginning with the subject concord of the third set, $3CS_{categ}$; both identify positive dynamic consecutive moods, as in Table 4.28.

Lastly, Table 4.29 shows the cases where a copulative begins with the infinitive or negation morphemes. Of these, the infinitive and the imperative are clear-cut cases, because no other constellation can begin with these morphemes. The other cases contain all three main categories (identifying, descriptive, and associative) and both sub-categories (stative and dynamic) of copulatives. However, each of them are negated copulatives in the indicative mood and in the present tense.

Table 4.26: Distribution of $2CS_{categ}$ followed by negation morphemes and negation clusters

preceding elements	**copulative**					**VCOP**
$2CS_{categ}$ sa_{MORPH_neg}	descriptive	dynamic	situative	pres	neg.	*be*
$2CS_{categ}$ sa_{MORPH_neg}	descriptive	dynamic	situative	perf.	neg.	*ba*
$2CS_{categ}$ sa_{MORPH_neg}	descriptive	dynamic	relative	pres.	neg.	*bego*
$2CS_{categ}$ sa_{MORPH_neg}	descriptive	dynamic	relative	perf.	neg.	*bago*
$2CS_{categ}$ sa_{MORPH_neg}	associative	dynamic	situative	pres	neg.	*be*
$2CS_{categ}$ sa_{MORPH_neg}	associative	dynamic	situative	perf.	neg.	*ba*
$2CS_{categ}$ sa_{MORPH_neg}	associative	dynamic	relative	pres.	neg.	*bego*
$2CS_{categ}$ sa_{MORPH_neg}	associative	dynamic	relative	perf.	neg.	*bago*
$2CS_{categ}$ se_{MORPH_neg}	descriptive	dynamic	subjunctive	–	neg.	*be*
$2CS_{categ}$ se_{MORPH_neg}	descriptive	dynamic	habitual	–	neg.	*be*
$2CS_{categ}$ se_{MORPH_neg}	associative	stative	situative	pres.	pos.	*na*
$2CS_{categ}$ se_{MORPH_neg}	associative	stative	relative	pres.	neg.	*nago*
$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	descriptive	dynamic	indicative	fut.	neg.	*be*
$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	descriptive	dynamic	situative	fut.	neg.	*be*
$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	associative	dynamic	indicative	fut.	neg.	*be*
$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	associative	dynamic	situative	fut.	neg.	*be*
$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	descriptive	dynamic	relative	fut.	neg.	*bego*
$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	associative	dynamic	relative	fut.	neg.	*bego*

Table 4.27: Distribution of $2CS_{categ}$ followed by auxiliary constellations

preceding elements	**copulative**					**VCOP**
$2CS_{categ}$ be_{V_aux} $2CS_{categ}$	descriptive	stative	situative	perf.	pos.	*le*
$2CS_{categ}$ be_{V_aux} $2CS_{categ}$	descriptive	stative	situative	perf.	neg	se_{VCOP_neg}
$2CS_{categ}$ be_{V_aux} $2CS_{categ}$	associative	stative	situative	perf.	pos.	*(e)na*
$2CS_{categ}$ be_{V_aux} $2CS_{categ}$ se_{MORPH_neg}	associative	stative	situative	perf.	neg.	*(e)na*
$2CS_{categ}$ $bego_{V\_aux}$ $2CS_{categ}$	descriptive	stative	relative	perf.	pos.	*le*
$2CS_{categ}$ $bego_{V\_aux}$ $2CS_{categ}$	descriptive	stative	relative	perf.	neg	se_{VCOP_neg}
$2CS_{categ}$ $bego_{V\_aux}$ $2CS_{categ}$	associative	stative	relative	perf.	pos.	*na*
$2CS_{categ}$ $bego_{V\_aux}$ $2CS_{categ}$ se_{MORPH_neg}	associative	stative	relative	perf.	neg.	*na*

Table 4.28: Distribution of constellations beginning with $3CS_{categ}$

preceding elements	**copulative**					**VCOP**
$3CS_{categ}$	identifying	dynamic	consecutive	–	pos.	*ba*
$3CS_{categ}$	descriptive	dynamic	consecutive	–	pos.	*ba*

Table 4.29: Constellations beginning with (negation) morphemes

preceding elements			**copulative**			**VCOP**
$go_{\text{MORPH\_cp15}}$			infinitive	–	pos.	*ba*
$go_{\text{MORPH\_cp15}}$ $se_{\text{MORPH\_neg}}$			infinitive	–	neg.	*be*
$se_{\text{MORPH\_neg}}$			imperative	–	neg.	*be*
$ga_{\text{MORPH\_neg}}$	identifying	stative	indicative	pres.	neg.	VCOP_pers
$ga_{\text{MORPH\_neg}}$	identifying	stative	indicative	pres.	neg.	$se_{\text{VCOP\_neg}}$
$ga_{\text{MORPH\_neg}}$	descriptive	stative	indicative	pres.	neg.	$2\text{VCOP}_{\text{categ}}$
$ga_{\text{MORPH\_neg}}$ CSPCSN	identifying	dynamic	indicative	pres.	neg.	*be*
$ga_{\text{MORPH\_neg}}$ $2\text{CS}_{\text{categ}}$	descriptive	dynamic	indicative	pres.	neg.	*be*
$ga_{\text{MORPH\_neg}}$ $2\text{CS}_{\text{categ}}$	associative	dynamic	indicative	pres.	neg.	*be*
$ga_{\text{MORPH\_neg}}$ $2\text{CS}_{\text{categ}}$	associative	stative	indicative	pres.	neg.	*na*

4.6 Conclusions

This chapter has shown that there are indeed some generalisations possible when examining elements or element clusters that are part of a VP. Unlike the main verb constellations, where only few signalling elements or element groups were found (e.g. the negation *ga se* solely appearing in the indicative mood), analyses of copulatives are supported by a number of indicating features, e.g. the subject concords of the second set ($2CS_{categ}$) followed by the negation morphemes *sa* or *se* that never appear in the indicative. A left-to-right parsing strategy of Northern Sotho verbal constellations might accelerate the disambiguation process.

In chapter 3, we defined the morphosyntactic description of - amongst other elements - Northern Sotho verbal phrases described in literature. From there, a hierarchical system of POS constellations was developed that describes these verbal categories in a top-down manner: sets of Northern Sotho verbal constellations are distinguished for each mood and tense, and a number of overviews are provided each summarising them (one example of such a summary is Table 3.26 on page 121). A partial implementation of Northern Sotho constellations according to these grammar fragment definitions of chapter 3 will be described in chapter 5.

Chapter 5

Implementation of a grammar fragment

5.1 Introduction

So far, this study has been concerned with a general description of a fragment of a Northern Sotho grammar. Chapters 2 and 3 provide the units and information on how they form morphosyntactic constellations, while chapter 4 contains an introduction to parsing and some basics for a generalisation from these constellations as a first step to developing a linguistic model of the language.

In this chapter, some of the Northern Sotho "groups of consecutive words" (Jurafsky and Martin, 2000, p. 421) or constituents are described more formally in a context-free grammar (CFG, see paragraph 1.4.4). For sake of demonstration, we opted for an implementation of some core parts of the grammar fragment, i.e. the basic verbal phrase, the imperative and indicative constellations in the constraint-based Lexical-Functional Grammar formalism (LFG[1]). LFG has been successfully utilised to model morphosyntactic phenomena for a number of languages, Bresnan (2001, pp. 148 to 160) describes - amongst others - phenomena of the Malawi Bantu language Chichewa, a language with a number of phenomena similar to Northern Sotho. This section contains a brief introduction to theory (paragraph 5.1.1) and formalisms (paragraph 5.1.2) of LFG, including a description of the environment used for implementation and testing (paragraph 5.1.2.4) and, lastly, a description of the partial implementation itself (section 5.2).

[1]See e.g. `http://www-lfg.stanford.edu/lfg`

5.1.1 Lexical-Functional Grammar (LFG)

In this paragraph, ideas and background to the theory of Lexical-Functional Grammar (LFG) are introduced, as described by e.g. Asudeh and Toivonen (2009), who state that

> "LFG is a theory of generative grammar, in the sense of Chomsky (1957, 1965). The goal is to explain the native speaker's knowledge of language by specifying a grammar that models the speaker's knowledge explicitly and which is distinct from the computational mechanisms that constitute the language processor."

They refer to Kaplan and Bresnan (1982), who introduced this formal system to describe the grammar of a language. LFG, according to Kaplan and Bresnan (1982, p. 173), supports the expression and explanation of generalisations that concern syntactic issues. It manages information on two levels: the lexicon, where semantic arguments are mapped to grammatical functions appearing at sentence level, and the syntactic rules that identify these functions with "particular morphological and constituent structure configuations" (Kaplan and Bresnan, 1982, p. 174). A constituent and a functional structure form the result of analysis, which together represent the knowledge of the system about a specific sentence, i.e. a surface form. To map these structures to surface sentences, the structure needs to be sufficiently well-formed: the necessary requirements for this will be explained in greater detail below.

5.1.2 The LFG formalism

5.1.2.1 Representations: constituent structure and functional structure

LFG provides two levels of analytic representation, the constituent (or c-)structure, and the functional (or f-)structure, while (predicate-)argument structure is an input from the lexicon. Butt et al. (1999, p. 3) describe phrasal dominance and precedence relations as being encoded in c-structure, while f-structure encodes syntactic predicate argument structure. C-structure is made visible as a tree-structure, f-structure as an attribute-value matrix. The grammar files themselves are (roughly) divided into a lexicon and a rules-section. Figure 5.1, based on (Bresnan, 2001, (13), p. 19) demonstrates the parallel structures of LFG.

5.1.2.2 Predicate argument structure versus syntactic structure

In LFG, predicate argument structure is disassociated from syntactic structure. The predicate argument structure, e.g. a verb's valency, is assigned to the lexicon entry. The

rules-section on the other hand, contains certain constituents relating to grammatical functions.

The lexicon would therefore contain information such as the verb stem *bolela* '[to] speak' requiring a subject in its cotext when used intransitively. (bolela ⟨SUBJ⟩), the specific constituent that carries this function, however, is found in the rules-section, where a sentence is e.g. defined as an NP carrying the subject function and a VP (S → $_{\text{SUBJ}}$NP VP). Any NP described in the rules section, e.g. a single noun (NP → N), may then possibly fill this subject slot. If the sentence in example (80) were to be grammatically analysed , *bolela* will have *monna* assigned in the subject role[2].

(80) *monna*$_{\text{N01}}$ *o*$_{\text{1CS01}}$ *a*$_{\text{MORPH\_pres}}$ *bolela*$_{\text{Vitr}}$
man `subj-cl1` `pres` speak
'(a) man speaks'

In paragraph 1.4.4 (page 13), it was argued that attribute-value pairs (functional schemata or functions) and the unification principle should be utilised in order to reduce the number of rules necessary for developing a grammar. LFG extends this principle alongside two others: completeness and coherence, summarised as the three principles of well-formedness, cf. (Kaplan and Bresnan, 1982, p. 211 et seq.). According to the grammaticality condition

[2]A Northern Sotho grammar rules section would also describe instances where the NP is omitted and the subject concord acquires the subject function, or the imperative case, where the subject does not appear at all.

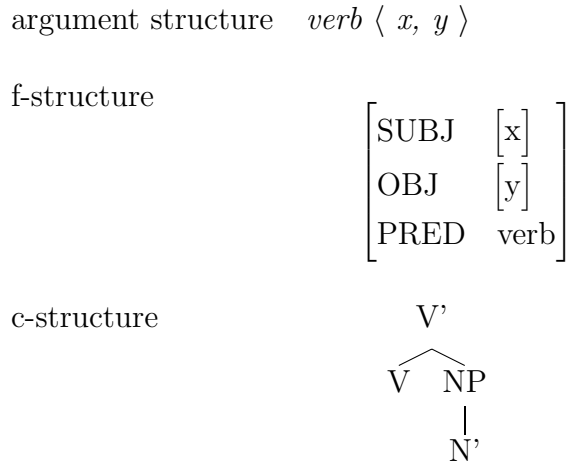


Figure 5.1: Parallel structures of LFG

(ibid. p. 212), "A string is grammatical only if it is assigned a complete and coherent f-structure."

Grammatical functions are assigned to constituents in the rules section in order to generate both a c(onstituent)-structure, and a f(unctional)-structure . The grammatical functions according to Bresnan (2001, p. 97 et seq.) are SUBJ(ect), OBJ(ect), OBL(ique), COMPL(ement) and ADJUNCT, TOP(ic) and FOC(us). Of these, ADJUNCT, TOP and FOC are non-argument functions and thus allow multiple instances, i.e the appearance of a zero element or other such items not governed by any other element. Any other grammatical function may only appear once per f-structure (functional uniqueness principle) and only if the sentence in question contains a unit that the lexicon states as being required (coherence principle). All of the grammatical functions assigned by the units contained in the sentence, i.e. the constituents that are governed by others, must however appear (completeness principle). The three LFG-principles may be summarised as follows (taken from Butt et al. (1999)):

- **Functional Uniqueness**: In a given f-structure, a particular attribute has a maximum of one value;
- **Completeness**: An f-structure is *locally complete* if and only if it contains all the governable grammatical functions that its predicate governs. An f-structure is *complete* if and only if it and all its subsidiary f-structures are locally complete;
- **Coherence**: An f-structure is *locally coherent* if and only if all the governable grammatical functions it contains are governed by a local predicate. An f-structure is *coherent* if and only if it and all its subsidiary f-structures are locally coherent.

5.1.2.3 An example analysis

Lexicon entries. The LFG formalism is best explained by an example analysis[3]. Therefore, we come back to the sentence *monna o reka apola* '(a) man buys (an) apple' mentioned in chapter 1 and begin with a description of the necessary LFG lexicon entries.

The singular noun *monna* 'man' is of noun class 1 and of the third person. These data are found in its lexicon entry (the meaning of ↑ will be explained in the paragraph on constructing c- and f-structure). The noun *apola* 'apple' is described similarly.

[3]We heavily rely on Wescoat (1989) when explaining the LFG formalism.

```
monna N *  (↑ PRED)='monna'
           (↑ CLASS)= 1
           (↑ NUM)= sg
           (↑ PERS) = 3.
apola N *  (↑ PRED)='apola'
           (↑ CLASS)= 9
           (↑ NUM)= sg
           (↑ PERS) = 3.
```

A subject concord is comparable to an inflectional prefix, it belongs to the verb, supplies agreement information (subject-verb agreement, cf. paragraph 2.4.2), and thus is not described with a PRED-value in the lexicon. The concord *o* occurs not only as a subject concord of the noun classes 1, 1a (1st set) and 3 (1st and 2nd set) and of the 2nd person singular (1st and 2nd set), it may also occur as a pronominal object concord of class 3. Person and number information on the subject are usually provided by the subject itself, however, as the subject concord may indeed acquire the subject's function when the respective NP is omitted, the appropriate grammar rule contains a disjunction describing these properties as well. The ambiguity of *o* is mirrored in the lexicon (the copulative use of *o* is not contained in the following lexicon entries):

```
o  1CS *  (↑ SUBJ CLASS)= 1 (↑ SUBJ NUM) = sg;
   1CS *  (↑ SUBJ CLASS)= 1a (↑ SUBJ NUM) = sg;
   1CS *  (↑ SUBJ PERS)= 2 (↑ SUBJ NUM) = sg;
   2CS *  (↑ SUBJ PERS)= 2 (↑ SUBJ NUM) = sg;
   1CS *  (↑ SUBJ CLASS)= 3 (↑ SUBJ PERS) = 3 (↑ SUBJ NUM) = sg;
   2CS *  (↑ SUBJ CLASS)= 3 (↑ SUBJ PERS) = 3 (↑ SUBJ NUM) = sg;
   CO *   (↑ CLASS)= 3 (↑ PERS) = 3 (↑ NUM) = sg.
```

The verb stem *reka* '[to] buy' is transitive, it therefore requires a subject and an object to appear in the sentence. As described in paragraph 2.7.2, certain verbal constellations may be solely identified by the verbal ending which need not be identical to the last letter(s) of the verb stem, like it is the case for *reka*. Verb stems like *re* '[to] say', however, appear as well. These behave syntactically like verbs ending in *–a*. Therefore, a specific attribute, Vend, was introduced which is added to each verb stem entry contained in the lexicon.

```
reka Vtr *  (↑ PRED)='reka ⟨(↑ SUBJ)(↑ OBJ)⟩'
            (↑ VEND)= a.
```

The rules section. LFG grammar rules look very similar to the CFG rules shown in paragraph 1.4.4, however, functional schemata may be added to their right side. Sentence (80) is a positive indicative of the present tense (as described in paragraph 3.2.5.1, page 97), it therefore requires the VIE to be annotated alongside[4]. The constraining equation '(↑VEND) =c a' means that in any such sentence, the attribute VEND must be *a*, i.e. only verb stems in the lexicon that have the attribute VEND defined and set to '−*a*' may appear in it, i.e. this equation must be satisfied by a f-structure (cf. (Kaplan and Bresnan, 1982, p. 207)). The rules may be constructed directly into an annotated tree (cf. Figure 5.2).

```
S    →  NP:   (↑ SUBJ) =↓;
        VP:   ↑=↓.

NP   →  N.

VP   →  VIE:  (↑ TNS-ASP MOOD)= indicative
              (↑ TNS-ASP TENSE)= pres
              (↑ TNS-ASP POL)= pos
        VBP:  (↑VEND) =c a.

VIE  →  1CS.
VBP  →  Vtr
        NP:   (↑ OBJ) =↓.
```

Constructing c- and f-structure When a sentence is analysed, the lexicon entries fill the predefined functions according to the well-formedness principles. The uniqueness principle avoids selecting the wrong lexicon entry of *o* ('CLASS 01' is defined by the lexicon entry of the subject noun *monna*). The completeness principle is fulfilled because all necessary arguments described by the lexicon entry of the predicate are present. The coherence principle is fulfilled because there are no constituents present that would unnecessarily add other arguments.

[4]The tense-aspect (TNS-ASP) attributes mood and tense appearing here are inspired by grammars developed in the pargram project, cf. section 5.1.2.4.

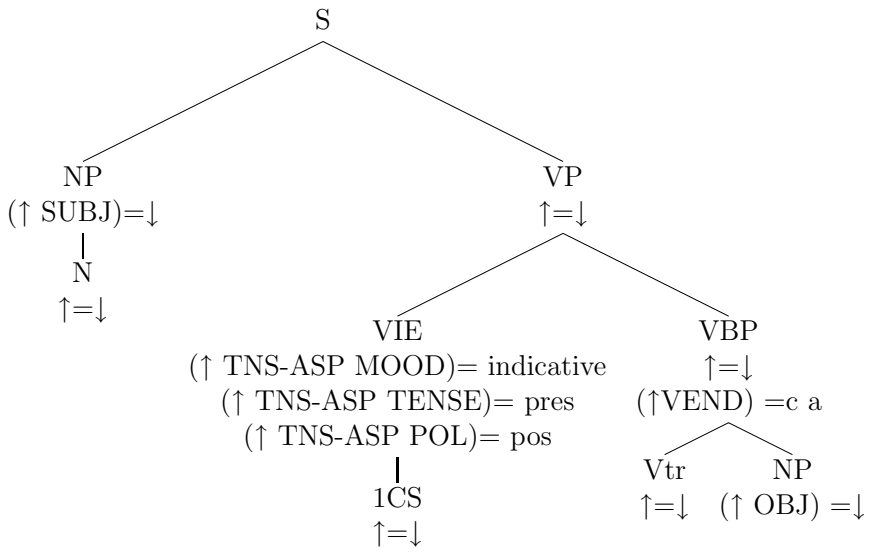


Figure 5.2: Basic sentence, consisting of a NP and a VP

The annotated c-structure shown in Figure 5.3 forms the basis for building an f-structure. This **instantiation** (cf. (Wescoat, 1989, p. 7)) "transforms the schemata into functional equations". Firstly, it is assumed that each node of the tree corresponds to one partial f-structure. Arbitrarily, one can give them numbers, as done in Figure 5.4. The arrows play an important role in identifying the referents, as they show correspondences between the f-structures. "↑" points to the tree node in which the respective node is contained, while ↓ points to the nodes contained in the respective node. As all nodes are instantiated in f-structures, "↑=↓" therefore means that the node's f-structure in which the node is contained is identical with the one the node's f-structure contains itself. In our example, $f2$=$f3$, $f4$=$f5$=$f6$=$f7$=$f8$, $f9$=$f10$.

The equation (↑ SUBJ)=↓ means that the node(s) below, i.e. the node(s) dominated by this node contains the subject of the node above. In our example, this means that the node found below the node $f2$ is the subject of S, i.e. ($f1$ SUBJ)=$f2$(=$f3$). Respectively, the object of S is found ($f7$(=$f4$=$f1$) OBJ)=$f9$(=$f10$), see Figure 5.5. Note that up-arrows in the lexicon entries of the words point to the respective node (and hence, the respective f-structure) where the word is to be contained in.

The set of all instantiated functional equations or **functional description** (cf. (Wescoat, 1989, p. 14)), see Table (5.1), is the only data for constructing the f-structures. For ease of

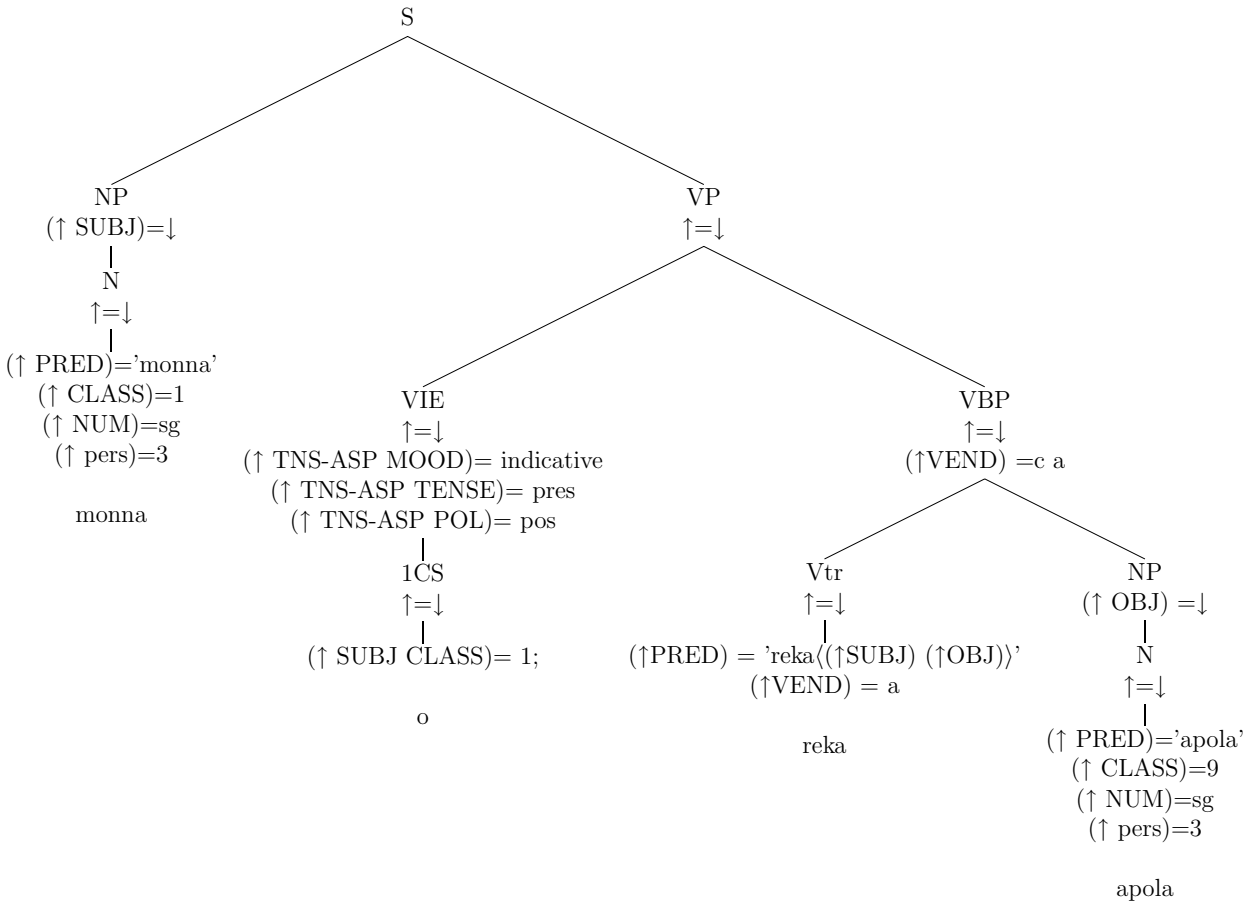


Figure 5.3: The c-structure of *monna o reka apola*

demonstration, we reduce the functional description by replacing the numbers of f-structures which are equated each with the lowest respective number, see Table (5.2). The reduction demonstrates that there are only four f-structures resulting from our analysis: *f*1 which contains a subject in f-structure *f*2 ((*f*1 SUBJ)=*f*2), an object in f-structure *f*9 ((*f*1 OBJ) =*f*9), and the f-structure TNS-ASP *f*4 created by the 2-part equations *f*1 TNS-ASP MOOD/TENSE/POL. All these data are contained in the thus simplified attribute-value matrix shown in Figure 5.6.

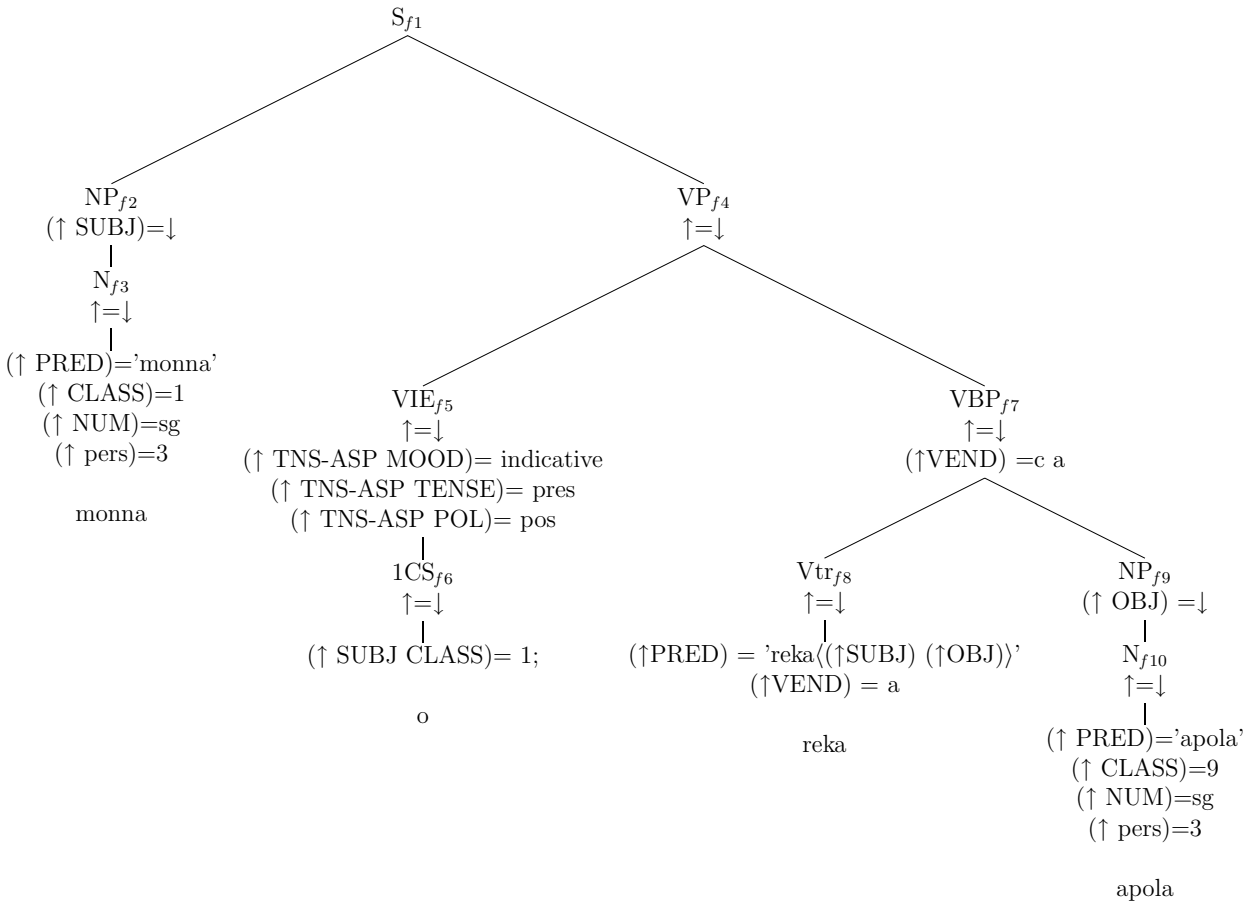


Figure 5.4: Numbered nodes at c-structure of *monna o reka apola*

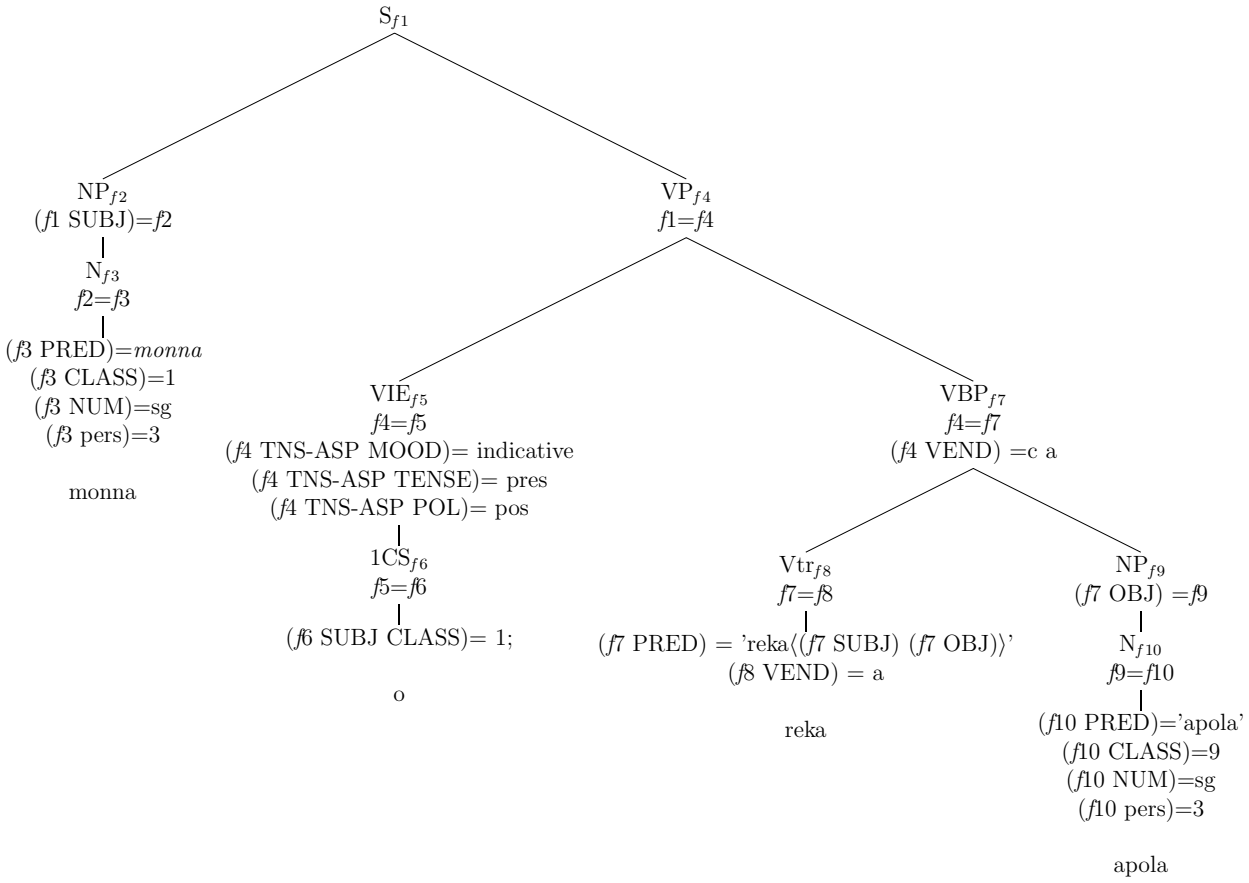


Figure 5.5: Instantiation of metavariables (↑ and ↓) of f-structures in the c-structure of *monna o reka apola*

a. (f1 SUBJ)=f2
b. f2=f3
c. (f3 PRED)=*monna*
d. (f3 CLASS)=1
e. (f3 NUM)=sg
f. (f3 pers)=3
g. f1=f4
h. f4=f5
i. (f4 TNS-ASP MOOD)= indicative
j. (f4 TNS-ASP TENSE)= pres
k. (f4 TNS-ASP POL)= pos
l. f5=f6
m. (f6 SUBJ CLASS)= 1
n. f4=f7
o. (f4 VEND) =c a
p. f7=f8
q. (f7 PRED) = 'reka⟨(f7 SUBJ) (f7 OBJ)⟩'
r. f8 VEND) = a
s. (f7 OBJ) =f9
t. f9=f10
u. (f10 PRED)='apola'
v. (f10 CLASS)=9
x. (f10 NUM)=sg
y. (f10 pers)=3

Table 5.1: Functional equations of *monna o reka apola*

a. (f1 SUBJ)=f2
c. (f2 PRED)=*monna*
d. (f2 CLASS)=1
e. (f2 NUM)=sg
f. (f2 pers)=3
i. (f1 TNS-ASP MOOD)= indicative
j. (f1 TNS-ASP TENSE)= pres
k. (f1 TNS-ASP POL)= pos
m. (f1 SUBJ CLASS)= 1
p. (f1 VEND) =c a
q. (f1 PRED) = 'reka⟨(f1 SUBJ) (f1 OBJ)⟩'
r. f1 VEND) = a
s. (f1 OBJ) =f9
u. (f9 PRED)='apola'
v. (f9 CLASS)=9
x. (f9 NUM)=sg
y. (f9 pers)=3

Table 5.2: Abbreviated version of functional equations of *monna o reka apola*

$$
\begin{bmatrix}
\text{PRED} & \text{'reka}\left\langle (f1\ \text{SUBJ})\ (f1\ \text{OBJ})\right\rangle\text{'} \\
\text{TNS-ASP} & \begin{bmatrix} \text{MOOD} & \text{indicative} \\ \text{TENSE} & \text{pres} \\ \text{POL} & \text{pos} \end{bmatrix} \\
\text{SUBJ} & \underset{f2}{\begin{bmatrix} \text{PRED} & \text{'monna'} \\ \text{CLASS} & 1 \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix}} \\
\text{OBJ} & \underset{f9}{\begin{bmatrix} \text{PRED} & \text{'apola'} \\ \text{CLASS} & 9 \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix}} \\
\underset{f1}{\text{VEND a}} &
\end{bmatrix}
$$

Figure 5.6: The f-structures of *monna o reka apola*.

5.1.2.4 The Pargram project

Concerning the implementation of the LFG formalism, we make use of the Xerox Linguistic Environment (XLE) provided by the Xerox Palo Alto Research Centre (PARC) in the framework of a research license. As foreseen in the LFG theory, XLE mainly provides two levels of descriptions: c-structure and f-structure. It also offers packed representations of ambiguous f-structures, the "fcharts". It may utilise a hand written full form lexicon or a morphological analyser. Rules are written in the LFG-formalism. Within the framework of the **Parallel Grammar** project (Pargram, cf. `http://www2.parc.com/isl/groups/nltt/pargram`) guidelines are provided that help to produce parallel representations of different languages. We will try to adhere to them from the start in order to provide the basics for the next step ahead (cf. section 6), the machine translation from Northern Sotho to English. Only a full form lexicon is however described, as there is ongoing work on automated morphological analysis of Northern Sotho (e.g. Anderson and Kotzé (2006)). At a later stage, such a morphological analyser may become part of a more substantial XLE implementation of Northern Sotho morphosyntax.

The implementation described in this chapter, adheres to the attributes suggested by the Pargram project, only few attributes were added, like, e.g. VEND to indicate the verbal ending of the verb stem in question. Another attribute added is TNS-ASP pol(arity), allowing for the two values positive and negative. TNS-ASP pol assists XLE in producing analyses containing the polarity information, hence making them consistent to the Northern Sotho modal system as described in Table (3.2) of paragraph 3.2.1.3 (Introduction to the modal system).

5.2 Implementation

5.2.1 The lexicon

The small, full form lexicon built for this study contains a number of verbs, nouns and other elements. In order to simplify the reading of the lexicon, some templates were defined[5]. Instead of having to add the attribute-value pair (↑ PERS) = 3 to all lexicon entries of nouns of the third person, one can define a template like e.g. PERS(_P) = (↑ PERS) =

[5]More information on XLE templates may be found at `http://www2.parc.com/isl/groups/nltt/xle/doc/walkthrough.html`.

_P. All respective lexicon entries may then have the person information added in a more human-readable form: '@(PERS 3)'. The same applies for other attributes occurring in this implementation, like noun class or number.

The lexicon, part 1 in (81) shows intransitive verb stems; the grammar fragment lexicon moreover contains saturated transitive, transitive, half saturated double transitive and double transitive verb stems (cf. (82)), as described in paragraph 3.2.1.6. Some past tense forms were added, all have the attribute VEND marked accordingly.

```
"intransitive:  [to] walk"
sepela    Vitr *  (↑ PRED) = 'sepela ⟨ (↑ SUBJ)⟩'  @(VEND a).
sepelang  Vitr *  (↑ PRED) = 'sepela ⟨ (↑ SUBJ)⟩'  @(VEND ang).
sepele    Vitr *  (↑ PRED) = 'sepela ⟨ (↑ SUBJ)⟩'  @(VEND e).
sepeleng  Vitr *  (↑ PRED) = 'sepela ⟨ (↑ SUBJ)⟩'  @(VEND eng).

"intransitive:  [to] speak"
bolela    Vitr *  (↑ PRED) = 'bolela ⟨ (↑ SUBJ)⟩'  @(VEND a).
bolelang  Vitr *  (↑ PRED) = 'bolela ⟨ (↑ SUBJ)⟩'  @(VEND ang).
bolele    Vitr *  (↑ PRED) = 'bolela ⟨ (↑ SUBJ)⟩'  @(VEND e).
boleleng  Vitr *  (↑ PRED) = 'bolela ⟨ (↑ SUBJ)⟩'  @(VEND eng).
boletše   Vitr *  (↑ PRED) = 'bolela ⟨ (↑ SUBJ)⟩'  @(VEND ile)
                  (↑ TNS-ASP TENSE) = past.

"intransitive:  [to] flee"
tšhaba    Vitr *  (↑ PRED) = 'tšhaba ⟨ (↑ SUBJ)⟩'  @(VEND a).
tšhabe    Vitr *  (↑ PRED) = 'tšhaba ⟨ (↑ SUBJ)⟩'  @(VEND e).
tšhabile  Vitr *  (↑ PRED) = 'tšhaba ⟨ (↑ SUBJ)⟩'  @(VEND ile)
                  (↑ TNS-ASP TENSE) = past.
```

(81) The intransitive verb stem entries of the lexicon (part 1)

In Northern Sotho, reflexivity is not explicitly expressed with a constituent carrying the object function, like in other languages ('oneself'), but encoded in the verb stem. The saturated transitive verb *ipshina* 'enjoy oneself' therefore does not require any such external object to appear. A correct analysis must however show that the object of the (originally transitive) verb is identical to the subject, like in (5.7) on page 240. According to the XLE "walkthrough"-page[6], however, the basic ontology of f-structures defined by Kaplan and Bresnan (1982) did not include the definition of an absent element carrying a grammatical

[6]The "walkthrough"-page gives comprehensive practical advice on how to develop a grammar in XLE, http://www2.parc.com/isl/groups/nltt/xle/doc/walkthrough.html.

function. An equation of the type (↑ SUBJ) = (↑ OBJ) is not acceptable, as the f-structure would become indeterminate. Therefore, in the lexicon, the subject's attributes like number, person and class are set equal to the object's respective attributes in the lexicon. The entry is additionally marked with the attribute 'TYPE' = refl(exive). In the rules section (cf. paragraph 5.2.2), objects of such verbs are declared as 'PRO(noun)'.

```
"monna o a ipshina."

    17[PRED    'ipshina<[1:monna], [3-OBJ:pro]>'      ]
    18[       1[PRED 'monna'                          ]
    14[SUBJ   2[CLASS 1, NUM sg, PERS 3               ]
    15[      29[                                      ]
    54[
     7[         [PRED 'pro'                           ]
     9[OBJ      [PRON [TYPE null]                     ]
     3[         [CLASS 1, NUM sg, PERS 3              ]
     6[
    51[TNS-ASP [FORM long, MOOD indicative, TENSE pres]
    44[CLAUSE-TYPE decl, TYPE refl, VEND a
    45[
    46[
    47[
    50[                                               ]
```

Figure 5.7: F-structure containing a saturated transitive verb *monna o a ipshina.* '(a) man enjoys himself.'

Lexicon entries describing a half-saturated double transitive verb stem, e.g. *mpha/mphe* '[to] give me', where the object concord of the first person singular (*N-*) is fused to the stem of *fa* '[to] give', are handled similarly, as shown in (82). As person, class and number of the verb's oblique is known, this data is contained in the entry.

```
"transitive or saturated transitive:  [to] enjoy (oneself)"
ipshina  Vsattr *     (↑ PRED) = 'ipshina ⟨ (↑ SUBJ) (↑ OBJ)⟩'  @(VEND a)
                      (↑ TYPE) = refl
                      (↑ OBJ CLASS) = (↑ SUBJ CLASS)
                      (↑ OBJ NUM) = (↑ SUBJ NUM)
                      (↑ OBJ PERS) = (↑ SUBJ PERS).

"transitive:  [to] buy"
reka     Vtr *        (↑ PRED)='reka ⟨(↑ SUBJ)(↑ OBJ)⟩'         @(VEND a).
rekang   Vtr *        (↑ PRED)='reka ⟨ (↑ SUBJ)(↑ OBJ)⟩'        @(VEND ang).
reke     Vtr *        (↑ PRED)='reka ⟨ (↑ SUBJ)(↑ OBJ)⟩'        @(VEND e).
rekeng   Vtr *        (↑ PRED)='reka ⟨ (↑ SUBJ)(↑ OBJ)⟩'        @(VEND eng).
rekile   Vtr *        (↑ PRED)='reka ⟨ (↑ SUBJ)(↑ OBJ)⟩'        @(VEND ile)
                      (↑ TNS-ASP TENSE) = past.

"transitive:  [to] close"
bula     Vtr *        (↑ PRED)='bula ⟨(↑ SUBJ)(↑ OBJ)⟩'         @(VEND a).
bule     Vtr *        (↑ PRED)='bula ⟨(↑ SUBJ)(↑ OBJ)⟩'         @(VEND e).
bulang   Vtr *        (↑ PRED)='bula ⟨(↑ SUBJ)(↑ OBJ)⟩'         @(VEND ang).
buleng   Vtr *        (↑ PRED)='bula ⟨(↑ SUBJ)(↑ OBJ)⟩'         @(VEND eng).

"half saturated double transitive:  [to] give me"
mpha     Vhsatdtr *   (↑ PRED)='mpha ⟨(↑ SUBJ)(↑ OBJ)(↑ OBL)⟩'  @(VEND a).
                      (↑ OBL NUM) = sg
                      (↑ OBL PERS) = 1
mphe     Vhsatdtr *   (↑ PRED)='mpha ⟨(↑ SUBJ)(↑ OBJ)(↑ OBL)⟩'  @(VEND e).
                      (↑ OBL NUM) = sg
                      (↑ OBL PERS) = 1

"double transitive:  [to] give"
fa       Vdtr *       (↑ PRED)='fa ⟨(↑ SUBJ)(↑ OBL)(↑ OBJ)⟩'    @(VEND a).
fe       Vdtr *       (↑ PRED)='fa ⟨(↑ SUBJ)(↑ OBL)(↑ OBJ)⟩'    @(VEND e).
```

(82) Other verb stem entries of the lexicon (part 2)

The lexicon also contains a number of nouns of all noun classes appearing in Northern Sotho, cf. (83) . Each noun is entered with its noun class, person and number. Note that the quoted translations in the last column of (83) are comments (marked by double quotes in XLE).

```
tate       N *   (↑ PRED) = 'tate' @(CLASS 1a) @(PERS 3) @(NUM sg).      "father"
monna      N *   (↑ PRED) = 'monna' @(CLASS 1) @(PERS 3) @(NUM sg).      "man"
banna      N *   (↑ PRED) = 'banna' @(CLASS 2) @(PERS 3) @(NUM pl).      "men"
bana       N*    (↑ PRED) = 'bana' @(CLASS 2) @(PERS 3) @(NUM pl).       "children"
motato     N *   (↑ PRED) = 'motato' @(CLASS 3) @(PERS 3) @(NUM sg).     "wire"
mothepa    N *   (↑ PRED) = 'mothepa' @(CLASS 3) @(PERS 3) @(NUM sg).    "young woman"
mmutla     N *   (↑ PRED) = 'mmutla' @(CLASS 3) @(PERS 3) @(NUM sg).     "hare"
merako     N *   (↑ PRED) = 'merako' @(CLASS 4) @(PERS 3) @(NUM pl).     "stone wall"
lemati     N *   (↑ PRED) = 'lemati' @(CLASS 5) @(PERS 3) @(NUM sg).     "door"
lengwalo   N *   (↑ PRED) = 'lengwalo' @(CLASS 5) @(PERS 3) @(NUM sg).   "letter"
lesogana   N *   (↑ PRED) = 'lesogana' @(CLASS 5) @(PERS 3) @(NUM sg).   "young man"
mangwalo   N *   (↑ PRED) = 'mangwalo' @(CLASS 6) @(PERS 3) @(NUM pl).   "letters"
masogana   N *   (↑ PRED) = 'masogana' @(CLASS 6) @(PERS 3) @(NUM sg).   "young men"
meetse     N *   (↑ PRED) = 'meetse' @(CLASS 6) @(PERS 3).               "water"
malekere   N *   (↑ PRED) = 'malekere' @(CLASS 6) @(PERS 3)@(NUM pl) .   "sweets"
sepetlele  N *   (↑ PRED) = 'sepetlele' @(CLASS 7) @(PERS 3) @(NUM sg).  "hospital"
ditho      N *   (↑ PRED) = 'ditho' @(CLASS 8) @(PERS 3) @(NUM pl).      "limbs"
puku       N *   (↑ PRED) = 'puku' @(CLASS 9) @(PERS 3) @(NUM sg).       "book"
dipuku     N *   (↑ PRED) = 'dipuku' @(CLASS 10) @(PERS 3) @(NUM pl).    "books"
```

(83) Noun entries of the lexicon

Next, some of the concords and morphemes and, finally, punctuation are listed in (84) to (87) . Note that object concords occur in a pronominal function replacing an omitted (or topicalised) object (cf. e.g. paragraph 3.2.1.1), therefore these entries have a PRED-value defined. Not all of these items' possible parts of speech are entered, *go*, for example is fourteenfold ambiguous[7]. Here, only six of the possible parts of speech are listed. Again, templates are used for a clearer overview, e.g. @C6P3Npl, an abbreviated entry of (↑ SUBJ CLASS) = 6 (↑ SUBJ PERS) = 3 (↑ SUBJ NUM) = pl. It is not possible to foresee the value of the attribute 'person' for entries of class 1, hence this attribute is not mentioned for these entries.

In paragraph 3.2.5.1, Table 3.15 describes the long form of the indicative. Only this constellation contains the present tense morpheme, namely *a*. As this form may only appear if the clause ends after the verb stem, the grammar rules described in paragraph 5.2.4.2 will make use of the attribute TNS-ASP FORM (with the value 'long').

[7] *go* seems to be the most ambiguous linguistic unit of Northern Sotho: it may occur as an object concord of class 15, object concord of the locative classes, object concord of the second person singular, subject concord of class 15 (set 1 and set 2), indefinite subject concord (set 1 and set 2), subject concord of the locative classes (set 1 and set 2), class prefix of class 15, locative particle, copulative indicating either an indefinite subject or a subject of class 15 or a locative subject.

```
"a is a subject concord of classes 1,6, an object concord of class 6"
"and a present or past tense morpheme."
a     2CS *     @C1Nsg;
      3CS *     @C1Nsg;
      1CS *     @C6NplP3;
      2CS *     @C6NplP3;
      3CS *     @C6NplP3;
      CO *      (↑ PRED) = 'pro' @(CLASS 6) @(NUM pl) @(PERS 3) ;
      MORPH *   (↑ TNS-ASP FORM) = long
                (↑ TNS-ASP TENSE) = pres;
      MORPH *   (↑ TNS-ASP TENSE) = past.
"ba is a subject concord of class 2, and an object concord of class 2"
ba    1CS *     @C2NplP3;
      2CS *     @C2NplP3;
      3CS *     @C2NplP3;
      CO *      (↑ PRED) = 'pro' @(CLASS 2) @(NUM pl) @(PERS 3) .
"bja is a subject concord of class 14 (3rd set)"
bja   3CS *     @C14NsgP3.
"bo is a subject concord of class 14, and an object concord of class 14"'
bo    1CS *     @C14NsgP3;
      2CS *     @C14NsgP3;
      CO *      (↑ PRED) = 'pro' @(CLASS 14) @(NUM sg) @(PERS 3).
"di is a subject and an object concord of classes 8 and 10"
di    1CS *     @C8NplP3;
      2CS *     @C8NplP3;
      1CS *     @C10NplP3;
      2CS *     @C10NplP3;
      CO *      (↑ PRED) = 'pro' @(CLASS 8) @(NUM pl) @(PERS 3) ;
      CO *      (↑ PRED) = 'pro' @(CLASS 10) @(NUM pl) @(PERS 3) .
```

(84) Concords: entries of the lexicon (part 1)

```
"e is a subject concord of classes 4 and 9, a neutral subject concord,"
"and an object concord of classes 4 and 9"
e    1CS *       @C4NplP3;
     2CS *       @C4NplP3;
     1CS *       @C9NsgP3;
     2CS *       @C9NsgP3;
     1CSNEUT *   (↑ SUBJ PERS) = 3;
     2CSNEUT *   (↑ SUBJ PERS) = 3;
     CO *        (↑ PRED) = 'pro' @(CLASS 4) @(NUM) = pl @(PERS 3) ;
     CO *        (↑ PRED) = 'pro' @(CLASS 9) @(NUM) = sg @(PERS 3) .
"ga is a negation morpheme, occurring alone and"
"as the first part of the negation cluster ga se"
ga   MORPH *     (↑ NEG) = ga (↑ TNS-ASP POL) = neg;
     MORPH *     (↑ NEG1) = ga (↑ TNS-ASP POL) = neg.
"go is a subject and an object concord of classes 15 and LOC"
"Note that a number of parts of speech of go are not listed here"
"Class 15 is the infinitite class, no person or number"
"Class LOC contains locatives which often are used adverbially"
go   1CS *       (↑ SUBJ CLASS) = 15;
     2CS *       (↑ SUBJ CLASS) = 15;
     1CS *       (↑ SUBJ CLASS) = LOC;
     2CS *       (↑ SUBJ CLASS) = LOC;
     CO *        (↑ PRED) = 'pro' @(CLASS 15);
     CO *        (↑ PRED) = 'pro' @(CLASS LOC).
"gwa is a subject concord of classes 15 and LOC (3rd set)"
gwa  3CS *       (↑ SUBJ CLASS) = 15;
     3CS *       (↑ SUBJ CLASS) = LOC.
"ka is the subject concord of the 3rd set of the 1st person"
"and a potential morpheme"
ka   3CS *       (↑ SUBJ PERS) = 1 (↑ SUBJ NUM) = sg;
     MORPHpot *  .
"la is a subject concord of class 5 and one of the 2nd person plural"
la   3CS *       @C5NsgP3;
     3CS *       (↑ SUBJ PERS) = 2 (↑ SUBJ NUM) = pl.
```

(85) Concords: entries of the lexicon (part 2)

```
le    1CS *       @C5NsgP3;
      2CS *       @C5NsgP3;
      1CS *       (↑ SUBJ PERS) = 2 (↑ SUBJ NUM) = pl;
      2CS *       (↑ SUBJ PERS) = 2 (↑ SUBJ NUM) = pl;
      CO *        (↑ PRED) = 'pro' @(CLASS 5) @(NUM sg) @(PERS 3).

o     1CS *       @C1Nsg;
      1CS *       @C1aNsgP3;
      1CS *       (↑ SUBJ PERS)= 2 (↑ SUBJ NUM) = sg;
      2CS *       (↑ SUBJ PERS)= 2 (↑ SUBJ NUM) = sg;
      1CS *       @C3NsgP3;
      2CS *       @C3NsgP3;
      CO *        (↑ PRED) = 'pro' @(CLASS 3) @(NUM sg) @(PERS 3).

mo    CO *        (↑ PRED) = 'pro' @(CLASS 1) @(NUM sg) @(PERS 3).

ra    3CS *       (↑ SUBJ PERS) = 1 (↑ SUBJ NUM) = pl.

re    1CS *       (↑ SUBJ PERS) = 1 (↑ SUBJ NUM) = pl;
      2CS *       (↑ SUBJ PERS) = 1 (↑ SUBJ NUM) = pl.

sa    3CS *       @C7NsgP3.

se    1CS *       @C7NsgP3;
      2CS *       @C7NsgP3;
      CO *        @C7NsgP3;
      MORPH *     (↑ NEG) = se (↑ TNS-ASP POL) = neg.
      MORPH *     (↑ NEG2) = se (↑ TNS-ASP POL) = neg.

tša   3CS *       @C8NplP3;
      3CS *       @C10NplP3.

wa    3CS *       @C3NsgP3;
      3CS *       (↑ SUBJ PERS) = 2 (↑ SUBJ NUM) = sg.

ya    3CS *       @C4NplP3;
      @C9NplP3;
      3CSNEUT *   (↑ SUBJ PERS) = 3.
```

(86) Concords: entries in the lexicon (part 3)

```
"future tense morphemes"
tlo   MORPH *       (↑ TNS-ASP TENSE) = fut.
tla   MORPH *       (↑ TNS-ASP TENSE) = fut.
"Punctuation"
.     PERIOD *.
,     COMMA *.
!     EXCLMARK *.
?     QUEMARK *.
```

(87) future tense morphemes and punctuation in the lexicon

5.2.2 The basic verbal phrase (VBP)

Table 5.3: Constellations forming the VBP

description	**VBP**			
	pos-1	**pos 0**	**pos+1**	**pos+2**
VBP		$\mathbf{V}^{\text{itr}}$		
VBP		$\mathbf{V}^{\text{tr}}$		$_{\text{OBJ}}\mathbf{NP}$
VBP		$\mathbf{V}^{\text{dtr}}$	$_{\text{OBJ}-\text{TH}}\mathbf{NP}$	$_{\text{OBJ}}\mathbf{NP}$
$\mathbf{VBP}^{\text{p}}$	$_{\text{OBJ}}\mathbf{CO}_{\text{categ}}$	$\mathbf{V}^{\text{tr}}$		
VBP	$_{\text{OBJ}-\text{TH}}\mathbf{CO}_{\text{categ}}$	$\mathbf{V}^{\text{dtr}}$		$_{\text{OBJ}}\mathbf{NP}$

In paragraph 3.2.2, the core element of all verbal phrases, the basic verbal phrase was defined[8]. Butt et al. (1999, p. 50) state that in English grammar, secondary objects are to be called OBL (not OBJ2 or OBJind), because they cannot undergo passivization. The XLE implemenation of an English grammar available from XEROX PARC however makes use of the argument OBJ-TH(ematic) for secondary objects subcategorised by double transitive verbs, e.g. 'I gave **him** a book' while other secondary objects, like the prepositional phrase in e.g. 'I am looking **for the book**' remain labeled as OBL. When preparing for a machine translation, similarity to the English grammar should be aimed for, the indirect object OBJind is hence renamed to OBJ-TH in Table 5.3.

Our VBP coding in XLE contains a number of options contained in braces ({,}); each

[8]As a reminder the contents of Table 3.14 are repeated in Table 5.3. Note that (as described in paragraph 3.2.5.1) according to our understanding, a sentence border should appear following the intransitive basic verbal phrase and the transitive basic verbal phrase where the object appears as an object concord.

option is separated from the other by the disjunction "|". Some verbal endings indicate a certain mood which is thus entered right away (the functional uniqueness principle then prohibits other analyses than those indicated).

The grammar rule describing the VBP can be split into three parts: verb stems subcategorising no arguments, verb stems subcategorising one and, lastly, two arguments, here described as NP[9]. As described in paragraph 3.2.2, the first case of a VBP processes the intransitive verbs. This VBP option firstly distinguishes the two cases where no object NP appears: both solely consist of the verb stem. Such a VBP may either consist of an intransitive verb or a saturated transitive verb. For the latter case, a pronominal object is defined ((↑OBJ PRED) = 'pro', (↑OBJ PRON TYPE)=null). This object will be added to the f-structure (see Figure 5.7), its attributes are stored with the verb entry in the lexicon. The verbal endings *-ang* or *eng* prescribe an imperative, in order to allow processing of the other verbal endings, these must be mentioned, too (the uniqueness principle would otherwise only allow for *-ang* or *eng* to appear with the attribute VEND).

```
VBP --> { { Vitr   : { (^ VEND) = ang  (^ TNS-ASP MOOD)= imperative
                     | (^ VEND) = eng  (^ TNS-ASP MOOD)= imperative
                     | (^ VEND) = a
                     | (^ VEND) = e
                     | (^ VEND) = ile
                     }
          | Vsattr : (^ OBJ PRED)= 'pro'
                     (^ OBJ PRON TYPE)= null
                     { (^ VEND) = ang  (^ TNS-ASP MOOD)= imperative
                     | (^ VEND) = eng  (^ TNS-ASP MOOD)= imperative
                     | (^ VEND) = a
                     | (^ VEND) = e
                     | (^ VEND) = ile
                     }
          }
...
```

The rule continues its description of the VBP with the two cases where one object is available, that of a transitive verb and that of a half saturated double transitive verb[10]. These are described similarly to the first two cases. The next item to be described is the case of an object concord preceding the transitive verb stem. The object concord has the object function assigned.

[9]At a later stage of the project, subcategorised clauses and adverbial attributes will be added.

[10]We will refer again to the line "e: (↑ TNS-ASP FORM)~= long" in paragraph 5.2.4.2.

```
...
        | e : (^ TNS-ASP FORM) ~= long;
          { Vtr    : { (^ VEND) = ang  (^ TNS-ASP MOOD)= imperative
                     | (^ VEND) = eng  (^ TNS-ASP MOOD)= imperative
                     | (^ VEND) = a
                     | (^ VEND) = e
                     | (^ VEND) = ile
                    }
          | Vhsatdtr: (^ OBJ-TH PRED)= 'pro'
                      (^ OBJ-TH PRON TYPE)= null
                      { (^ VEND) = ang  (^ TNS-ASP MOOD)= imperative
                      | (^ VEND) = eng  (^ TNS-ASP MOOD)= imperative
                      | (^ VEND) = a
                      | (^ VEND) = e
                      | (^ VEND) = ile
                       }
          }
          NP: (^ OBJ)=!

          | CO: (^ OBJ)=!;

          Vtr :  { (^ VEND) = ang  (^ TNS-ASP MOOD)= imperative
                 | (^ VEND) = eng  (^ TNS-ASP MOOD)= imperative
                 | (^ VEND) = a
                 | (^ VEND) = e
                 | (^ VEND) = ile
                 }
...
```

The third part of the VBP-rule describes the two cases of the double transitive verbs with two object NPs or with object concord and object NP. For the first case, the order of objects is predefined as thematic (indirect) object preceding the direct object (cf. paragraph 3.2.1.1). As the object concord may in theory stand in for each of them (though a preference for the indirect object has been observed), the latter case would have to be divided into two further options. Note that such a rule leads to a double analysis of all double transitive verbs occurring with an object concord. In this implementation, however, only the case expected to occur more frequently is described: the object concord replacing an indirect (thematic) object.

```
...
        | e : (^ TNS-ASP FORM) ~= long;
          Vdtr:  { (^ VEND) = ang  (^ TNS-ASP MOOD)= imperative
```

```
            | (^ VEND) = eng  (^ TNS-ASP MOOD)= imperative
            | (^ VEND) = a
            | (^ VEND) = e
            | (^ VEND) = ile
            };
      NP: (^ OBJ-TH)=!;
      NP: (^ OBJ)=!

    | e : (^ TNS-ASP FORM) ~= long;
      CO: (^ OBJ-TH)=!;
      Vdtr
      NP: (^ OBJ)=!
    }.
```

In this paragraph, we have implemented a set of optional rules for the VBP that mirror the summarised VBP in Table 3.14. Paragraphs 5.2.3 and 5.2.4 show how the VBP appears in the rules defining the verbal phrases.

5.2.3 The imperative

The imperative is the simplest form of a verbal phrase (cf. paragraph 3.2.3). The non-predicative 'VPimp' either contains solely a VBP for the imperative positive, or additionally an imperative VIE (VIEimp containing the negation morpheme *se*) used for the negative form of this VP for which the condition is set that the verb stem must be described in the lexicon as ending in '*–e*' or '*–eng*' (parameter "Vend"). Sentences 88 to 92 are to be analysed by the grammar.

(88) *Bolela !*
speak !
'speak!'

(89) *Se bolele !*
speak !
'Do not speak!'

(90) *Bulang lemati !*
Close door !
'Close (the) door!'

(91) *Efa monna puku !*
Give man book !
'Give (a) man (a) book!'

(92) *Se fe monna puku !*
neg give man book !
'Do not give (a) man (a) book!'

The rule describing the positive imperative VP (VPimp) extends the VBP only to the extent that the number of the subject (↑SUBJ NUM) is described. The suffix *-ng* is used when several people are addressed (cf. paragraphs 2.7.2 and 3.2.3), therefore the respective verb stems are marked with (↑SUBJ NUM)=pl. The negated form of the imperative additionally uses the imperative verbal inflectional phrase (VIEimp) which solely contains the negation *se*. Note that the mood's polarity (TNS-ASP POL) is set to "neg(ative)" by a respective entry in the lexicon describing *se*.

```
VPimp --> { VBP:(^ TNS-ASP POL) = pos
                { (^ VEND) =c a (^ SUBJ NUM) = sg
                | (^ VEND) =c e (^ SUBJ NUM) = sg
                | (^ VEND) =c ang (^ SUBJ NUM) = pl
                | (^ VEND) =c eng (^ SUBJ NUM) = pl
                }
          | VIEimp: ^=!;
            VBP : { (^ VEND) =c e (^ SUBJ NUM) = sg
                  | (^ VEND) =c eng (^ SUBJ NUM) = pl
                  }
          }.

VIEimp -->
          MORPH: (^ NEG) =c se.
```

In the current implementation, imperative and indicative VPs are described, cf. paragraph 5.2.4). The imperative VP is contained in S. An overall root node, "Root", is defined, containing S and possibly punctuation (the parenthesis "()" occuring in the rule defining ROOT indicate optionality of the described punctuation) .

```
VP -->
   { VPimp: (^ TNS-ASP MOOD)= imperative " imperative"
   | VPpred
   }.

S -->
    VP: (^ SUBJ PRED)='pro'                " imperative "
        (^ SUBJ PRON TYPE)= null
```

```
        (^ SUBJ PERS)= 2
        (^ TNS-ASP MOOD)=c imperative.
ROOT -->
       S
       ({PERIOD
        |QUEMARK: (^ TNS-ASP MOOD)= question
        |EXCLMARK
        }).
```

Figures 5.8 and 5.9 show a XLE-analysis of a positive imperative, making use of the VBP definition of the intransitive verb in (88). The following figures, 5.10 and 5.11 demonstrate the analysis of a transitive verb of example (90). Next, figures 5.12 and 5.13 demonstrate a double transitive VBP contained in the imperative (91), and its negated form in (92), cf. figures 5.14 and 5.15.

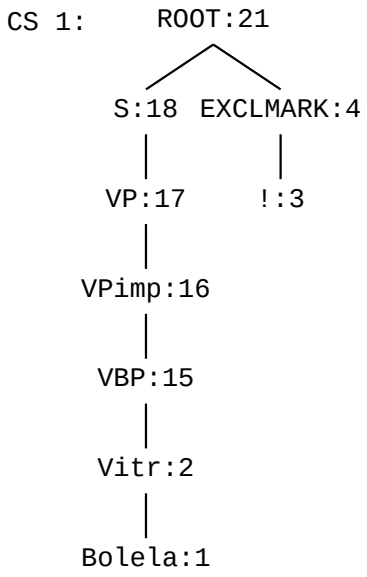


Figure 5.8: C-structure of a positive imperative intransitive *Bolela!* 'Speak!'

```
"Bolela!"

 3 PRED     'bolela<[1-SUBJ:null_pro]>'
 4          [PRED 'null_pro'          ]
 1 SUBJ     [PRON [TYPE null]         ]
 2          [NUM sg, PERS 2           ]
15
16 TNS-ASP  [MOOD imperative, POL pos]
17
18 VEND     a
21
```

Figure 5.9: F-structure of a positive imperative intransitive *Bolela!* 'Speak!'

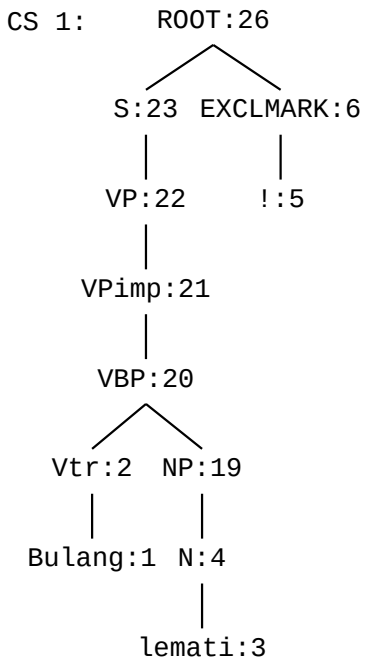


Figure 5.10: C-structure of a positive imperative transitive *Bulang lemati!* 'Close (the) door!'

```
"Bulang lemati!"

 ⎡PRED    'bula<[1-SUBJ:null_pro], [3:lemati]>'⎤
 ⎢        ⎡PRED 'null_pro'                     ⎤
5⎢SUBJ    ⎢PRON [TYPE null]                    ⎥
6⎢        ⎣NUM pl, PERS 2                      ⎦
1⎢
2⎢          3⎡PRED 'lemati'                    ⎤
20⎢OBJ      4⎢CLASS 5, NUM sg, PERS 3          ⎥
21⎢        19⎣                                ⎦
22⎢
23⎢TNS-ASP  [MOOD imperative, POL pos]
26⎣VEND     ang                                ⎦
```

Figure 5.11: F-structure of a positive imperative transitive *Bulang lemati!* 'Close (the) door!'

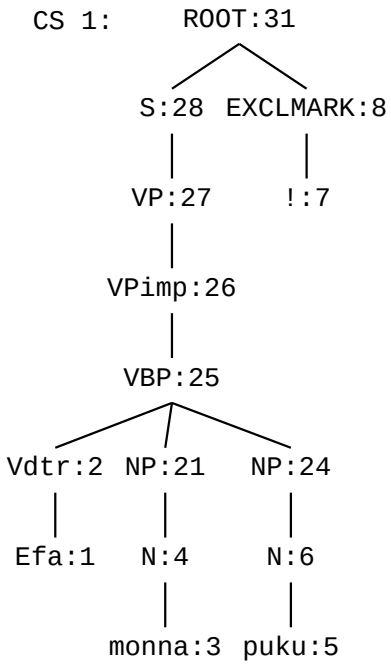


Figure 5.12: C-structure of a positive imperative double transitive *Efa monna puku!* 'Give (a) man (a) book!'

```
"Efa monna puku!"

⎡PRED     'fa<[1-SUBJ:null_pro], [3:monna], [5:puku]>'⎤
⎢         ⎡PRED 'null_pro'                            ⎤
⎢SUBJ     ⎢PRON ⎡TYPE null⎤                           ⎥
⎢         ⎣     ⎣NUM sg, PERS 2 ⎦                     ⎦
⎢       3 ⎡PRED 'monna'                               ⎤
⎢OBJ-TH 4 ⎢CLASS 1, NUM sg, PERS 3                    ⎥
⎢      21 ⎣                                           ⎦
⎢       5 ⎡PRED 'puku'                                ⎤
⎢OBJ    6 ⎢CLASS 9, NUM sg, PERS 3                    ⎥
⎢      24 ⎣                                           ⎦
⎢TNS-ASP  ⎡MOOD imperative, POL pos⎤
⎣VEND     ⎣a                       ⎦
```

Figure 5.13: F-structure of a positive imperative double transitive *Efa monna puku!* 'Give (a) man (a) book!'

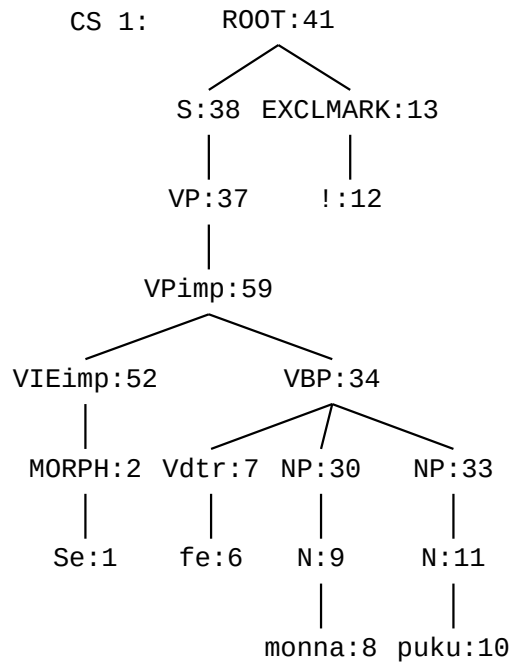


Figure 5.14: C-structure of a negated imperative double transitive *Se fe monna puku!* 'Do not give (a) man (a) book!'

```
"Se fe monna puku!"

    [PRED     'fa<[1-SUBJ:null_pro], [8:monna], [10:puku]>'
               [PRED 'null_pro'                              ]
12 SUBJ        [PRON [TYPE null]                             ]
13             [NUM sg, PERS 2                               ]
 6
 7           8 [PRED 'monna'                                 ]
34 OBJ-TH    9 [CLASS 1, NUM sg, PERS 3                      ]
 1          30 [                                            ]
 2          10 [PRED 'puku'                                 ]
52 OBJ      11 [CLASS 9, NUM sg, PERS 3                     ]
59         33  [                                            ]
37
38 TNS-ASP    [MOOD imperative, POL neg                     ]
41 [NEG se, VEND e                                          ]
```

Figure 5.15: F-structure of a negated imperative double transitive *Se fe monna puku!* 'Do not give (a) man (a) book!'

5.2.4 The predicative independent indicating mood: the indicative

5.2.4.1 General rules for the indicative

The indicative is a predicative mood, it therefore always contains (at least) a subject concord which is found in the VIE. The VPind is defined to be of the clause type declarative ((↑ CLAUSE-TYPE) = decl) in order to conform with English grammar terminology.

```
VPpred -->
       VPind: (^ CLAUSE-TYPE) = decl
              (^ TNS-ASP MOOD)= indicative.

VPind -->
       VIE
       VBP.
```

5.2.4.2 The imperfect indicative

The imperfect indicative may consist of the positive short present tense and long present tense, or its negated form. In paragraph 3.2.5.1, where this indicative is described, we modified the VBP in order to indicate that its long form ends with the verb stem. The XLE parsing algorithm however does not need such markers as the respective rule does not allow any other use of *a*. This case is determined by the constraint (↑ TNS-ASP FORM)=c long. Secondly, the VBPs where one or more NPs follow the verb stem, are exluded from the long form with the constraint "e : (↑ TNS-ASP FORM) ∼= long" ("∼=" is read as "not equal").

Furthermore, a continuation of the rule is indicated with "...", as the perfect tense and the future forms will be described in the paragraphs (5.2.4.3 and 5.2.4.4). There is no element in the short present tense form that indicates that the constellation is of the present tense. The subject concord of the first set, which is the only element of this constellation, can occur with other tenses as well (e.g. the perfect positive, cf. Table 4.7 in paragraph 4.4 on page 210 for an overview). Such missing indication is a quite regular phenomenon in Northern Sotho, as usually the VIE as a whole is seen as indicating information, like, e.g., tense. Therefore, XLE rules often contain symbolic empty elements ("e"), these are inserted in the grammar wherever a surface marker cannot be identified. These empty elements hence contain constraints that cannot be assigned to single elements.

```
"Verbal Inflectional Elements : VIE"
VIE -->   " short present tense form"
       {  1CS : (^ TNS-ASP FORM) = short;
          e : (^ TNS-ASP TENSE) = pres
              (^ VEND) =c a
          "long form "
       | 1CS
          MORPH: (^ TNS-ASP FORM) =c long
                 (^ TNS-ASP TENSE) =c pres;
          e : (^ VEND) =c a
          " negated present tense"
       |  MORPH:  (^ NEG) =c ga;
          e : (^ TNS-ASP TENSE) = pres
              (^ VEND) =c e;
          2CS
       ...
       }.
```

Sentences where the subject is not present often have several readings, as the subject concords are often ambiguous. Here, XLE offers a packed f-structure that indicates all possible readings, e.g. *o a bolela* '(s)he/it speaks' or 'you speak', where the subject concord *o* may refer to a omitted noun of the noun classes 1, 1a or class 3, or to a 2nd person singular, cf. Figure 5.16. The c-structure, however, is not affected by this ambiguity (cf. Figure 5.17).

5.2.4.3 The perfect indicative

The perfect indicative occurs in five different constellations, a positive and four negated forms (cf. paragraph 3.2.5.2). The VIE-rule of the previous paragraph is hence extended respectively.

```
"Verbal Inflectional Elements : VIE"
VIE -->
       { ...
          " perfect tense "
       |  1CS
          e : (^ VEND) =c ile
        " negated perfect tense constellations"
       |  MORPH : (^ NEG1) =c ga;
          MORPH : (^ NEG2) =c se;
          e : (^ TNS-ASP TENSE) = past
              (^ VEND) =c a;
```

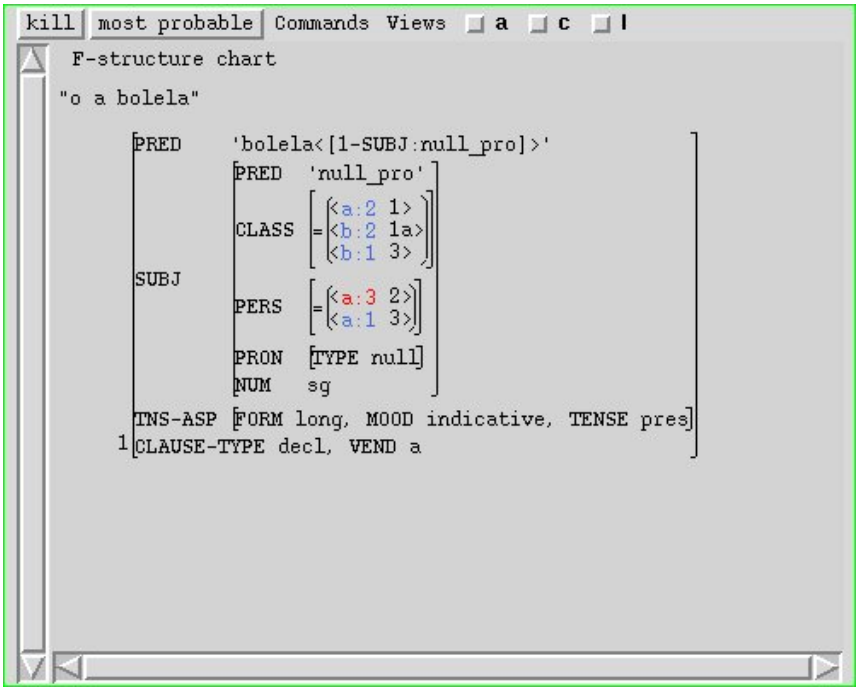


Figure 5.16: Packed f-structure of a positive indicative, present tense, subject not present *o a bolela.* '(s)he/it speaks / you speak.'

```
   3CS
|  MORPH : (^ NEG1) =c ga;
   MORPH : (^ NEG2) =c se;
   e : (^ TNS-ASP TENSE) = past
       (^ VEND) =c e;
   2CS
|  MORPH : (^ NEG) =c ga;
   e : (^ TNS-ASP TENSE) = past
       (^ VEND) =c a;
   3CS
|  MORPH : (^ NEG) =c ga;
   1CS
   MORPH:  (^ TNS-ASP TENSE) = past;
   e : (^ VEND) =c a;
...
}.
```

The following c- and f-structures show a number of example sentences of the perfect tense. The positive form is demonstrated by *lesogana le e rekile* '(a) young man bought it/them', an ambiguous sentence, for the object concord *e* may either belong to noun class 4 (plural)

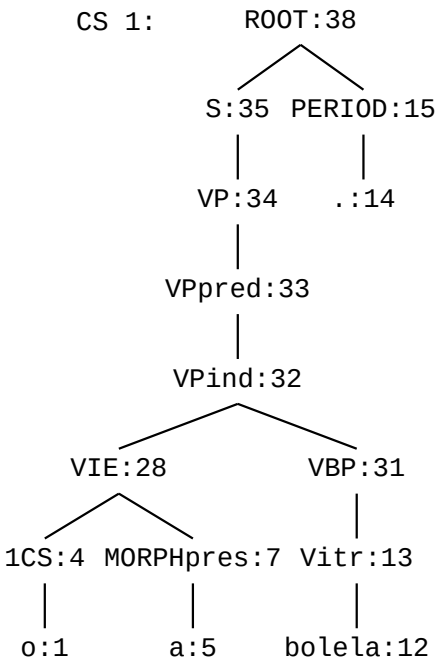


Figure 5.17: C-structure of a positive indicative, present tense, subject not present *o a bolela.* '(s)he/it speaks / you speak.'

or 9 (singular). Hence, the analysis results in two f-structures, packed in Figure 5.19. However, as the constituent analysis is not affected by this issue, only one c-structure results (cf. Figure 5.18). The negated examples are *lesogana ga se la bolela. / lesogana ga se le bolele. / lesogana ga la bolela. / lesogana ga le a bolela.* '(a) young man did not speak.' in figures 5.20/5.21, 5.22/5.23, 5.24/5.25, and 5.24/5.25.

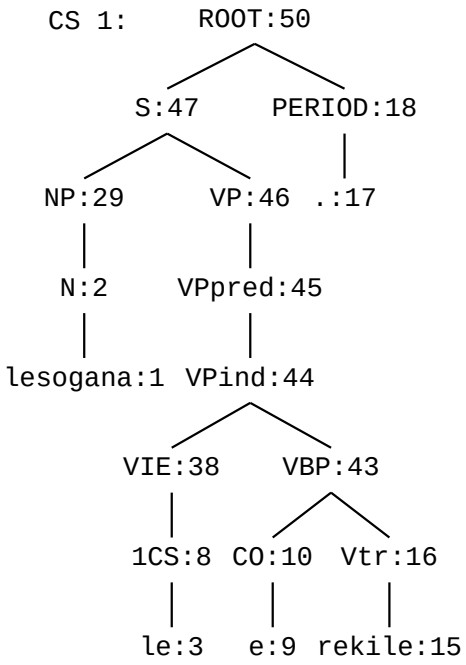


Figure 5.18: C-structure of a positive indicative, perfect tense *lesogana le e rekile.* '(a) young man bought it/them.'

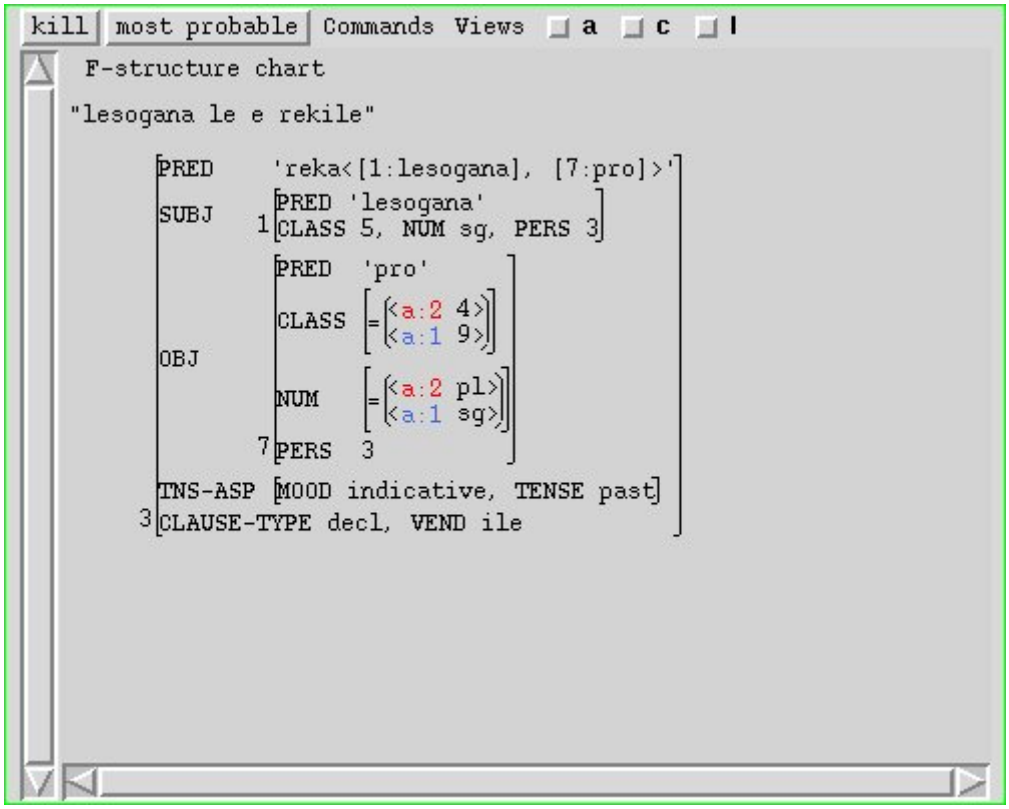


Figure 5.19: Packed f-structure of a positive indicative, perfect tense *lesogana le e rekile.* '(a) young man bought it/them.'

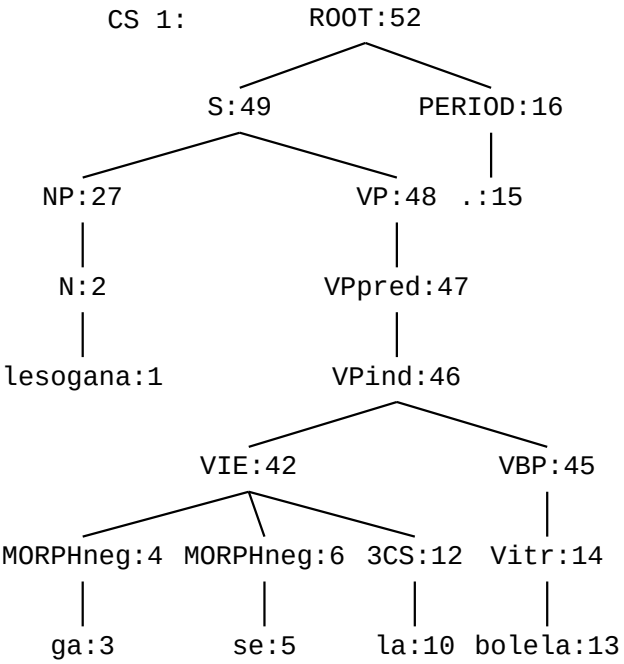


Figure 5.20: C-Structure of a negated indicative, perfect tense *lesogana ga se la bolela.* '(a) young man did not speak.'

```
"lesogana ga se la bolela."

 15 PRED     'bolela<[1:lesogana]>'
 16        1 PRED 'lesogana'
 13 SUBJ   2 CLASS 5, NUM sg, PERS 3
 14       27
 45
 10 TNS-ASP [MOOD indicative, POL neg, TENSE past]
 12 CLAUSE-TYPE decl, NEG1 ga, NEG2 se, VEND a, VTYPE main
  5
  6
  3
  4
 42
 46
 47
 48
 49
 52
```

Figure 5.21: F-Structure of a negated indicative, perfect tense *lesogana ga se la bolela.* '(a) young man did not speak.'

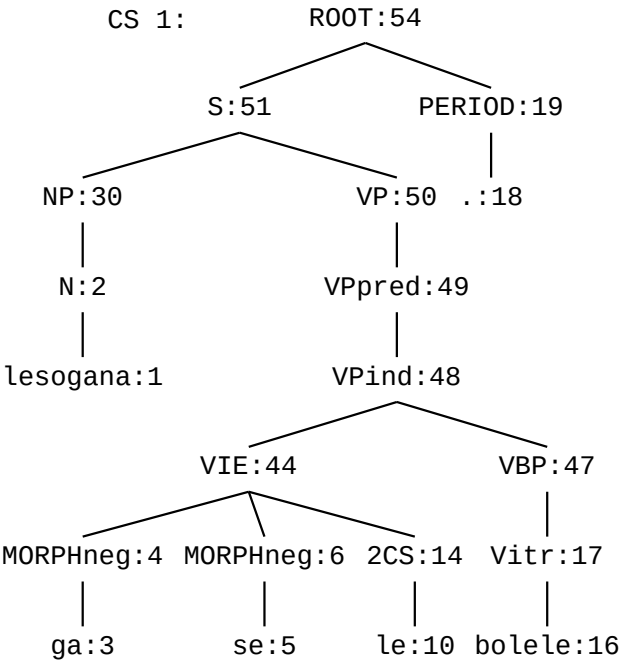


Figure 5.22: C-Structure of a negated indicative, perfect tense *lesogana ga se le bolele.* '(a) young man did not speak.'

```
"lesogana ga se le bolele."

 18 PRED     'bolela<[1:lesogana]>'
 19        1 PRED 'lesogana'
 16 SUBJ   2 CLASS 5, NUM sg, PERS 3
 17       30
 47
 10 TNS-ASP [MOOD indicative, POL neg, TENSE past]
 14 CLAUSE-TYPE decl, NEG1 ga, NEG2 se, VEND e, VTYPE main
  5
  6
  3
  4
 44
 48
 49
 50
 51
 54
```

Figure 5.23: F-Structure of a negated indicative, perfect tense *lesogana ga se le bolele.* '(a) young man did not speak.'

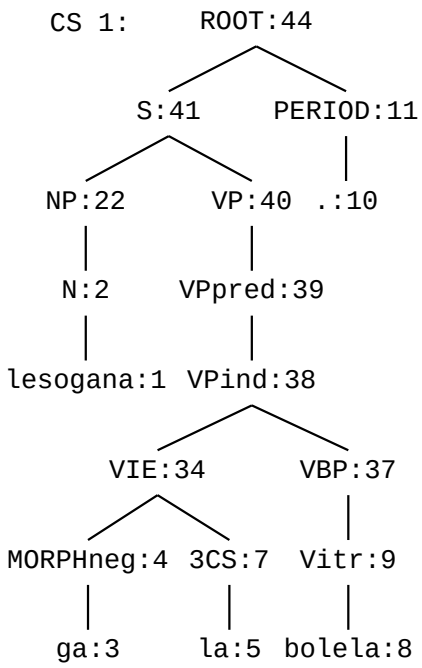


Figure 5.24: C-Structure of a negated indicative, perfect tense *lesogana ga la bolela.* '(a) young man did not speak.'

```
"lesogana ga la bolela."

    10[PRED    'bolela<[1:lesogana]>'                    ]
    11[       1[PRED 'lesogana'                          ]
     8[SUBJ   2[CLASS 5, NUM sg, PERS 3                  ]
     9[      22[                                         ]
    37[
     5[TNS-ASP [MOOD indicative, POL neg, TENSE past]
     7[CLAUSE-TYPE decl, NEG ga, VEND a, VTYPE main
     3[
     4[
    34[
    38[
    39[
    40[
    41[
    44[                                                  ]
```

Figure 5.25: F-Structure of a negated indicative, perfect tense *lesogana ga la bolela.* '(a) young man did not speak.'

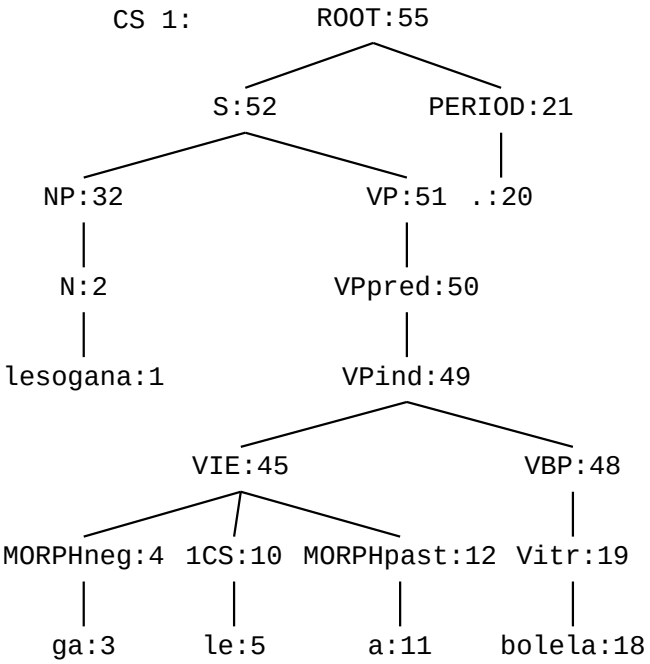


Figure 5.26: C-Structure of a negated indicative, perfect tense *lesogana ga le a bolela.* '(a) young man did not speak.'

```
"lesogana ga le a bolela."

 20 [PRED     'bolela<[1:lesogana]>'                          ]
 21 |        1 [PRED 'lesogana'                                |
 18 |SUBJ     2 [CLASS 5, NUM sg, PERS 3                      |
 19 |        32 [                                             |
 48 |
 11 |TNS-ASP [MOOD indicative, POL neg, TENSE past]
 12 |CLAUSE-TYPE decl, NEG ga, VEND a, VTYPE main
  5 |
 10 |
  3 |
  4 |
 45 |
 49 |
 50 |
 51 |
 52 |
 55 |
```

Figure 5.27: F-Structure of a negated indicative, perfect tense *lesogana ga le a bolela.* '(a) young man did not speak.'

5.2.4.4 The predicative independent indicating mood: The future indicative

Finally, we add the future indicative, as described in paragraph 3.2.5.3. The positive form contains one of the future tense morphemes *tlo* or *tla*, the negated form contains the potential morpheme *ka* and the negation *se*. Concerning the negated form, none of its elements can clearly be identified to indicate the future tense, therefore, again an empty element "e" is defined that inserts this information into the f-structure.

```
"Verbal Inflectional Elements : VIE"
VIE -->
       { ...

         " future tense "
       |  1CS
          MORPH : (^ TNS-ASP TENSE) = fut
                  (^ TNS-ASP POL) = pos
          " negated future tense "
       |  2CS
          MORPHpot
          MORPH : (^ NEG) =c se;
          e: (^ TNS-ASP TENSE) = fut
       }.
```

Example analyses of *mmutla o tlo tšhaba.* '(a) hare will flee.' (figures 5.28 and 5.29) and its negated form *mmutla o ka se tšhabe* '(a) hare will not flee.' (figures 5.30 and 5.31) conclude this section on implementation.

5.2.5 Summary

The purpose of this chapter was to show the possibility of an implementation of part of the Northern Sotho grammar fragment defined in chapter 3. Our "toy"-grammar will be extended in the future. We have defined the imperative and indicative constellations making use of a full form lexicon. At a later stage, this full form lexicon may be replaced by a morphological analyser. The current implementation, however, may already form the basis for an experimental machine translation into English, which will be the subject of the next chapter.

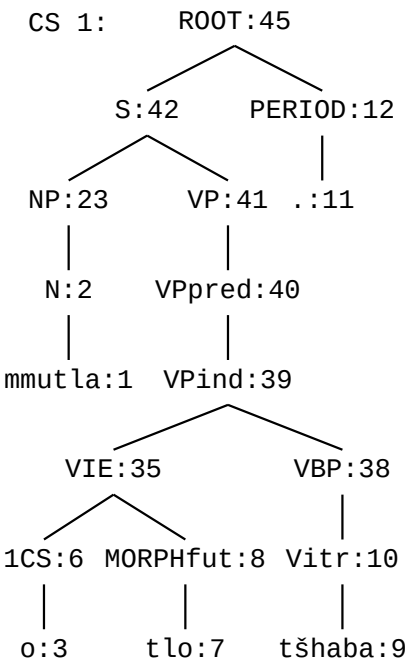


Figure 5.28: C-structure of a positive indicative, future tense *mmutla o tlo tšhaba.* '(a) hare will flee.'

"mmutla o ka se tšhabe."

```
17 PRED     'tšhaba<[1:mmutla]>'
18        1 [PRED 'mmutla'                          ]
15 SUBJ   2 [CLASS 3, NUM sg, PERS 3                ]
16       29 [                                        ]
45
10 TNS-ASP [MOOD indicative, POL neg, TENSE fut]
11 CLAUSE-TYPE decl, NEG se, VEND e
 7
 8
 3
 5
42
46
47
48
49
52
```

Figure 5.29: F-structure of a positive indicative, future tense *mmutla o tlo tšhaba.* '(a) hare will flee.'

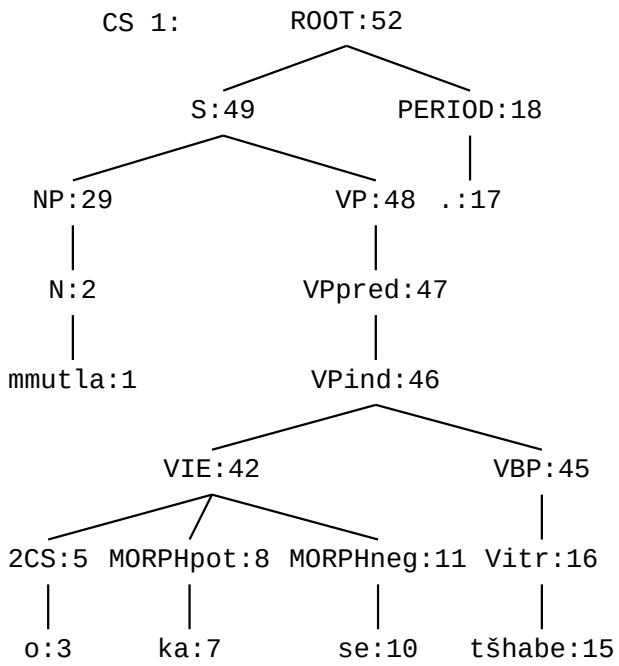


Figure 5.30: C-structure of a negated indicative, future tense *mmutla o ka se tšhabe.* '(a) hare will not flee.'

```
"mmutla o ka se tšhabe."

17 [PRED     'tšhaba<[1:mmutla]>'                          ]
18 [        1[PRED 'mmutla'                                ]
15 [SUBJ    2[CLASS 3, NUM sg, PERS 3                      ]
16 [       29[                                             ]
45 [
10 [TNS-ASP [MOOD indicative, POL neg, TENSE fut          ]
11 [CLAUSE-TYPE decl, NEG se, VEND e
 7 [
 8 [
 3 [
 5 [
42 [
46 [
47 [
48 [
49 [
52 [                                                       ]
```

Figure 5.31: F-structure of a negated indicative, future tense *mmutla o ka se tšhabe.* '(a) hare will not flee.'

Chapter 6

A basis for an automated translation

6.1 Introduction

As far as we know, there has only been one attempt so far to translate automatically from a Bantu Language spoken in South Africa into a European Language or vice versa[1]. While the syntactic features and the lexical stock of Northern Sotho have been deeply investigated on a linguistic and at least to some extent on a computational linguistic level (cf. Roux and Bosch (2006) for an overview), machine translation is currently not a major issue in South African linguistics.

In Europe, there have been a number of projects on Machine Translation (MT)[2], Bantu Languages however have not been targeted by computational linguistics so far – a rising interest[3] can however be noted.

Some earlier European projects on MT, such as EUROTRA (Hutchins and Somers, 1992, pp. 239–257), may not have reached all their development goals. However, at least one important aspect of developing MT was always proved: One will inevitably gain a deep insight into the source language when describing it in order to make a computer translate it into another language. "Contrastive Knowledge" (Jurafsky and Martin, 2000, p. 807) about both languages is also enhanced significantly. And knowledge gained in such manner

[1] Jordaan-Weiss (1996) reported on a system called *EPI-use*, translating administration documents between Setswana, English and Afrikaans.

[2] A summary of MT systems of the past is found e.g. in Hutchins and Somers (1992), for an explanation of current developments cf. e.g. METIS-II, cf `http://www.ccl.kuleuven.be/ws-metis/`.

[3] Cf. `http://www.aflat.org/?q=node/322` about a workshop on Human Language Technologies for the African Languages, held in Athens in April 2009 in the framework of the 12^{th} Conference of the European Association of Computational Linguistics, EACL.

will be useful for further linguistic work, be it lexicography or projects on Computer Aided Language Learning (CALL).

This study describes preparatory steps on the way towards a rule-based machine translation from Northern Sotho to English, but it does not present such a system itself. In this chapter, MT is introduced in general (section 6.2). Some specific Northern Sotho to English translation challenges and suggestions how to address them follow in section 6.3. Lastly, constrastive descriptions for translating Northern Sotho expressions from different word classes into their English equivalents are described in section 6.4.

6.2 An Introduction to Machine Translation

MT is an automation of the translation process, it is however not supposed to be a complete substitute for a human translator, as in the analysis of text of any source language, there are lexical and structural ambiguities to be resolved, language specific idioms to be noted, and anaphora to be resolved, just to state a few of the many problems that are difficult for automatic systems. Furthermore, the same efforts are necessary for the target language, too, in order to avoid producing ambiguous output.

In rule-based MT, appropriate electronic mono-, bi- or even multilingual dictionaries and grammars need to be developed before it is possible to design proper translations into a target language. For such a resource building procedure, the first technical question is how finely an MT system should analyse the source language before the translation is done. The more detailed the analysis is, the more monolingual resources will have to be built.

Dorna (2001, p. 516) (referring to Vauquois (1968)), demonstrates with a triangle (cf. Figure 6.1) that rule-based MT-systems have been categorised according to the abstraction level of the representation used for contrastive mappings, i.e. from "direct" MT that analyses and translates phrases in one step, to "interlingua" MT that analyses source language sentences to a representation level that – in theory – is language independent. From there, it generates target language sentences, skipping a translation step. The higher in the triangle the translation process begins, the more stringent is the analysis of the source sentence, which in turn requires further monolingual resources for the source language. The lower in the triangle the translation process ends, the less effort is required in analysing the target language, requiring fewer monolingual resources to generate the translated sentence. The

arrow leading from source to target language stands for the contrastive description, i.e. the transfer itself: the longer it is, the greater the effort necessary to transfer sentence representations from source to target language.

Another aspect of MT is also demonstrated by this triangle: direct MT is placed at the base of the triangle. Here, few monolingual resources are necessary for source and target language, the main focus lying in the development of transfer lexicons and rules. Moving up the triangle, monolingual analysis is expected to become more and more language-independent, the highest point representing the "interlingua", a representation level that is supposed to represent all summarised knowledge of a sentence. This representation is independent of any language-specific structural information[4], Vauquois (1968, p. 207) defines it as the "representation of meaning". From this top representation shown in Figure 6.1, it should in theory be possible to generate equivalent sentences in any language, a transfer step is no longer necessary.

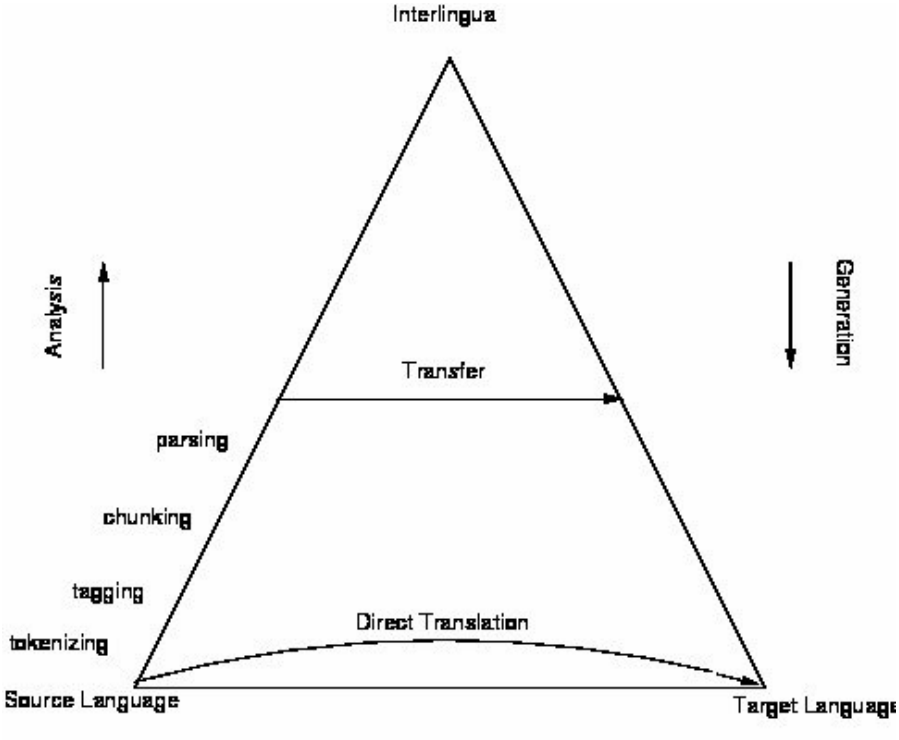


Figure 6.1: The MT triangle

[4]An example of interlingua in use is described by Traum and Habash (2000).

All systems inbetween "direct" and "Interlingua" MT are called "transfer" MT for they analyse the source language to a certain extent, transfer the results of analysis to an adequate representation of the target language and then generate output in the target language.

6.2.1 System architecture and interfaces between modules

In general, analysing the source and generating the target language is implemented as two different processes. Each step of the analysis receives knowledge gained by the previous steps and enriches the input text with more knowledge, resulting in the next higher level of knowledge-representation. Where analysis ends, transfer begins; and it will depend on the monolingual knowledge on the target language present in the system as to how many translation tasks will be performed.

The analysis of an input text begins with tokenization[5] (cf. paragraph 1.4.2.1 on page 9), where each language unit, i.e. each token, is identified and marked, in most cases by line breaks. Sentence borders are usually also identified and labelled by the tokenizer. A part of speech (POS) tagger adds word class information to each token of a sentence leading to the next level of representation. The following module may be a chunker adding non-recursive, structural information in determining chunks[6].

On the level of such chunks, a successful transfer to the target language is already possible and although such systems (e.g. SYSTRAN, cf. (Hutchins and Somers, 1992, p. 175 et seq.)) are known to be robust, the lack of syntactic and semantic knowledge often leads to an inaccurate translation. Therefore, analysis preferably goes further, at least to a complete syntactic representation of the source sentence. The modular EUROTRA system (cf. (Hutchins and Somers, 1992, p. 239 et seq.) and Figure 6.2) analyses the source text in a series of cascading steps. It begins with morphosyntactic representation, continues with a constituent structure and a "relational" representation containing information on the grammatical functions of the units of the sentence. It then adds an interface-structure, "based on semantic interdependency" (Hutchins and Somers, 1992, p. 244), making the transfer step less difficult, as it is independent of both word order and other syntactic issues. LFG (in its implementation XLE in, cf. paragraph 5.1.2.1 and section 6.5) usually processes the

[5]In this study, formatting issues do not play a role (i.e. how to handle text produced by different word processors).

[6]'Chunking' is also known as 'shallow parsing'. This processing usually results in a flat, non-recursive parse, cf. paragraph 4.2 on page 192.

transfer step (the τ-function) on the level of f-structure[7].

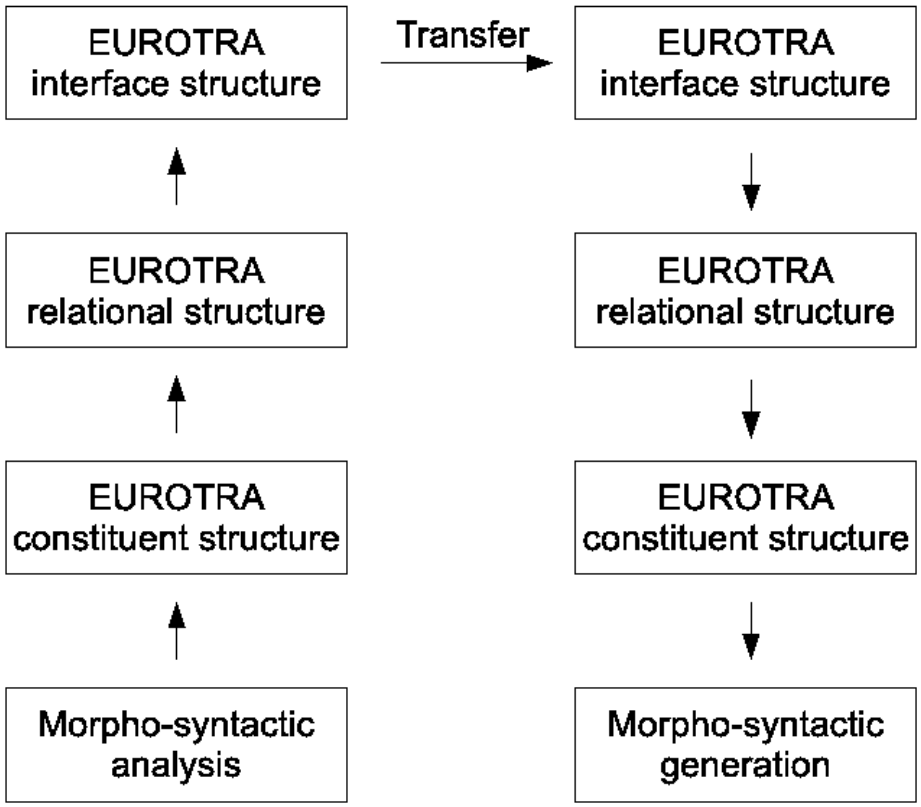


Figure 6.2: Translating with the transfer system EUROTRA

The tasks of the transfer step in MT Systems depend to a huge extent on the level where analysis ends and generation begins; therefore, the task division between transfer and generation cannot be defined clearly on the level of a general description of MT processes. A guide rather than hard and fast rules might describe issues like the translation of idioms or the mapping of f-structures to be processed during transfer while word order and morphosyntactic issues are catered for by the generation step. While in general, transfer

[7]Note that Fenstad et al. (1987) describe the possibility of an additional semantic representation, (developing the σ-function that produces a semantic structure from the f-structure). Kaplan et al. (1995), referring to Fenstad, adds a second transfer step (the τ'-function) in parallel to that of f-structure, which enables transfer of this semantic representation.

might partially integrate generation tasks, EUROTRA is known for its strict stratificational design.

The critical issue of generation is that of underspecification: when generating from e.g. an f-structure, often a huge number of possible surface sentences are possible. Here, the generating process must be restricted to only one or very few options that the user is then supposed to choose from. Restriction in generation can be "pre-determined" (Hutchins and Somers, 1992, p. 137 et seq.), state however, that such generators produce rather monotonous structures. A better solution is a strict preservation of the source sentence's structure, if ever possible, leading to literal translations. This puts MT, as Hutchins and Somers (1992, p. 138) describe it, "in direct contrast with human translation", but it leads to a higher level of accuracy. Current systems, however, can learn from existing translations and thus create more variability (cf. paragraph 6.2.3).

6.2.2 Reversibility of resources and processes

A number of resources are necessary to support the processes in analysis, transfer and generation, and it should be the aim of a designer to keep these resources reversible, whenever possible. Lexicons may in general be handled as reversible resources, however, some entries have to be handled as exceptions: a word that is rather unambiguous in one language, might become ambiguous when being viewed from another. The verb *kwa*, for example, is to be translated either into '[to] feel', '[to]taste' or '[to] smell', depending on the context in which it appears[8]. Consequently, conditions (i.e. constraints on contextual data) have to be taken into account when translating *kwa* from Northern Sotho to English, while vice versa, this should not be necessary.

Any modern parser can analyse and generate using the same grammatical resources. However, issues like the necessity to restrict the number of possible translations (paragraph 6.2.1) can make it necessary for analysis and generation to use different parts of the grammar[9].

Concerning the transfer rules, one can indeed assume that if all transfer rules are reversible, the transfer system as a whole is reversible, too.

[8]Translations from the Oxford School dictionary, Northern Sotho - English (cf. De Schryver (2007)).

[9]There have been systems such as ROSETTA (Hutchins and Somers, 1992, pp. 279 et seq.) that used one grammar for both directions.

6.2.3 Developments in MT

Earlier machine translation systems were designed as rule-based, however, throughout the last decade, statistical methods gained more influence and nowadays either support rule-based MT significantly like, e.g. in METIS[10] or form the basis of new systems, e.g. the (commercial) system Language Weaver (Fraser and Wong, 2008). This statistical machine translation (SMT) makes use of machine learning algorithms, i.e. it can remember previous choices and based on this data, its translation results continuously improve (cf. e.g. the "T(ranslation) M(emory) Generator" described by Fraser and Wong (2008, p. 16)).

Representations of linguistic knowledge about source and target language and contrastive knowledge form the basis of any rule-based MT. SMT on the other hand requires statistical models instead (for source and target language, and transfer), which are developed on the basis of corpora. These text collections of the languages in question are generally rather large, e.g. Koehn (2002) uses over 20 million words per language. In general, the size of the text collection and its quality both play a role, as the models are generated on the basis of word sequences and distributional issues. There is only a little linguistic processing in this kind of SMT, therefore it can be categorised as direct MT (Dorna, 2001, p. 519).

Recently, experiments with statistical machine translation on the basis of linguistic representations, e.g. f-structures of LFG have successfully been performed. Such an approach, as described by Riezler and Maxwell III (2006) or Graham et al. (2009), achieves a significant improvement of the quality of the translated sentences compared to standard SMT.

In the case of Northern Sotho, the lack of parallel, comparable or even monolingual corpora necessary for developing such a statistical language and/or translation model, makes it advisable at the current point in time to begin with a rule-based approach that may be enhanced at a later stage with heuristics based on corpus data.

6.3 MT from Northern Sotho to English: general lexical and structural issues

This section describes problems to expect when automatically translating from Northern Sotho to English. We begin with lexical ambiguities of the source language, and continue

[10] cf. `http://www.ccl.kuleuven.be/ws-metis/`.

with lexical mismatches or gaps between the languages in question, paragraph 6.3.1 also shows examples. Concerning structural ambiguities, a typical example is explained in paragraph 6.3.2. Structural differences (divergences) include differences in argument structure which will also be described alongside an example in paragraph 6.3.3. The lack of adjectives in Northern Sotho leads to the prominent use of verbal relatives and possessives, which need specific attention when being translated into English adjectives; an example is demonstrated in paragraph 6.3.4. Differences in word order between source and target may be separated from these issues, a possible handling of those is described in paragraph 6.3.5. Finally, the lack of determiners in Northern Sotho may force an insertion of determiners during transfer, cf. 6.3.6.

6.3.1 Lexical ambiguities in the source language

To demonstrate a problematic case of lexical ambiguity, we utilise the Online Northern Sotho to English dictionary[11], translating the Northern Sotho word *ja* into English; the results are shown in Table 6.1.

- Word senses 1–9 can be divided into four groups, where a general translation of *ja* as 'eat' may summarise both of the first two groups;
  - transitive verb
    eat, devour, consume, cost, despoil
  - intransitive verb
    eat, cohabit
  - paraphrased
    have sex
  - paraphrased, where a possible object of *ja* could be translated as the object of a prepositional phrase, to be added later.
    take nourishment (of), partake of,
- word senses 10 to 19 show *ja* as a part of multiword constellations. Some are directly translatable, others will have to be paraphrased;
- word sense 20 results from adding the suffix *-go* to the verb and is to be understood as an verb stem appearing in an indirect relative clause, here as 'who is/are eating'.

[11] cf. `http://africanlanguages.com/sdp/`.

Table 6.1: Translations of the Northern Sotho verb *ja* into English

no.	**grouped**	**Norther Sotho**	**English**
		ja	
1	1a		eat
2	1b		devour
3	1c		consume
4	1d		take nourishment
5	1e		partake of
6	2		cost
7	3		despoil
8	4a		have sex
9	4b		cohabit
		ja hlogo	
10	1		ponder
11	2		think
		ja moretlwa	
12	1		get a hiding
		ja motho leonyane	
13	1		shadow a person
14	2		follow stealthily
		re ja	
15	1a		beat
16	1b		slap
17	1c		strike
18	1d		ache
19	2		confiscate
		jago	
20	1		who is eating

Lexical ambiguities may be handled in different ways. Corpus data may be studied to find the most frequent word senses of *ja*, less frequent word senses may then be ignored by the system. The source language data could be enhanced with semantic information, helping to resolve ambiguities when transferring to the target language. Thirdly, the use of *ja* in certain multiword units, e.g. *ja moretlwa* 'get a hiding' can be stored in a bilingual collocation lexicon, which is then taken into account by the transfer system prior to any further lexical or morphosyntactic processing of single units.

Lexical gaps in English are the cause for single words of Northern Sotho having to be translated as phrases of English, like in example 93 (a) and (b). On the other hand, some English colour differentiations do not exist in Northern Sotho, as example 93 (c) shows. An MT-lexicon should cater for such transitions[12]

93(a) *lebese*$_{\text{N05}}$
fresh milk
'fresh milk'

(b) *Leburu*$_{\text{N05}}$
white Afrikaans-speaking person
white Afrikaans-speaking person'

(c) *tše*$_{\text{CDEM09}}$ *khubedu*$_{\text{ADJ09}}$
`dem-cl09` red/orange
'red/orange'

6.3.2 Structural ambiguities in the source language

Paragraph 4.2 on page 192 describes one of the biggest problems in analysing indo-European languages like English: 'PP-attachment'. This issue may be demonstrated by the sentence *The toy rocket flew to the planet with lights on*, cf. Figures 6.3 and 6.4.

Coordination can lead to another structural ambiguity occurring regularly in these languages: for example, the analysis of *mothers and children under (the age of) 13* (cf. paragraph 4.2) will also result in several trees, of which two are shown in Figures 6.5 and 6.6.

The same applies for Northern Sotho (for sake of convenience, we repeat example (77) of page 193 as (94), demonstrating the case): although the concordial system often supports avoiding structural ambiguity, sentences like (94) are as ambiguous as they are in English, whenever the referents belong to the same class. Monolingual analysis results in several trees, where the respective VP is attached to either the top node, the first nominal of the coordination, or the second one. Figures 6.7 and 6.8 show the first and the last case. Concerning an MT system, for such 'VP-attachment' ambiguity[13] occurring in Northern Sotho there does not seem a necessity to resolve it during analysis and it could very well be that it remains during transfer to English where the same ambiguity is present as a

[12]The examples in (93) are taken from the Oxford School dictionary, cf. De Schryver (2007).

[13]The same could apply for particle phrase attachment in Northern Sotho, i.e. if nominals of a coordination have a particle phrase (cf. paragraph 3.9 on page 185) in the appendix.

pp attachment. However, more research on such structures contained in text collections is deemed necessary, before such an assumption can be made with more assurance.

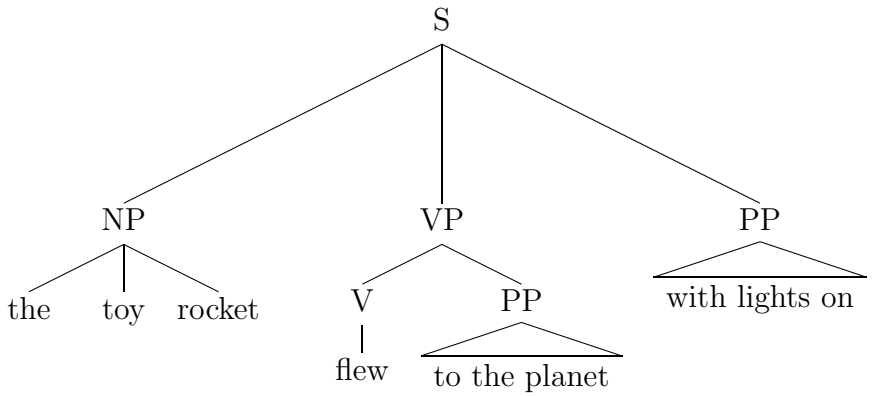


Figure 6.3: Example analysis: PP attachment (top node)

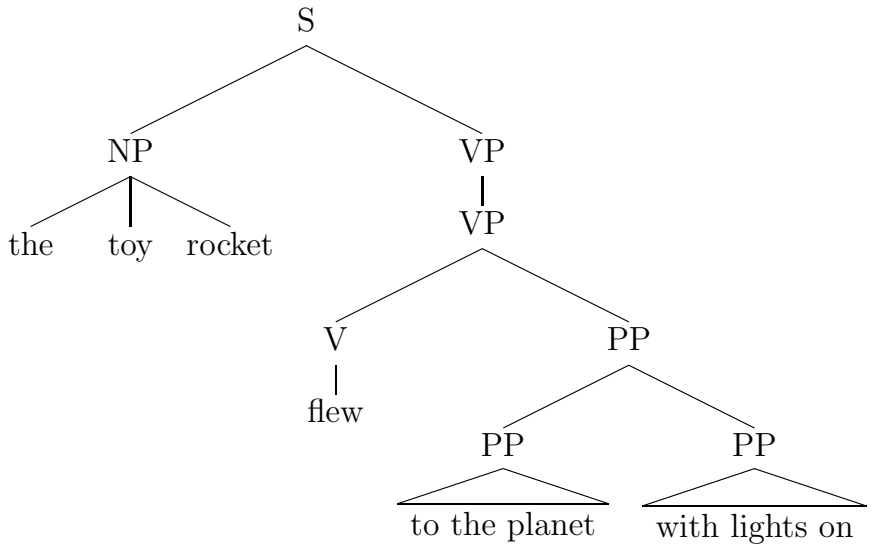


Figure 6.4: Example analysis: second PP attached to first PP

(94) *bomme*$_{\text{N02b}}$ *le*$_{\text{PART\_con}}$ *bana*$_{\text{N02}}$ *ba*$_{\text{CDEM02}}$ *ba*$_{\text{CS02}}$ *lego*$_{\text{VCOP}}$ *fase*$_{\text{NLOC}}$
mothers `con` children `dem-3rd-cl2` `subj-3rd-cl2` who are under
ga$_{\text{CPOSSLOC}}$ *mengwaga*$_{\text{N04}}$ *ye*$_{\text{CDEM04}}$ *13*$_{\text{NUM}}$
of years `dem-3rd-cl4` 13
'mothers and children under 13'

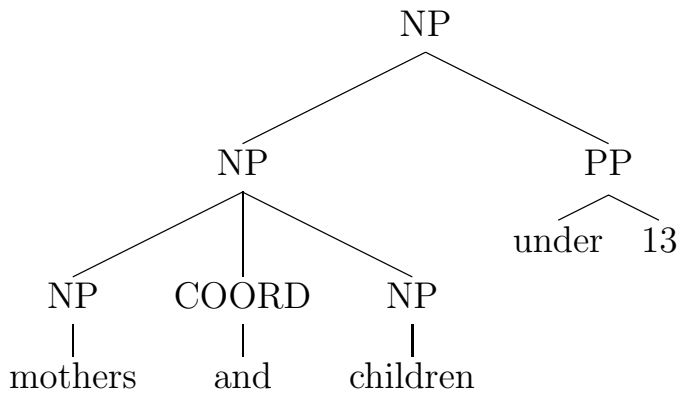


Figure 6.5: Example analysis: pp attachment (top node)

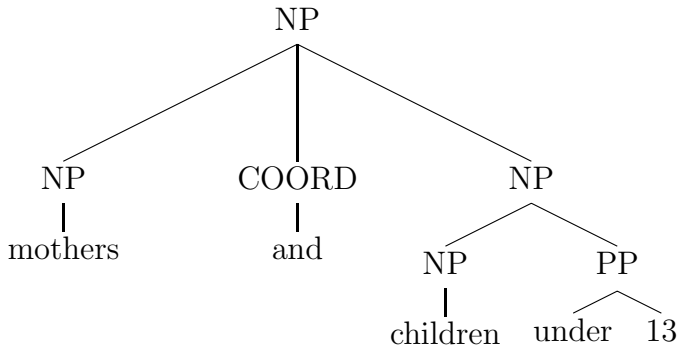


Figure 6.6: Example analysis: pp attachment (second NP of coordination)

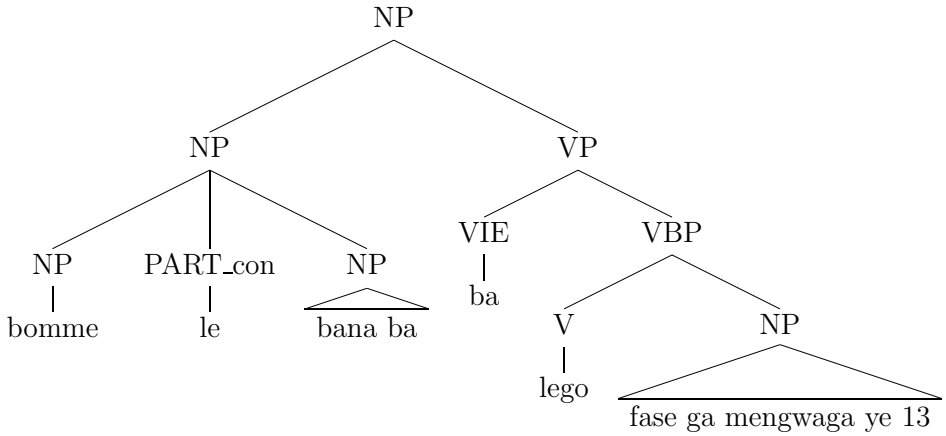


Figure 6.7: Example analysis: 'VP attachment' (top node)

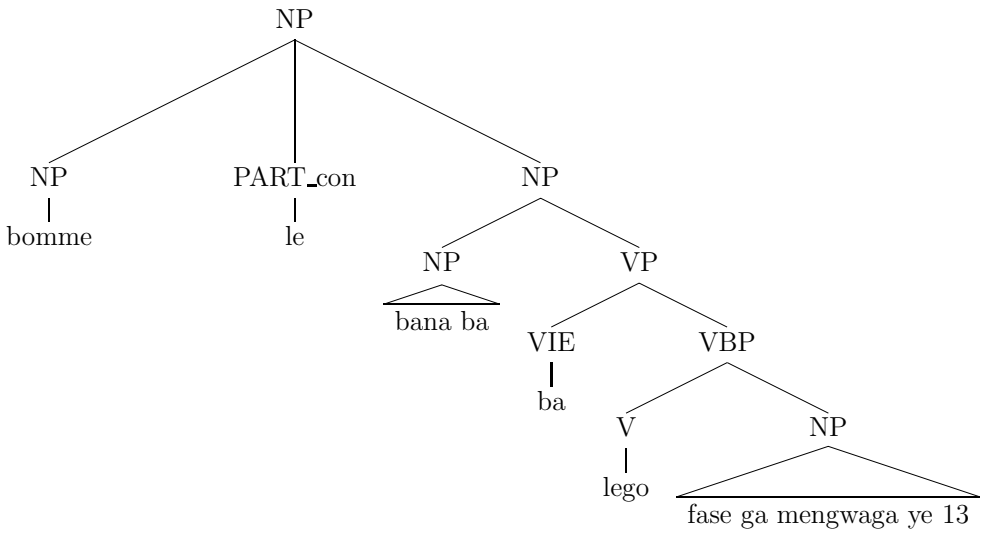


Figure 6.8: Example analysis: 'VP attachment' (second NP of coordination)

6.3.3 Differences in argument structure

Example (95) (Lombard, 1985, p. 110) demonstrates that in Northern Sotho, the applied verbal extension may be added to the verb stem, i.e. the infix *-el-*. The verb *nyaka*, for example, generally means '[to] want'. In our example, it appears in the word sense 'to look for' and may sub-categorise a direct object, e.g. *malekere* 'sweets'. This verb stem may also appear in the form *nyakela* meaning 'looking for on behalf of'. Adding the Northern Sotho applied verbal extension to *nyaka* therefore leads to the requirement that a second object be present in the clause. This object, e.g. *banna*, then appears as the first argument following the verb stem[14].

The respective English verb '[to] look' can similarly be supplemented by the oblique prepositional phrase 'for sweets', extending the original semantics of '[to] look' with the (direct) object 'sweets', that someone searches for. A second, adjunctive prepositional phrase containing the thematic object 'for the children' may be added as well, to express that someone is searching for sweets on behalf of the children.

(95) *tate*$_{\text{N01a}}$ *o*$_{\text{1CS01}}$ *nyakela*$_{\text{V\_dtr}}$ *bana*$_{\text{N02}}$ *malekere*$_{\text{N06}}$
Father `subj-01` looking for on behalf of children sweets
'father is looking for sweets for the children'

[14]In Northern Sotho, the indirect object usually precedes the direct object, cf. paragraph 3.2.1.1, referring to (Ziervogel, 1988, p. 82).

Figures 6.9 and 6.10 demonstrate the differences of f-structures[15] generated by XLE grammars of English and Northern Sotho resulting from example (95). The most apparent problem is that while the Northern Sotho verbs may have argument nominals added directly when being extended e.g. with the applied infix -*el*-, these arguments appear in their English translations as objects of prepositions. The prepositional phrases are then represented as non-mandatory obliques or as adjuncts. In the case of translating the argument into an oblique, a simple transfer rule is sufficient, while a translation into (a possible set of) adjuncts is no trivial task, which is however solvable (see Emele and Dorna (1998) who make use of XLE's packed representations). In XLE, such differences in functional structure can be catered for in the transfer lexicon. However, there are also a number of other issues to be considered there when translating automatically into English; these will be listed in section 6.4.

[15]We have simplified these f-structures for the sake of demonstration.

$$\begin{bmatrix}
\text{PRED} & \text{'nyakela } \langle \text{SUBJ,OBJ-TH,OBJ} \rangle\text{'} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'tate'} \\ \text{CLASS} & \text{N01a} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'malekere'} \\ \text{CLASS} & \text{N06} \\ \text{NUM} & \text{pl} \\ \text{PERS} & 3 \end{bmatrix} \\
\text{OBJ-TH} & \begin{bmatrix} \text{PRED} & \text{'bana'} \\ \text{CLASS} & \text{N02} \\ \text{NUM} & \text{PL} \\ \text{PERS} & 3 \end{bmatrix} \\
\text{TNS-ASP} & \begin{bmatrix} \text{TENSE} & \text{PRES} \\ \text{MOOD} & \text{INDICATIVE} \end{bmatrix}
\end{bmatrix}$$

Figure 6.9: Simplified f-structure of *Tate o nyakela bana malekere*).

$$\begin{bmatrix}
\text{PRED} & \text{'look } \langle \text{SUBJ OBL} \rangle\text{'} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'father '} \\ \text{CASE} & \text{nom} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{'for } \langle \text{OBJ} \rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'sweets'} \\ \text{CASE} & \text{obl} \\ \text{NUM} & \text{pl} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix} \\
\text{ADJUNCT} & \begin{bmatrix} \text{PRED} & \text{'for} \langle \text{OBJ} \rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'children'} \\ \text{SPEC} & \begin{bmatrix} \text{PRED} & \text{'the'} \end{bmatrix} \\ \text{CASE} & \text{obl} \\ \text{NUM} & \text{pl} \\ \text{PERS} & 3 \end{bmatrix} \end{bmatrix} \\
\text{TNS-ASP} & \begin{bmatrix} \text{TENSE} & \text{PRES} \\ \text{MOOD} & \text{INDICATIVE} \end{bmatrix}
\end{bmatrix}$$

Figure 6.10: Simplified f-structure of 'Father looks for sweets for the children').

6.3.4 Translating the verbal relative / possessive

As described in section 2.5 on page 45, adjectives in Northern Sotho form a closed class. Therefore, a number of noun properties are expressed utilising verbal relatives (as in 96 (a)) or as possessive structures (cf. 96 (b)). While the verbal relative can be translated isomorphically, a literal translation of possessives of the kind shown in 96 (b) fails. For such cases, a general transfer rule could be designed stating that possessive phrases containing infinitive forms of such intransitive verbs[16] are to be translated as adjectives.

[16]Verbs that semantically contain a copula, like *seleka* or *thaba*.

96(a) *mošemane*$_{\text{N01}}$ *yo*$_{\text{CDEM01}}$ *a*$_{\text{2CS01}}$ *selekago*$_{\text{V-itr}}$
Boy `dem-3rd-cl01` `subj-3rd-cl01` who-be-naughty
'(a) boy who is naughty'

96(b) *mošemane*$_{\text{N01}}$ *wa*$_{\text{CPOSS01}}$ *go*$_{\text{MORPH\_cp15}}$ *seleka*$_{\text{V-itr}}$
boy of to be-naughty
'(a) naughty boy'

6.3.5 Differences in word order

English and Northern Sotho are both SVO languages. In English, subjects and sub-categorised objects may not be omitted without substitution by a pronoun. In Northern Sotho, the subject concord obtains pronominal status if the subject is omitted. If a sub-categorised object does not appear in its usual position following the verb stem, the pronominal object concord will have to fill the position immediately preceding the verb stem, therefore, word order is changed in this case to SOV, cf. section 3.1.

Any transfer system should cater for differences in word order. XLE handles such challenges by transferring from source to target language on the level of functional structure, where word order plays no role. It then generates the target language sentence using a monolingual grammar. We describe this feature of XLE in more detail in section 6.5.

6.3.6 Lack of determiners

The lack of determiners in Northern Sotho poses a problem of translation, as determiners play an important role in English. Not only are they often necessary from a syntactic perspective, moreover, definite and indefinite articles provide different discourse information as well. Definite articles appearing with a noun suggest that the reader/listener already knows the entity the noun refers too. In Northern Sotho, on the other hand, a known noun is rather omitted while the respective concord takes its syntactic function. Such concords should be translated as pronouns (we will explain this issue in more detail in paragraph 6.4.2). It may therefore be assumed that most nouns appearing in Northern Sotho text introduce a new entity to the discourse, while the appearance of a pronominal subject concord rather signals a known entity. Therefore, a routine inserting the indefinite article 'a(n)' during transfer whenever syntactically necessary (i.e. for singular nouns) is suggested.

6.4 MT from Northern Sotho into English: parts of speech

So far, a number of examples of Northern Sotho words and their constellations together with their English translations have been mentioned. This section will describe some phenomena of Northern Sotho on the level of parts of speech (POS) which have to be considered when transferring these constellations into English, i.e. a contrastive description.

6.4.1 Transferring nouns and pronouns

Usually, a noun may simply be translated utilising a bilingual lexicon, e.g. *monna* → '(a) man'. The nominal attributes 'person' and 'number' are to be transferred unchanged as they are necessary for the English grammar to generate the correct word form. The monolingual lexical attribute 'noun class', however, being irrelevant for the English translation, may be deleted during transfer.

In paragraph 2.2.2.5 (page 28), it was argued that nouns of class 15 should rather be analysed as infinitive verbal phrases, identical to the English 'to'-infinitive constructions. Such a method allows for an easy, isomorphic transfer possibility on the level of f-structure. For sake of demonstration, we repeat example 4 (a) from page 29 as (97) and show a respective simplified f-structure (6.11).

(97)

ba	*rata*	*go*	*bala*	*dipuku*
`subj-3rd-cl2`	like	to	read	books

'they like to read books '

Emphatic pronouns find their equivalents in English where they either appear as deictic determiners (e.g. *tše* 'these') or as pronouns (e.g. *nna* 'I', or 'me' respectively). However, in the first case, the position in which they appear relative to the noun they refer to is relevant for a correct interpretation. As described in paragraph 3.8.3, depending on this position, they either express a contrast or a specification. However, in the grammar fragment described so far, no specific attribute is defined that could preserve respective information. Therefore, *tše dimpša* and *dimpša tše* would both be translated as 'these dogs'. To solve this problem, several NP rules could be introduced to be used during monolingual analysis. The NP that is to be translated differently, is then to be marked,

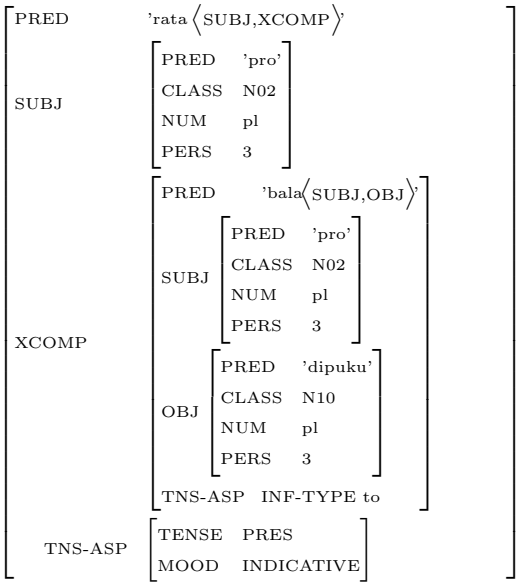


Figure 6.11: Simplified f-structure of *ba go bala dipuku*).

e.g. with an additional attribute. An NP as shown in example (98)[17] could then trigger a specific transfer rule that would insert a preceding 'as for'.

(98) *mošemane*$_{N01}$ ***yena***$_{\text{PROEMP01}}$ *o*$_{\text{1CS01}}$ *rata*$_{\text{V\_tr}}$ *diapola*$_{\text{N10}}$,
boy `emp-cl1` `subj-3rd-cl1` like apples,
'as for the boy, he likes apples,'

basetsana$_{\text{N02}}$ ***bona***$_{\text{PROEMP02}}$ *ba*$_{\text{1CS02}}$ *rata*$_{\text{V\_tr}}$ *dinamune*$_{\text{N10}}$
girls `emp-cl2` *subj-3rd-cl2* like oranges
'(but) as for the girls, they like oranges'

Paragraph 2.3.2 described possessive pronouns (e.g. *gago*$_{\text{PROPOSSPERS\_2sg}}$ 'your(s)$_{\text{sg}}$') as never appearing with the noun they refer to. They instead always occur in a pronominal function and may be translated into English accordingly. The case of quantitative pronouns is even easier, as all of them may be translated as 'all of'.

6.4.2 Transferring concords

Generally, concords are to be deleted during transfer, as most of them only provide morphological information in terms of attribute-value pairs (e.g. '*le* pers=3 num=sg). However,

[17] Example 73 (a) of page 180 is repeated here for sake of convenience.

like all elements that agree with a noun (in terms of noun class agreement), they may acquire a pronominal function when the noun they refer to is omitted (cf. paragraph 2.3 referring to (Louwrens, 1991, p. 154)). Therefore, an alternative lexical entry is necessary as shown in paragraph 5.1.2.3, which needs to be transferred to the respective pronoun(s) of English, as in (99). The latter also applies for all object concords. Note again that concords do not specify gender, therefore the English grammar will – in the case of singular – generate several possible translations because of the unification principle, cf. paragraph 1.4.4 (page 13): If an attribute-value pair is missing in a source language f-structure to be used for generation into the target language, XLE will create as many possible f-structures as pre-defined in the respective grammar (cf. the translation in (99)).

(99) $o_{1CS01/1CS03/..}$ a_{MORPH_pres} $opela_{V\_itr}$
`subj-cl1/3` `pres` sing
'(s)he/it is singing

Possessive concords introduce possessive noun phrases, similar to English prepositional 'of'-phrases; they basically could be translated as 'of' and change their word class respectively. In English, as there is no agreement between 'of' and its argument, all these attributes could therefore be deleted during transfer. However, like all concords, the possessive may also acquire a pronominal function whenever the possession is omitted, hence the monolingual lexicon should – like for the subject concords – contain two entries for each of them. Transfer to English should be done in a similar way to that of subject (and object) concords.

Demonstrative concords appear in three functions[18], they are not only deictic determiners (cf. paragraph 2.4.5 on page 43) which may easily be transferred to their English equivalents; their second function (as described in paragraph 3.2.7 on page 112) mirrors that of the English relative pronouns 'which' or 'who', which both introduce relative clauses. A transfer of the relative pronoun function should always lead to this word class. In their third function, these concords introduce adjective phrases (cf. section 2.5 and paragraph 3.8.4.2). Such concords are not yet represented in the monolingual analysis and hence play

[18]A problematic case of demonstrative concords concerning MT is that of deletion during transfer. In a number of clauses shown in this study, e.g. (94) on page 279, a demonstrative concord appears in the Northern Sotho clause, but is obviously not translated, for no determiner is present in the English translation. This study, being the first attempt to provide support for a rule-based MT system from Northern Sotho to English, cannot cater for such a case, because no rules seem to be available to identify it; the case is not described in any of the literature consulted. Comparable/parallel corpora would ease the task of finding such rules, however at present, we cannot assign more than three categories of demonstrative concords.

no role in transfer.

In the third case, the demonstrative copulative (cf. paragraph 2.4.6) may basically be treated like a copulative verb stem during transfer (cf. paragraph 6.4.10). However, to add the deictic meaning of these concords (for sake of convenience, we repeat example 14 (a) of page 44 as 100 (a)), the respective adverb of English (e.g. 'here' should be added during transfer. Again, a pronominal function (we repeat example 14 (b) as 100 (b)), makes a second entry in the lexicon necessary, similar to that of the other concords.

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

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

6.4.3 Transferring morphemes

The temporal group of the Northern Sotho morphemes (MORPH_pres, MORPH_past, MORPH_fut) only cater for the tense attribute in monolingual analysis, and therefore usually do not provide predication values. Transfer rules for the negation morphemes and their clusters (*ga, ga se, se, sa*) depend on the implementation of negations in the English grammar. If a negation there has its own predication value, the tense attribute will have to be transferred into that predication value (e.g. as the negation 'not'). Otherwise, an attribute 'neg' might be sufficient to trigger insertion of a respective negation in the target language.

The potential morpheme MORPH_pot *ka* should trigger the appearance of the modal verb 'may', as example 101[19], it is taken from (Lombard, 1985, p. 190), demonstrates.

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

One must however make sure that the negated future tense of the indicating mood (cf. paragraph 3.2.5.3 on page 102) which also makes use of the potential morpheme will be

[19]For sake of convenience, we repeat example 23 of page 58 (repeated on (69) on page 171) again here.

recognised as such and correctly translated, cf. example 24 (b) of page 59 repeated here as (102). Note that NP-negation (as e.g. in 'I see no reason') does not occur in Northern Sotho.

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

Lastly, the morpheme MORPH_cp15 *go* is to be translated directly as the infinitive particle 'to', cf. paragraph 6.4.1.

6.4.4 Transferring particles

Northern Sotho particles in transfer generally become prepositions, as, e.g. $ka_{\text{PART\_ins}}$ 'with' or $ke_{\text{PART\_agen}}$ 'by' in passive constellations. As hortative particles like, e.g., *a* introduce constellations similar to English, it is possible to translate them directly as the verb 'let'. Other such particles, e.g. *anke* may simply be translated as the interjection 'please'. The occurrence of question particles in Northern Sotho text should trigger a do-insertion[20] as they introduce yes-no questions.

6.4.5 Transferring adjectives and enumeratives

As described in sections 2.5 and 2.6, adjectives and enumeratives each form a closed class, of which most are easily transferred into their English equivalents, e.g. all forms of the stem *-bedi*, i.e. *babedi, mebedi, mabedi, pedi* and *dipedi* are translated into the English numeral 'two' (cf. Table 2.13 on page 48).

6.4.6 Transferring adverbs

Adverbs of Northern Sotho and English fulfil similar functions and may be transferred in most cases isomorphically. A special case, however, are the locative adverbs, e.g. *gabo-mogolo* that are to be expressed as prepositional phrases in English: 'at the elder sibling's'.

6.4.7 Transferring question words

Question words like, e.g. *bjang* 'how' or *mang* 'who' may both be translated directly to their English counterparts.

[20] "Do-Insertion" is a monolingual issue when generating English questions and therefore not described here.

6.4.8 Transferring verb stems

Main basic verb stem forms (Vstem+verbal ending) can be transferred with their attributes into their English translations, e.g. *otlela* '[to] drive', *otlelwa* '[is] driven' etc. The argument structure of more complex Northern Sotho verbs (e.g. the applicative forms), however, might undergo significant changes when being transferred to English. Noun phrases might find their expression in English prepositional phrases and should be described accordingly, cf. Figures 6.9 and 6.10 on page 283, where the subcategorization frame of *nyakela*, i.e. its valency describes subject, thematic object and object, all appearing as NPs, while its translation, 'search' only describes subject and object NPs. Here, the thematic object is to be transferred not only as a PP, is also becomes a non-argument adjunct (as mentioned in paragraph 6.3.3, Emele and Dorna (1998) describe the procedures necessary for transfer).

6.4.9 Transferring auxiliaries

There is a variety of auxiliary verbs in Northern Sotho of which this study only describes a few (cf. paragraph 2.7.3 on page 51). More research will be necessary before this word class can be extensively described. Concerning their transfer into English, basically four categories may be distinguished:

- temporal modifier
- modal modifier
- relative copulative
- adverbial modifier

Temporal modifiers set the verbal constellation into past tense. As such, only their tense attribute is important for transfer (e.g. *be* 'was/were'). A similar treatment is possible for the modal modifiers like *ke* 'should'. The auxiliaries *bego* 'who did/was/were' and *kego* 'who once did/was/were' however, need specific attention as they contain a copulative and a relative element and refer to humans only. These auxiliaries should not be mistaken for copulatives, as they – like all auxiliaries – are linked to their subject by a subject concord and followed by a full verb, cf. example (103), taken from (Louwrens, 1991, p. 52). Transfer in XLE maps a source f-structure with a target f-structure, therefore, the fact that they are to be treated similarly to the past tense morpheme 'a', may already be implemented for monolingual analyses, as the (simplified) f-structure of 6.12 shows.

(103) ***ba***$_{\text{1CS02}}$ ***bego***$_{\text{V\_aux}}$ *ba*$_{\text{2CS02}}$ *bolela*$_{\text{V\_itr}}$.
`subj-cl2` `who-past` `subj-cl2` talk
'those who were talking'

$$\begin{bmatrix} \text{PRED} & \text{'bolela}\left\langle \text{SUBJ} \right\rangle\text{'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{CLASS} & \text{N02} \\ \text{NUM} & \text{pl} \\ \text{PERS} & 3 \end{bmatrix} \\ \text{TNS-ASP} & \begin{bmatrix} \text{TENSE} & \text{PAST} \\ \text{MOOD} & \text{RELATIVE} \end{bmatrix} \end{bmatrix}$$

Figure 6.12: Simplified f-structure of *ba bego ba bolela*.

Finally, the adverbial modifiers, e.g. *tšama* 'continually' or *ešo* 'not yet', as well as verbs that may appear as auxiliaries, like, e.g. *šetše* 'already' can be translated into the respective English adverbs.

6.4.10 Transferring copulatives

Section 3.3 (page 125 et seq.) showed that there are many different copulative verbal constellations. All have their own specific translation, therefore, a transfer lexicon must basically cater for each of them. This study does not define such extensive rules as this would exceed its scope. However, describing groups of copulas (as e.g. done in paragraph 3.3.2, cf. Table (3.31)) may support the reduction of necessary transfer rules.

6.4.11 Transferring other parts of speech

As conjunctions, interjections and negations appear in similar roles in Northern Sotho and English, they may be transferred without changes (e.g. *gore* as 'that').

6.5 XLE in machine translation

Bresnan describes LFG in the introductory chapter of Bresnan (2001) as possibly solving the fundamental problems of designing a 'universal grammar', i.e. a grammar formalism that is able to describe concepts all languages may have in common. Different languages may express the same phenomena in different surface constellations, however, in terms of grammatical constraints, similar conceptual units can be identified. Such similarity in representing sentences of different languages leads to a lower effort when LFG is used for machine translation. For example, the kinds of grammatical functions are pre-defined

and thus used in the same way for different languages. Secondly, as f-structure abstracts from surface form, representing morphosyntactical information as feature structures, a close similarity in the representation of the same sentence in different languages results.

The application of XLE as an MT-system appears as a transfer system. Transfer functions in the MT-lexicon describe the translation of f-structures of the source language into f-structures of the target language, therefore contrastive knowledge of the units described in the f-structures is necessary.

The implementation of this method of transferring these functional units is called 'term rewriting', in other words, the system is told to replace specific units with others. Term rewriting may be used on lexical or on phrasal level. Again, it must be stressed that this study is not implementing machine translation from Northern Sotho to English, it may only be used as an input for such a project. As such, the following description only shows a simple example of the term rewriting methodology for sake of demonstrating the issue. More information on how to write transfer rules can be found at `http:/www2.parc.com/isl/groups/nltt/xle/doc/transfer-manual.html`. Newer versions of XLE automatically create the term rewriting rules themselves, if parallel corpora and similarly constructed monolingual grammars of the respective languages are both available.

Transfer rules of lexicon entries in XLE are based on the f-structure generated during monolingual analysis, therefore need to not only contain the translation of single words, but also information on how to modify the f-structure in which they appear in a way that the monolingual grammar of English will be able to generate a c-structure and a surface sentence from it. Consequently, information contained in the Northern Sotho f-structure which is not relevant for English must be deleted while absent information necessary in the English f-structure must be inserted.

Lexicon entries in XLE as described in chapter 5 are described as feature structures. A typical noun is described as follows:

```
tate  N *  (↑ PRED) = 'tate' @(CLASS 1a) @(PERS 3) @(NUM sg).
```

It is represented as the f-structure in figure 6.13.

The transfer lexicon for such entries should not only make sure that *tate* is translated to 'father', it should also delete the noun class information on class, as it is not relevant

$$\begin{bmatrix} \text{PRED} & \text{'tate'} \\ \text{CLASS} & \text{N01a} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix}$$

Figure 6.13: f-structure of *tate*

for English generation. Respective term rewriting rules[21] are therefore to be written that delete the argument `class`. The application of the transfer function leads to the English f-structure shown in figure 6.14.

$$\begin{bmatrix} \text{PRED} & \text{'father'} \\ \text{NUM} & \text{sg} \\ \text{PERS} & 3 \end{bmatrix}$$

Figure 6.14: f-structure of 'father'

6.6 Summary

As the previous paragraphs have shown, a number of the Northern Sotho word classes may in general be translated one-to-one into the word classes of English. In some cases, structural information might have to be added (like for the arguments of the verb that might translate into prepositional phrases, or possessive phrases containing infinitives of specific verb stems to be translated into adjectives of English), however a number of word classes (like, e.g. the concords) are usually not to be transferred at all. A typical feature of transfer rules is the change of the POS, e.g. particles of which most are to be transferred into prepositions, in other words, most particle phrases of Northern Sotho are isomorphical to prepositional phrases of English. Copulatives, however, remain a problem, as each of the hundreds of possible constellations are to be translated separately, most of them within fixed expressions. More research will be necessary to summarise and group the possible constellations from a translation perspective, i.e. in a way that allows the number of necessary transfer rules

[21] see `http://www2.parc.com/isl/groups/nltt/xle/doc/transfer-manual.html` for technical details on writing term rewriting rules.

to be kept low. One rule-based system supporting machine translation from one language to another is XLE, where an f-structure representing a source sentence is mapped to an f-structure of the same sentence of the other language.

This chapter has provided some basic contrastive knowledge concerning the translation from Northern Sotho into English and thus concludes the study. It is followed by a summary and conclusions resulting from the present work.

Chapter 7

Summary and conclusions

7.1 Aims of this study

Overall, this study is a first attempt of describing, from a computational perspective, a significant grammar fragment of Northern Sotho with the view on parsing; it thus makes a contribution to a formal description of the morphosyntax of the language. Because of their similarity, its results could be also used as a basis for a morphosyntactic description of other Sotho languages like Setswana. Moreover, as Taljard and Bosch (2006) have shown, the conjunctively written languages of South Africa, like Zulu have a similar morphosyntactic structure, when being described on a morpheme basis. This study could therefore be utilized as a first draft for describing the morphosyntax of some of these languages, too. The computational perspective entails separating data from rules, in other words, as we remain in the framework of generative grammar, separating a lexicon (chapter 2) containing Northern Sotho's linguistic units and their word classes (parts of speech (POS)) from morphosyntactic rules (chapter 3) that make use of them. An additional aim of the study is to provide some basics for describing features of verbal phrases by examining distributional patterns of the many ambiguous units contained in Northern Sotho's (specifically) verbal constellations. Such generalisation may be of use not only when working towards a more general linguistic model of the language, but also for part of speech disambiguation at an early stage of morphosyntactic analysis (chapter 4).

Lexical-Functional Grammar (LFG) is a well-known approach that has proven its value also for Bantu-Languages, the study therefore also aims at showing a possible implementation, i.e. developing a parser of at least a fragment of the descriptions utilising LFG (chapter 5). The final aim of this study (chapter 6) is to demonstrate possibilities and challenges when

translating automatically from Northern Sotho to English on the basis of the descriptions provided in this book.

7.2 Summary of results

7.2.1 Chapter 2: The word classes of Northern Sotho

As grammar rules usually describe linguistic constellations on the level of word classes or parts of speech (POS), a first step towards a morphosyntactic description of a language is the sorting of the language's words into such categories. Northern Sotho is a disjunctively written language: a single linguistic word may contain more than one orthographic word. Some of these orthographic "tokens" are bound morphemes that may not appear alone as they are dependent on others, while others can be categorised as being free, i.e. independent morphemes that may reign bound ones. When orthographic items appear "equally", i.e. next to each other on the surface level, while having a different status of independence, parsing is often preceded by a morphological analysis, identifying linguistic words (as e.g. Anderson and Kotzé (2006) describe it). In this study (cf. paragraph 1.4.2.1), we have however opted for not making any difference between bound and free morphemes for POS categorisation and therefore describe all orthographic tokens on one level, relying on Taljard et al. (2008). Such methodology, i.e. assigning parts of speech on the level of parsing, supports the disambiguation of the many ambiguous closed-class morphemes that appear in Northern Sotho.

Chapter 2 therefore introduces the most relevant linguistic units based on the set of Northern Sotho word classes, basically as defined by Taljard et al. (2008), however some of the word classes are described and thus labeled in more detail. As our study moreover aims at providing a detailed and more general overview of the parts of speech of Northern Sotho, a number of other publications were also considered of which the most important were the prescriptive descriptions of Northern Sotho grammar by Lombard (1985), Van Wyk et al. (1992), and Poulos and Louwrens (1994). These have been examined extensively and a number of our definitions are inherited from them, with one exception: a change of perspective when viewing noun class 15. This class contains "infinitives", which, if being examined from their internal structure, differ significantly from all other Northern Sotho nouns: they do not contain a noun stem and therefore, in terms of our approach, rather constitute verbal phrases. Therefore, we do not consider them as "nouns", i.e. substantives

in the sense of the word (cf. paragraph 2.2.2.5). A "noun" from class 15, as defined by all of our sources, e.g. Taljard et al. (2008), is not used in this study and is not contained in our tagset. However, if such a phrase should appear with the grammatical function of a subject, we view it as a nominal resulting from a conversion (a derivation process which does not add or delete affixes). This process is known for e.g. the German nominalisation of infinitives, e.g. in *Schwimmen ist gesund*, 'Swimming is healthy' (the Gerund of English, cf. paragraph 3.2.4 and Faaß and Prinsloo (forthcoming)).

Another important issue concerns the Northern Sotho subject concords, which are described with a finer distinction here than in existing literature (e.g. Poulos and Louwrens (1994)). Usually, one set of subject concords is defined that contains two different concords to be used for noun class 1: o_{CS01} and a_{CS01}. As it is described by the same sources, these are not interchangeable, each one appears in specific constellations. To ease the task of formulating unambiguous morphosyntactic rules on the basis of parts of speech, the set is therefore split into two (cf. Table 2.8 on page 41), thus extending the labels of the respective concords described in Taljard et al. (2008) to $o_{1\mathrm{CS01}}$ and $a_{2\mathrm{CS01}}$. The concords contained in the second set (the "consecutive" subject concords) as described by e.g. Poulos and Louwrens (1994) are labelled $3\mathrm{CS}_{\mathrm{class}}$ here. The whole set of word classes of Northern Sotho as used in the study is shown in Tables. 7.1 and 7.2.

Table 7.1: The tagset of Northern Sotho 1 / 2

Description	**tag 1st level**	**tag 2nd level**
concords		
1st subject class 1 – 10,14,15	1CS01 – 1CS10, 1CS14, 1CS15	–
2nd subject class 1 – 10,14,15	2CS01 – 2CS10, 2CS14, 2CS15	–
3rd subject class 1 – 10,14,15	3CS01 – 3CS10, 3CS14, 3CS15	–
1st personal subject	1CSPERS	1sg,2sg,1pl,2pl
2nd personal subject	2CSPERS	1sg,2sg,1pl,2pl
3rd personal subject	3CSPERS	1sg,2sg,1pl,2pl
1st locative subject	1CSLOC	–
2nd locative subject	2CSLOC	–
3rd locative subject	3CSLOC	–
1st indefinite subject	1CSINDEF	–
2nd indefinite subject	2CSINDEF	–
3rd indefinite subject	3CSINDEF	–
1st neutral subject	1CSNEUT	–
2nd neutral subject	2CSNEUT	–
3rd neutral subject	3CSNEUT	–
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	–

Table 7.2: The tagset of Northern Sotho 2 / 2

Description	**tag 1st level**	**tag 2nd level**
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
adjectives		
class 1 – 10, 14, 15	ADJ01 – 10, ADJ14, ADJ15	–, dim
locative	ADJLOC	–
verbals		
verb stem	V	itr, tr, dtr, sat-tr, hsat-dtr, aux
copula	VCOP	–, 01 – 10, 14
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 (page 65)	

Table 7.3: Lombard's modal system

General	Dependency	Ind./Mod.	Mood	Comments
Predicative				refers to a subject
	Independent			not dependent on other information, distinguishes tenses
		Indicating	Indicative	in main clauses
		Modifying		not in main clauses
			Situative	modifies the verb
			Relative	modifies the noun
	Dependent			dependent on other information, does not distinguish tenses
			Consecutive	chronologically dependent
			Subjunctive	causatively dependent
			Habitual	habitually dependent
Non-predicative				does not refer to a subject
			Imperative	
			Infinitive	

7.2.2 Chapter 3: A fragment of the grammar of Northern Sotho

A significant part of the grammar of Northern Sotho concerns a variety of verbal constellations, therefore the definitions of verbal phrases are discussed in great detail in chapter 3. We rely on Lombard's modal system (Lombard, 1985, p. 144) when identifying the categories of the different verbal phrases containing main verbs (cf. Table 3.1 on page 75, repeated as Table 7.3).

When examining the constellations of morphemes that precede the verb stem, morpheme clusters can be identified that contain information on verb-subject agreement, mood, tense, and positiveness or negativeness ("actuality", (Lombard, 1985, p. 139 et seq.) or "polarity", cf. paragraph 5.2.3). We group these elements as a Verbal Inflectional element (VIE). It depends solely on the semantics of the verb stem whether there are objects required

(valency). The given discourse is then responsible for the form (nominal or pronominal object concord) in which they appear[1]. We analyse the verb stems and their possible object(s) as building a Verbal Basic Phrase (VBP). In the case of a positive imperative mood, this VBP is the sole component of the Verbal Phrase (VP); in all other main verb constellations, the VP consists of a VIE followed by a VBP.

We also describe positional slots for VPs which are to be filled with specific elements: slot zero forming the VBP contains four positions: pos-1 contains an optional object concord, pos-0 contains the main verb stem, pos+1 and pos+2 may contain the verb stem's objects. The VIE consists of two slots: slot zero-1 may either contain a tense marker or remain empty, while slot zero-2 contains a subject concord and/or negation morphemes, as in Table 3.4 (page 79), repeated in Table 7.4.

Table 7.4: A schematic representation of the slot system

The slot system					
VIE		**VBP**			
zero-2 subject and/or negation marker	**zero-1** tense marker	**slot zero** **verb stem and its object(s)**			
		pos-1 object concord	**pos-0** **verb stem**	**pos+1** object 1	**pos+2** object 2

Verbal endings often are a decisive factor when determining the mood of Northern Sotho verbal phrases, therefore they form part of the morphosyntactic rules defined in chapter 3. Northern Sotho's main verb stems basically show four different endings (cf. paragraph 2.7.2): *–a* which constitutes a base form usually appearing in positive constellations, and *–e* appearing inter alia in negative and dependent constellations. The verbal ending *–go* appears with the relative, while endings *–ang* and *–eng* are found in the relative and im-

[1]Note that in paragraph 3.2.1.7, Northern Sotho verb stems are taken into consideration not only in the categories intransitive (no object required), transitive (one object required), or double transitive (two objects required), but also in fused forms of object concord and verb. *Nthuše!* 'help me!', for example, is regarded as a "saturated" transitive verb stem that does not require any external object, while "half saturated" double transitive verb stems like, e.g. *Mphe (puku)!* 'Give me (the book)!' still require one object to appear.

perative forms. However, *–e* also appears as the ending of the past tense forms, where the morpheme *–il–* is inserted between the verb stem and *–e*, as in *rekile* which is the past tense form of *reka* 'buy'. On the other hand, some non-standard verb stems can cause problems when being analysed on the basis of such descriptions. The main verb stem *re* 'say', for example, occurs in the same constellations that require the verb to end in *–a*. An additional problem is the past tense ending *–il–e*, because it appears in a number of allomorphs, e.g. in *les–itš–e*, the past tense form of *les–a* 'let loose/free' (cf. e.g. (Van Wyk et al., 1992, p. 47)). To solve this problem, this study introduces an additional attribute, "Verbal ending" (Vend), that is contained in the defined grammar rules and is hence to be assigned to each main verb stem entry of the lexicon with an appropriate value, e.g. Vend="*–a*" for the verbs *reka* and *re*, or Vend="*–ile*" for the verb stems *rekile* and *lesitše*. For sake of completeness, this attribute is also assigned to verb stems ending in *–ang* or *–eng* or *–go* with the respective values. Following this procedure, all verbal endings can be identified correctly by the parsing process (cf. chapter 4), independent of their surface form.

Based on these definitions and categorisations, paragraphs 3.2.3 to 3.2.11 describe all main verb constellations of Lombard's modal system. The summaries of these morphosyntactic rules are distributed over several tables: Table 7.5 (based on Table 3.19 of page 104) contains the independent indicative forms; Table 3.26 (page 121), repeated as Table 7.6 contains all modifying moods and, lastly, Table 3.29 (page 125), repeated as Table 7.7 describes the dependent constellations.

Table 7.5: A summary of the independent indicative forms

$^{\text{INDPRES}}\text{VP}$

	VIE		**VBP**		
descr.	**zero-2**	**zero-1**	**zero**	**zero+1**	**Vstem ends in**
ind.pres.pos.long	$1\text{CS}_{\text{categ}}$	MORPH_pres	VBP^{p}	\$.	*–a*
ind.pres.pos.short	$1\text{CS}_{\text{categ}}$		VBP	¬\$.	*–a*
ind.pres.neg.	$ga_{\text{MORPH\_neg}}$ $2\text{CS}_{\text{categ}}$		VBP		*–e*
ind.perf.pos.	$1\text{CS}_{\text{categ}}$		VBP		*–ile*
ind.perf.neg. 1	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$		VBP		*–a*
ind.perf.neg. 2	$ga_{\text{MORPH\_neg}}$ $se_{\text{MORPH\_neg}}$ $2\text{CS}_{\text{categ}}$		VBP		*–e*
ind.perf.neg. 3	$ga_{\text{MORPH\_neg}}$ $3\text{CS}_{\text{categ}}$		VBP		*–a*
ind.perf.neg. 4	$ga_{\text{MORPH\_neg}}$ $1\text{CS}_{\text{categ}}$ MORPH_past		VBP		*–a*
ind.fut.pos	$1\text{CS}_{\text{categ}}$	*tlo/tla* $_{\text{MORPH\_fut}}$	VBP		*–a*
ind.fut.neg	$2\text{CS}_{\text{categ}}$ $ka_{\text{MORPH\_pot}}$ $se_{\text{MORPH\_neg}}$		VBP		*–e*

Table 7.6: A summary of the modifying moods

	^{MOD}VP			
	^{MOD}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
sit.pres.pos.	$2CS_{categ}$		VBP	*–a*
sit.pres.neg.	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–e*
sit.perf.pos.	$2CS_{categ}$		VBP	*–ile*
sit.perf.neg.1	$2CS_{categ}$ se_{MORPH_neg} $3CS_{categ}$		VBP	*–a*
sit.perf.neg.2	$2CS_{categ}$ se_{MORPH_neg} $1CS_{categ}$		VBP	*–a*
sit.perf.neg.3	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–a*
sit.fut.pos.	$2CS_{categ}$	*tlo/tla* $_{MORPH\_fut}$	VBP	*–a*
sit.fut.neg.	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}		VBP	*–e*
rel.pres.pos.	$2CS_{categ}$		VBP	*–a* + ‘relative’
rel.pres.neg.	$2CS_{categ}$ sa_{MORPH_neg}		VBP	*–e* + ‘relative’
rel.perf.pos.	$2CS_{categ}$		VBP	*–ile* + *–a* + ‘relative’
rel.perf.neg.1	$2CS_{categ}$ *sego/seng* $_{MORPH\_neg}$ $3CS_{categ}$		VBP	*–a*
rel.perf.neg.2	$2CS_{categ}$ *sego/seng* $_{MORPH\_neg}$ $2CS_{categ}$		VBP	*–e*
rel.fut.pos.1	$2CS_{categ}$	*tlago/tlogo* $_{MORPH\_fut}$	VBP	*–a*
rel.fut.pos.2	$2CS_{categ}$	*tla/tlo* $_{MORPH\_fut}$	VBP	*-a* + ‘relative’
rel.fut.neg.1	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	*tlago/tlogo* $_{MORPH\_fut}$	VBP	*–a*
rel.fut.neg.2	$2CS_{categ}$ ka_{MORPH_pot} se_{MORPH_neg}	*tla/tlo* $_{MORPH\_fut}$	VBP	*–a* + ‘relative’
rel.fut.neg.3	$2CS_{categ}$ ka_{MORPH_pot}	se_{MORPH_neg}	VBP	*–a* + ‘relative’

Table 7.7: Summary of the dependent moods

	^{DEP}VP			
	^{DEP}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
cons.pos.	$3CS_{categ}$		VBP	*–e*
cons.neg.	$3CS_{categ}$ se_{MORPH_neg}		VBP	*–e*
suha.pos.	$2CS_{categ}$		VBP	*–e*
suha.neg.	$2CS_{categ}$ se_{MORPH_neg}		VBP	*–e*

The Northern Sotho copula, being the core of the copulative constellations, can be described by the following properties (see also paragraph 3.3.1 on page 125 and (Prinsloo, 2002, p. 28 et seq.)):

- copulas express relations between a subject and a complement, namely identification, description or association;
- there are two types of copulas for each of the defined relations: a stative and a dynamic type;
- copulas can contain the copulative particle ke_{PART_cop}, or a subject concord referring to a person or class;
- copulas can be multiword expressions like *e le*, *e se*, *o ba*, *o na*, etc., (we should like to add: of which some contain auxiliaries, like e_{CSNEUT} be_{V_AUX} e_{CSNEUT} le_{VCOP});
- copulas occur in moods.

Like the main verb stem, the copulative of Northern Sotho (VCOP) appears in a variety of moods and tenses which are summarised in Table 3.30, repeated here as Table 7.8. This study provides morphosyntactic rules for all of these from Table 3.32 (page 133) to Table 3.52 (page 167). These will not be repeated at this point for reasons of space.

Table 7.8: Overview of copulative constellations

Copulative	Identifying		Descriptive		Associative	
Category	stative	dynamic	stative	dynamic	stative	dynamic
Tense						
pres	×	×	×	×	×	×
perfect	×	×	×	×	×	×
fut		×		×		×
Mood						
indicative (pos/neg)	×	×	×	×	×	×
situative (pos/neg)	×	×	×	×	×	×
relative (pos/neg)	×	×	×	×	×	×
consecutive (pos/neg)		×		×		×
subjunctive (pos/neg)		×		×		×
habitual (pos/neg)		×		×		×
infinitive (pos/neg)		×		×		×
imperative (pos/neg)		×		×		×

As far as verbal phrases are concerned, this study also describes morphosyntactic rules for auxiliary verbal phrases (cf. paragraph 3.4) which take the previously described main verbal phrases as their complements, and also rules for hortative forms (paragraph 3.5.1), which show similar features. In order to describe the latter, the slot system is extended with an additional slot (zero-3) appearing to the left of the previously defined ones, as shown in Table 3.54 (page 171), repeated here as Table 7.9.

Table 7.9: The hortative constellation

description	**zero-3**	**zero-2 to zero**
	PART_hort	$^{\text{subjunctive}}$VP
Example	$a_{\text{PART\_hort}}$	$re_{\text{CSPERS\_2sg}}$ $reke_{\text{V\_tr}}$ $dipuku_{\text{N10}}$ 'let us buy books'

This study moreover contains a description of the potential (paragraph 3.5.2), the respective rules are contained in Table 3.55 (page 173), repeated as Table 7.10 on page 308.
In summary, chapter 3 shows that it is indeed possible to analyse a substantial part of Northern Sotho grammar with a view to parsing by describing the token constellations in a two-part positional slot system. In most cases, the value of the lexical attribute "Vend" assigned by the verb stem contained in the VBP combined with the POS/token constellation found in the VIE unambiguously identifies a specific mood (and tense) of the main verbal phrase in question. Chapter 3 moreover contains definitions of other constellations: noun phrases (NPs, section 3.8), and also the adjective phrases that may appear as nominals (APs, paragraph 3.8.4.2). Finally, particle phrases (paragraph 3.9) are described. Specifications of sentences of Northern Sotho conclude the chapter (section 3.10).

7.2.3 Chapter 4: Features of verbal phrases

So far, the Northern Sotho constellations have been described in a top-down manner, taking Lombard's system as a basis and defining the linguistic objects and how they combine to form them. Chapter 4 introduces electronic grammars (parsers) and begins with a right-to-left, bottom-up analysis of a complete sentence making use of a parts of speech lexicon and of the rules developed in the previous chapters. It thereby demonstrates that a bottom-up parser basically begins its analysis not with the rules present in its grammar, but with the tokens of the sentence in question.

	^{pot}VP			
	^{pot}**VIE**		**VBP**	
descr.	**zero-2**	**zero-1**	**zero**	**Vstem ends in**
pot.pres.ind/sit.pos.	**2CS$_{categ}$ *ka*/subsMORPH_pot**		VBP	*–a*
	example: *(ge)a*$_{2CS01}$ *ka*$_{MORPH\_pot}$ (when) `subj-3rd-cl1 pot` '(when) (s)he may speak'		*bolela*$_{V\_itr}$ speak	
pot.fut.pos	nonexistent			
pot.neg. 1 (ind.fut.neg) (sit.fut.neg)	**2CS$_{categ}$ *ka*/subsMORPH_pot *se*$_{MORPH\_neg}$**		VBP	*–e*
	example: *a*$_{2CS01}$ *ka*$_{MORPH\_pot}$ *se*$_{MORPH\_neg}$ `subj-3rd-cl1 pot neg` '(s)he might not speak'		*bolele*$_{V\_itr}$ speak	
pot.neg. 2	**2CS$_{categ}$ *ka*/subsMORPH_pot *se*$_{MORPH\_neg}$ *ke*$_{V\_aux}$ 3CS$_{categ}$**		VBP	*–a*
	example: *a*$_{2CS01}$ *ka*$_{MORPH\_pot}$ *se*$_{MORPH\_neg}$ *ke*$_{V\_aux}$ *a*$_{3CS01}$ `subj-3rd-cl1 pot neg neg subj-3rd-cl1` '(s)he might not speak'		*bolela*$_{V\_itr}$ speak	

Table 7.10: The potential forms

As many of the morphemes appearing in verbs of Northern Sotho are highly ambiguous concerning their parts of speech, such a parser can be expected to find a number of possible illicit analyses that should be abolished as soon as possible during the process. Chapter 4 is therefore an attempt to find generalisations on the contextual distribution of ambiguous morphemes in order to support their POS-disambiguation. The grammar rules described previously are explored in order to find patterns in the co-occurrence of parts of speech that could support the elimination of illicit analyses at an early stage of parsing. Such distributional data is however not only aiding disambiguation, it also may support the design of a future, more general linguistic modelling of Northern Sotho verbs.

7.2.4 Chapter 5: Implementation of a grammar fragment

For the sake of demonstration, parts of the grammar fragment described in chapter 3 have been implemented in the framework of Lexical-Functional Grammar (LFG, Kaplan and Bresnan (1982)). Section 5.1, contains a brief introduction to this constraint-based grammar theory and its formalism, demonstrated with an example analysis. This study utilises the Xerox Linguistic Environment (XLE, `http://www-lfg.stanford.edu/lfg`) as an implementation of LFG, and section 5.2 continues with the description of a lexicon and the necessary rules for this parser, defining the basic verbal phrase, the imperative, and the predicative independent mood (imperfect, perfect and future indicative) of Northern Sotho[2]. The chapter also contains a number of example analyses processed by the system.

7.2.5 Chapter 6: A basis for an automated translation

The final aim of this study is to provide a contrastive description of Northern Sotho and English that could be utilised for the development of a future machine translation (MT) system translating from the first to the latter language. Chapter 6 briefly introduces MT in terms of system architecture, interfaces between modules, and reversibility of resources and processes. Some current developments in MT are described as well. Beginning with section 6.3, general lexical and structural issues concerning MT from Northern Sotho to English are described alongside challenges and possible solutions, followed by translation descriptions for relevant parts of speech contained in the tag sets (Tables 7.1 and 7.2 on pages 298 and 7.2). Finally, chapter 6 offers a brief introduction to machine translation

[2]Note that an implementation of the infinitive constellations is added to the parser in Faaß and Prinsloo (forthcoming)

utilising XLE.

7.3 Conclusions and future work

This study contains a description of a grammar fragment of Northern Sotho, a partial implementation and a brief description of the challenges and possible solutions for translating Northern Sotho sentences into English. A number of problems have been solved, including defining an initial morphosyntactic description of the 'nouns' of class 15, and the creating the definition of an additional category of subject concords in order to provide a less ambiguous set of morphosyntactic rules. A slot system in which a number of Northern Sotho verbal constellations can be placed, was also designed. For machine translation into English, contrastive knowledge has been developed and a number of proposals were made that demonstrate at least some basics for translating Northern Sotho sentences into English. However, a few issues have been omitted: this study does not contain a number of other possible constellations, e.g. verbal phrases containing the deficient morphemes (MORPH_def) or the progressive morpheme $sa_{\text{MORPH\_prog}}$ (cf. paragraph 2.9.7). "Aspect prefixes" as described by Poulos and Louwrens (1994, p. 289 et seq.) have also been left out, hence the grammar fragment described is far from being complete and should be enhanced in the future. We expect the methodology defined here to also cater for these issues.

Only a small fragment of the morphosyntactic rules developed have been implemented in XLE so far, the current electronic grammar thus needs to be brought to a greater stage of completion. We however do not expect to have to change the methodology or the basic definitions developed in this study.

In general, the major obstacle for defining grammar rules on the basis of textual data is the lack of resources, i.e. tagged corpora (using an appropriate tagset). Within the framework of this study, we have already utilised some small parts of the *University of Pretoria Sepedi Corpus* (PSC), (cf. De Schryver and Prinsloo (2000)), however, the bulk of these data is still in the process of being cleaned up and annotated with parts of speech. When this goal is reached, a future project should look at corpus-based research on morphosyntactic issues of Northern Sotho aiming at the definition of grammar rules covering constellations of Northern Sotho on the basis of their frequency of occurrence. In other words: in parallel to the further development of parser rules based on language theory, more research on distributional information will lead to the development of data-based parser rules.

The adaption of Northern Sotho's morphosyntactic rules to simlar languages, like, e.g. Setswana or Southern Sotho, might not only add to the contrastive knowledge described so far, but could also form the basis of parsing systems being made available for all of the Sotho languages. On a longer time span, the rules could be of use for the description of the conjunctively written languages of South Africa, too.

Finally, a future machine translation project could make use of the contrastive descriptions given in this study. In combination with the use of parallel corpora, a hybrid system could be developed, translating bi-directionally between English and Northern Sotho.

Bibliography

W. Anderson and P.M. Kotzé. Finite state tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of northern sotho. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC-2006*, Genova, Italia, 2006. LREC. [CD-ROM].

A. Asudeh and I. Toivonen. Lexical-functional grammar. In B. Heine and H. Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 425 – 458. Oxford University Press, Oxford, 2009.

J.W. Backus. The syntax and semantics of the proposed international algebraic language of the ZURICH ACM-GAMM Conference. In *Information Processing: Proceedings of the International Converence on Information Processing, Paris*, pages 125 – 132. UNESCO, 1959.

L. Bloomfield. *Language*. Henry Holt and Co., New York, USA, 1933.

J. Bresnan. *Lexical-Functional Syntax*. Blackwell Publishing, Maiden, USA, Oxford, UK, Victoria, Australia, 2001.

H. Bußmann, editor. *Lexikon der Sprachwissenschaft*. Kröner, Stuttgart, Germany, 3rd edition, 2002.

M. Butt, T. Holloway King, M.E. Niño, and F. Segond. *A Grammar Writer's Cookbook*. CSLI Lecture Notes Number 95. CSLI Publications, Stanford, California, USA, 1999.

N. Chomsky. *Syntactic Structures*. Mouton, The Hague/Paris, 1957.

N. Chomsky. *Knowledge of Language*. Praeger, New York, 1986.

N. Chomsky. Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3):113 – 124, September 1956.

G.G. Corbett. Agreement: Terms and boundaries. In *Surrey Scholarship Online, Surrey Morphology Group (SMG) conference papers*, Surrey, England, 2001. University of Surry.

GM De Schryver, editor. *Oxford School Dictionary*. Oxford University Press South Africa, Cape Town, South Africa, 1st edition, 2007.

G.M. De Schryver and D.J. Prinsloo. The compilation of electronic corpora with special reference to the African languages. *Southern African Linguistics and Applied Language Studies*, 18(1 – 4):89 – 106, 2000.

G.M. De Schryver and E. Taljard. Locative trigrams in northern sotho, preceded by analyses on formative bigrams. *Linguistics: an interdisciplinary journal of the language sciences*, 44(Jan – Feb), 2006.

C.M. Doke. *The Problem of Word Division in Bantu with Special Reference to the Languages of Mashonaland*. Southern Rhodesia: Department of Native Development, 1921.

C.M. Doke. *The southern Bantu languages*. Published for the International African Institute by the Oxford University Press, London, New York, 1954.

M. Dorna. Maschinelle Übersetzung. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, and H. Langer, editors, *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Spektrum Akademischer Verlag, Heidelberg, 2001.

D. Embick and R. Noyer. Distributed Morphology and the Syntax-Morphology Interface. In G. Ramchand and C. Reiss, editors, *The Oxford Handbook of Linguistic Interfaces*, pages 289 – 324. Oxford University Press, Oxford, England, 2007.

M.C. Emele and M. Dorna. Ambiguitiy Preserving Machine Translation using Packed Representations. In *COLING-ACL*, pages 365 – 371, 1998.

K. Endemann. *Wörterbuch der Sotho-Sprache (Südafrika)*. Abhandlungen des Hamburgischen Kolonialinstituts, Band VII. L.Friedrichsen and Co. (Dr. L. & R. Friedrichsen), Hamburg, Germany, 1911.

G. Faaß. Eine LFG-Lernergrammatik für japanische DeutschlernerInnen (LFG - as an error-tolerant grammar for Japanese learners of German). Master's thesis, University of Stuttgart, 2005. URL `http://www.ims.uni-stuttgart.de/lehre/studentenarbeiten/fertig/Diplomarbeit\_Faasz.pdf`.

G. Faaß, U. Heid, E. Taljard, and D.J. Prinsloo. Part-of-Speech tagging in Northern Sotho: disambiguating polysemous function words. In *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages – AfLaT 2009*, pages 38 – 45. The 12th Conference of the European Chapter of the Association for Computational Linguistics, 30th March to 3rd April 2009.

G. Faaß and D.J. Prinsloo. A morpho-syntactic view on the infinitive of Northern Sotho. forthcoming.

J.E. Fenstad, P-K Halvorsen, T. Langholm, and J. van Benthem. *Situations, Language and Logic*. D. Reidel, Dordrecht, 1987.

A. Fraser and W. Wong. The Language Weaver Statistical Machine Translation Software System. In A. Farghaly, editor, *Arabic Computational Linguistics: Current Implementations*, chapter 4. CSLI Publications, 2008.

T. Givón. Some historical changes in the noun-class system of Bantu, their possible causes and wider implications. In K. Chin-Wu and H. Stahlke, editors, *Papers in African linguistics*, pages 33 – 54. Linguistic Research Inv., Carbondale and Edmonton, 1971.

H. Glück, editor. *Metzler Lexikon Sprache*. Metzler Verlag, Stuttgart, Germany, 1st edition, 2000.

Y. Graham, A. Bryl, and J.van Genabith. F-Structure Transfer-Based Statistical Machine Translation. In *Proceedings of the 14th International LFG Conference*, Cambridge, UK, 2009.

G Grefenstette. Tokenization. In H. van Halteren, editor, *Syntactic Wordclass Tagging*. Kluwer Academic Publishers, Dordrecht / Boston / London, 1994.

G. Grefenstette and P. Tapanainen. What is a word, what is a sentence? Technical report, Rank Xerox Research Center, Grenoble Labatory, 1994.

M. Guthrie. *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages, vol 1*. Gregg International, Farnborough, 1967.

M. Guthrie. *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages, vol 2*. Gregg International, Farnborough, 1971.

U. Heid, Prinsloo D.J., G. Faaß, and E. Taljard. Designing a noun guesser for part of speech tagging in Northern Sotho. *South African Journal of African Languages (SAJAL)*, 29 (1):1 – 19, 2009.

P. Hellwig. Parsing natürlicher Sprachen: Realisierungen. In I.S. Bátori, W. Lenders, and W. Putschke, editors, *Computational Linguistics - Computerlinguistik (HSK 4)*, pages 378 – 432. Walter de Gruyter, Berlin, New York, 1989.

W. Hutchins and H. Somers. *An Introduction to Machine Translation*. Academic Press Ltd., London, UK, 1992.

M. Jordaan-Weiss. Machine translation, terminology and the African languages in South Africa: An overview. In *Proceedings of the EAMT Machine Translation Workwhop TKE'96*, pages 95 – 98, Vienna, Australia, August 1996. European Association for Machine Translation.

D. Jurafsky and J.H. Martin. *Speech and Language Processing*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, New Jersey, USA, 2000.

R.M. Kaplan and J. Bresnan. Lexical-functional grammar: a formal system for grammatical representation. In J. Bresnan, editor, *The mental representation of Grammatical Relations*, pages 173 – 281. MIT Press, Cambridge, MA, 1982.

R.M. Kaplan, K. Netter, J. Wedekind, and A. Zaenen. Translation by structural correspondences. In M. Dalrymple, R.M. Kaplan, J.T. Maxwell III, and A. Zaenen, editors, *Formal Issues in Lexical-Functional Grammar*, number 47 in CSLI Lecture notes, pages 311 – 329. CSLI Publications, 1995.

F. Katamba. *Morphology*. Modern Linguistic Series. St. Martin's Press, New York, USA, 1993.

I.M. Kgosana. Aspects of pronominalisation in Northern Sotho. Master's thesis, University of Pretoria, Department of African Languages, 2005.

P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*, 2002. URL `http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl.pdf`.

I. Kosch. Die Onvoltooidheits-a in Noord-Sootho. Master's thesis, University of South Africa, 1985. unpublished MA dissertation.

I.M. Kosch. *Topics in Morphology in the African Language Context*. UNISA Press, South Africa, 2006.

B. Levin. Approaches to Lexical Semantic Representation. In D.E. Walker, A. Zampolli, and N. Calzolari, editors, *Automating the Lexicon*, pages 53 – 92. Oxford University Press, New York, USA, 1992.

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

L.J. Louwrens. *Aspects of the Northern Sotho Grammar*. Via Afrika, Pretoria, South Africa, 1991.

L.J. Louwrens. *Dictionary of Northern Sotho Grammatical Terms*. Via Afrika, Pretoria, South Africa, 1994.

J. Lyons. *Introduction to Theoretical Linguistics*. Cambridge University Press, London, England, 1968.

O.K. Matsepe. *Tša Ka Mafuri*. J.L. van Schaik, Pretoria, South Africa, 1974.

N.N. Mutaka. *An Introduction to African Linguistics*. LINCOM Handbooks in Linguistics 16. LINCOM EUROPA, München, 2000.

F.R. Palmer. *Semantics*. Cambridge University Press, 1986.

S. Petrick. Parsing. In S. Shapiro, D. Eckroth, and E. Vallasi, editors, *Encyclopedia of Artificial Intelligence*, volume II, pages 687 – 696. Wiley, New York, 1989.

C. Pollard and I.A. Sag. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press, Chicago, USA, 1994.

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

D.J. Prinsloo. The Lemmatization of Copulatives in Northern Sotho. In *Lexikos 12*, pages 21 – 43. Buro van die Wat, Stellenbosch, South Africa, 2002.

D.J. Prinsloo, G. Faaß, E. Taljard, and U. Heid. Designing a verb guesser for part of speech tagging in Northern Sotho. *South African Linguistics and Applied Language Studies (SALALS)*, 26(2):185 – 196, 2008.

D.J. Prinsloo and U. Heid. Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. In Isabella Ties, editor, *LULCL Lesser used languages and computational linguistics*, pages 97 – 113, Bozen/Bolzano, 27/28-10-2005 2005. Bozen: Eurac.

A. Ramsay. Computer-Aided Syntactic Design of Language Systems (computergestützte Verfahren zur syntaktischen Beschreibung von Sprache). In I.S. Bátori, W. Lenders, and W. Putschke, editors, *Computational Linguistics - Computerlinguistik (HSK4)*, pages 204 – 219. Walter de Gruyter, Berlin, New York, 1989.

S. Riezler and J.T. Maxwell III. Grammatical machine translation. In *Proceedings of HLT-ACL*, pages 248 – 255, New York, NY, 2006.

J. Roux and S.E. Bosch. Language resources and tools in Southern Africa. In *Proceedings of the Workshop on Networking the Development of Language Resources for African Languages. 5th International Conference on Language Resources and Evaluation*, pages 11 – 15, Genova, Italy, 22 May 2006 2006.

I.A. Sag, T. Wasow, and E.M. Bender. *Syntactic Theory*. CSLI Lecture Notes Number 152. CSLI, Stanford, California, USA, 2nd edition, 2003.

H. Schmid. Unsupervised Learning of Period Disambiguation for Tokenisation, 1994. URL `http://www.ims.uni-stuttgart.de/~schmid/`. Internal Report, IMS, University of Stuttgart.

C.T. Schütze. PP Attachment and Argumenthood. In *MIT Working Papers in Linguistics*, number 26 in Papers on Language Processing and Acquisition, pages 95 – 151. Massachusetts Institute of Technology (MIT), 1995.

L.P. Shapiro. Tutorial: An Introduction to Syntax. *Journal of Speech, Language and Hearing Research*, 40, April 1997. URL `http://chs.sdsu.edu/slhs/publications/shapiro533.pdf`.

E. Taljard. Unpublished study material for ba (hons) african languages programme. university of pretoria. BA(HONS) Sepedi, University of Pretoria, Department of African Languages, 1995.

E. Taljard. *Die kopulatief van Noord-Sotho: 'n nuwe perspektief*. PhD thesis, University of Pretoria, 1999.

E. Taljard and S.E. Bosch. A Comparison of Approaches to Word Class Tagging: Distinctively Versus Conjunctively Written Bantu Languages. *Nordic Journal of African Studies*, 15(4):428 – 442, 2006.

E. Taljard, G. Faaß, U. Heid, and D.J. Prinsloo. On the development of a tagset for Northern Sotho with special reference to the issue of standardization. *Literator – special edition on Human Language Technologies*, 29(1):111 – 137, 2008.

R.M. Thobakgale. *Khuetso ya O.K. Matsepe go bangwadi ba Sepedi*. PhD thesis, University of Pretoria, 2005.

D. Traum and N. Habash. Generation from lexical conceptual structures. In *NAACL-ANLP 2000 Workshop: Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*, pages 52 – 59, Seattle, May 2000. URL `http://www.mt-archive.info/ACL-2000-Traum.pdf`.

E.B. Van Wyk. *Woordverdeling in Noord-Sotho en Zulu (Word division in Northern Sotho and Zulu)*. PhD thesis, DLitt thesis. University of Pretoria, Pretoria, South Africa, 1958.

E.B. Van Wyk. Proclitic *bo* of Northern Sotho. *South African Journal of African Languages*, 7(1):34 – 42, 1987.

E.B. Van Wyk, P.S. Groenewald, D.J. Prinsloo, J.H.M. Kock, and E. Taljard. *Northern Sotho for first years*. J.L. van Schaik, Pretoria, South Africa, 1992.

B. Vauquois. A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress 1968*, pages 201–213, Edinburgh, 1968.

M. Wescoat. Practical Instructions for Working with the Formalism of Lexical Functional Grammar. Technical report, MS, Xerox Parc, California, 1989.

A. Wilkes. Oor die voornaamwoorde van zulu met besondere verwysing na die sogenaamde Demonstratiewe en Absolute voornamwoorde. *Studies in Bantoetale*, 3(1), 1976.

S. Zerbian. Questions in Northern Sotho. *ZASPIL - ZAS Papers in Linguistics*, 2006.

D. Ziervogel. *A Handbook of the Northern Sotho Language*. J.L. van Schaik, Pretoria, South Africa, 3rd edition, 1988.

D. Ziervogel and P.C. Mokgokong. *Groot Noord-Sotho-Woordeboek*. J.L. van Schaik, Pretoria, South Africa, 1st edition, 1975.