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.

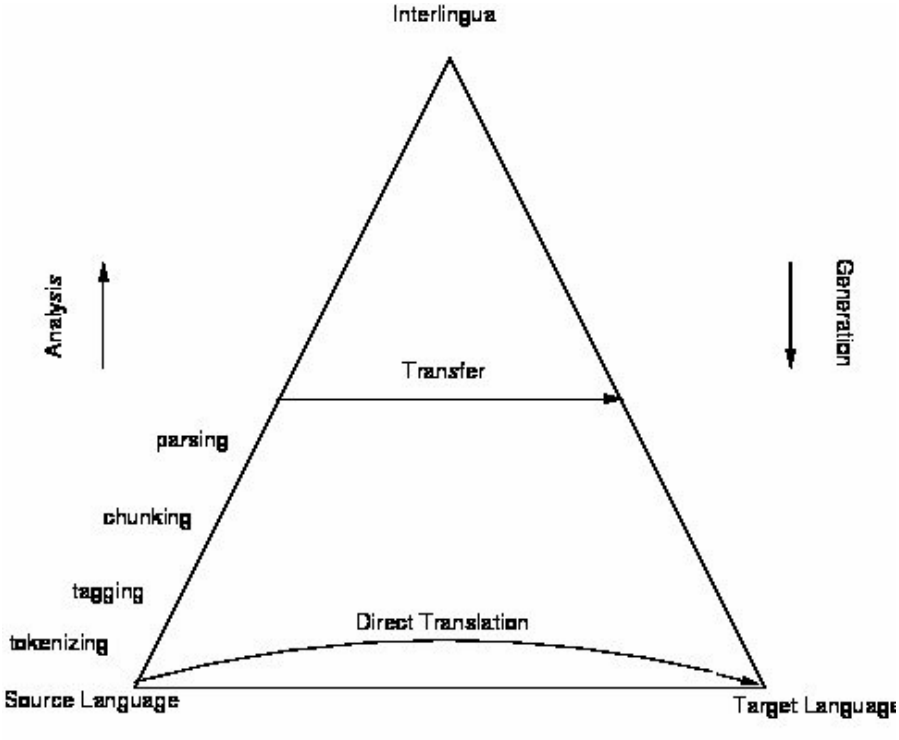

Figure 6.1: The MT triangle

[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].

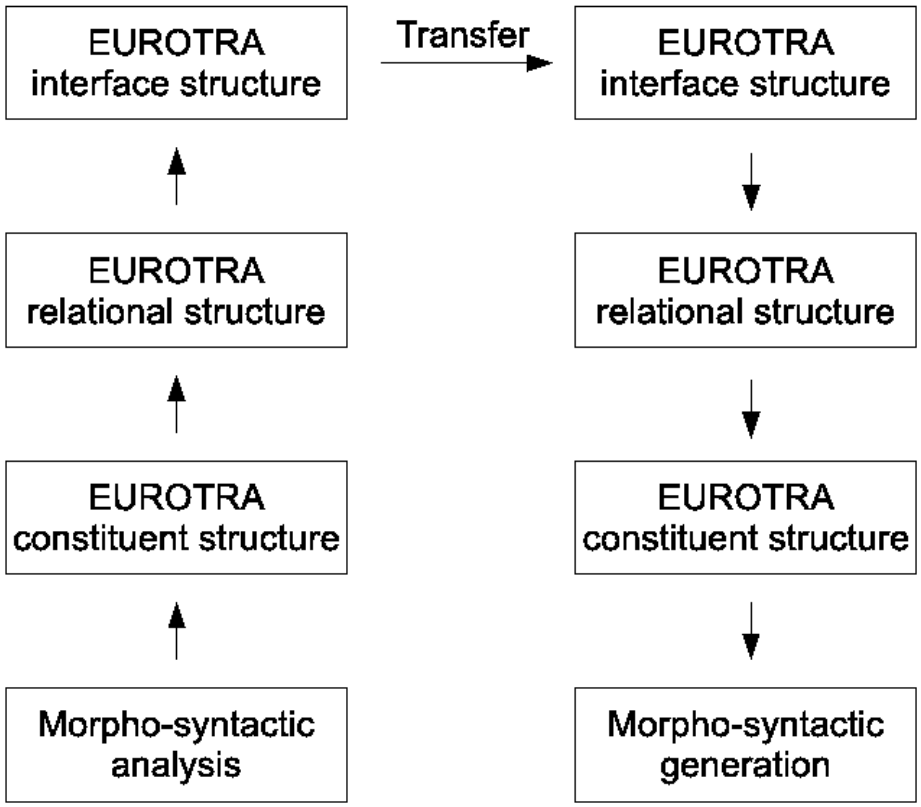

Figure 6.2: Translating with the transfer system EUROTRA

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.

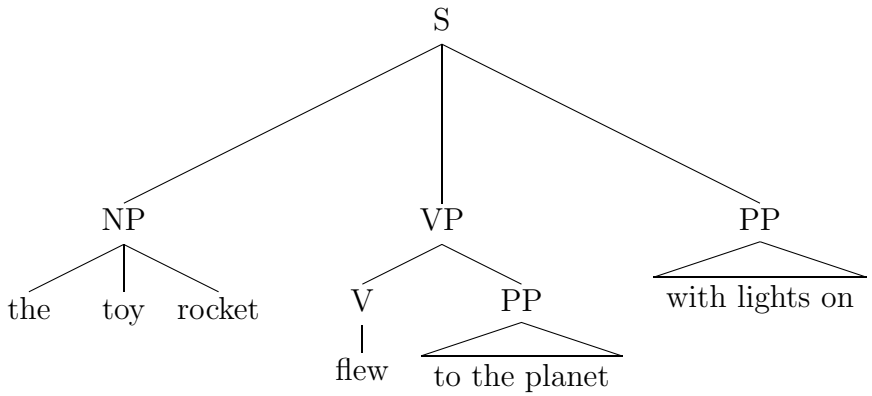

Figure 6.3: Example analysis: PP attachment (top node)

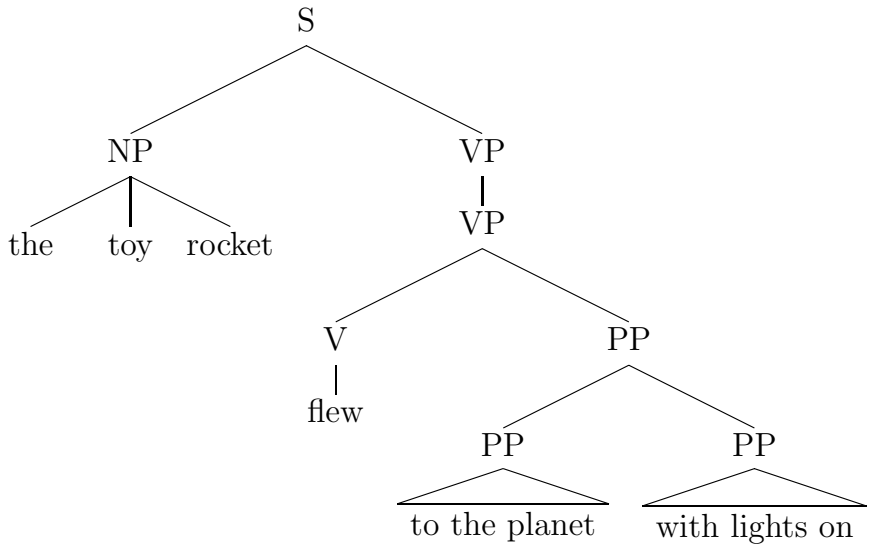

Figure 6.4: Example analysis: second PP attached to first PP

(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'

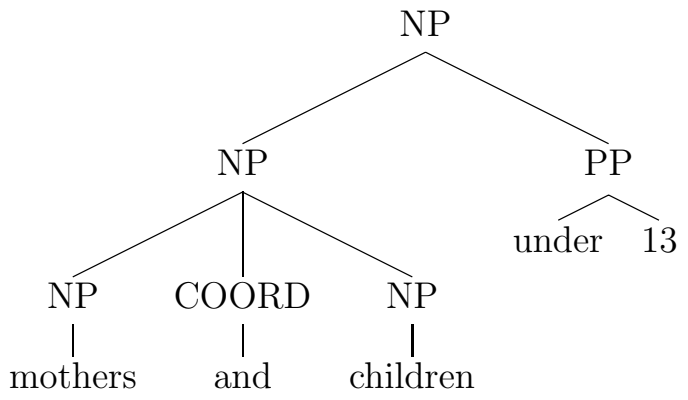

Figure 6.5: Example analysis: pp attachment (top node)

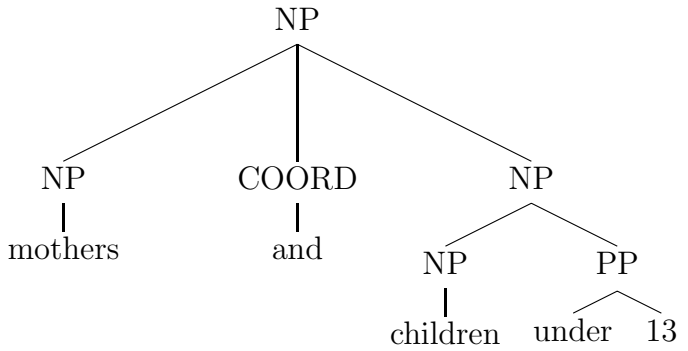

Figure 6.6: Example analysis: pp attachment (second NP of coordination)

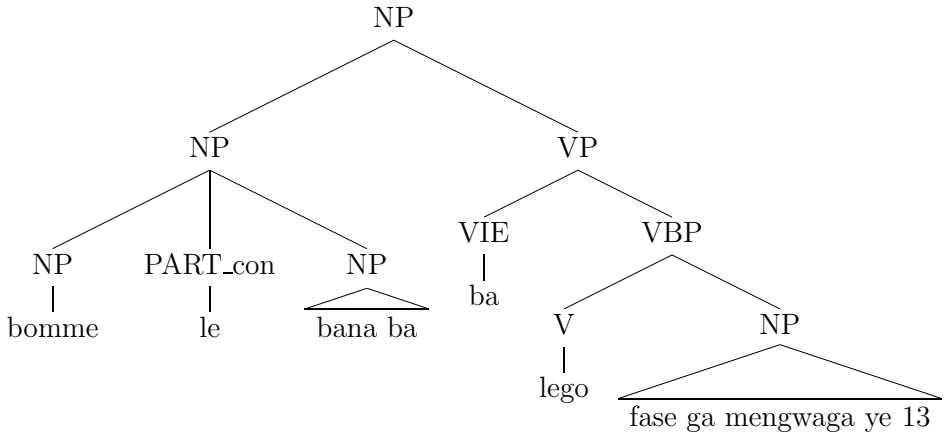

Figure 6.7: Example analysis: 'VP attachment' (top node)

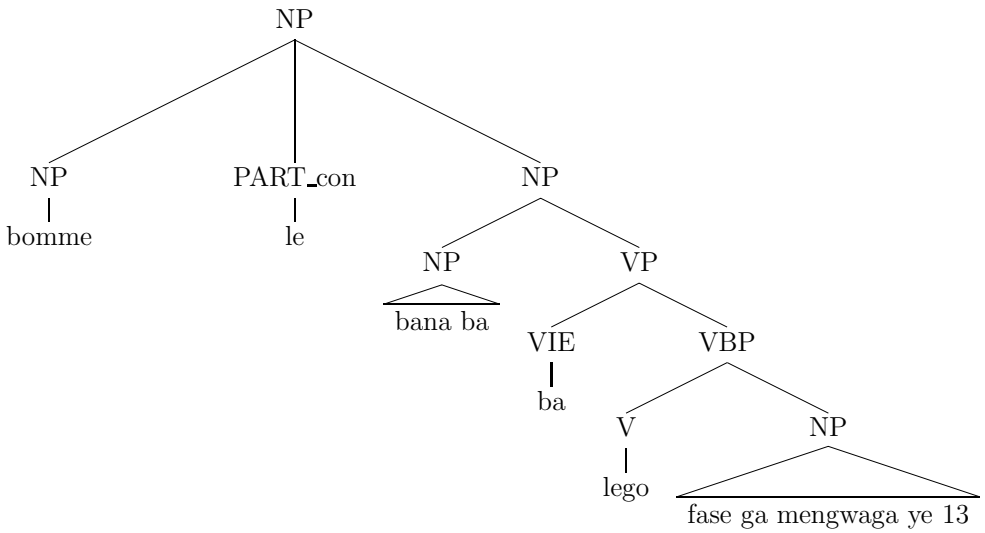

Figure 6.8: Example analysis: 'VP attachment' (second NP of coordination)

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,

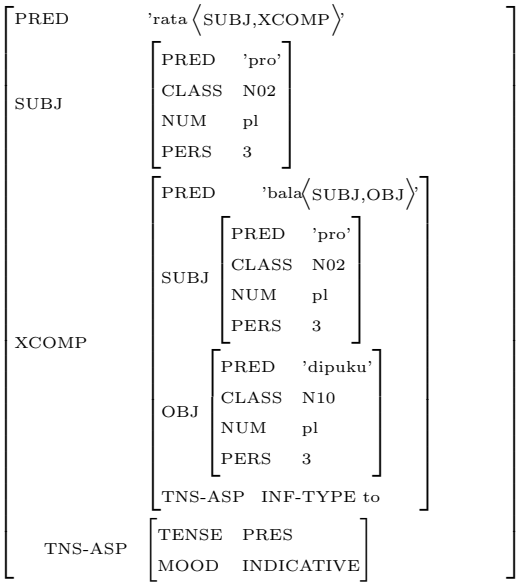

Figure 6.11: Simplified f-structure of *ba go bala dipuku*).

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.