Towards a computational morphological analysis of Setswana compounds

L. Pretorius & B. Viljoen
School of Computing
University of South Africa
PRETORIA
E-mail: pretol@unisa.ac.za
viljoe@unisa.ac.za

R. Pretorius & A. Berg
School of Languages
Potchefstroom Campus
North-West University
POTCHEFSTROOM
E-mail: Rigardt.Pretorius@nwu.ac.za
Ansu.Berg@nwu.ac.za

Abstract

Towards a computational morphological analysis of Setswana compounds

The development of a computational morphological analyser for Setswana necessitates the accurate modelling and implementation of, among others, compounding as a word formation process. Compounding is known to be an area of Setswana morphology that has sadly been neglected and still requires much investigation and research. The main purpose of this article is to investigate the formation of noun + noun compounds by computational morphological means in order to understand how this process should be formalised, modelled and subsequently implemented. In particular, an empirical study based on a collection of Setswana noun + noun compounds is reported on. The computational morphological analysis of these compounds revealed linguistic deviations from the standard morphological rules governing the formation of nouns and deverbatives. Examples, computational results and a discussion of the main findings are included.

Opsomming

'n Rekenaarmatige morfologiese analise van samestellings in Setswana

Die ontwikkeling van 'n rekenaarmatige morfologiese analiseerder vir Setswana noodsaak die akkurate modellering en implementering van, onder andere, samestelling as 'n woordvormingsproses. Die bestudering van samestellings in die Setswa-

namorfologie is 'n navorsingsveld wat agterweë gebly het en heelwat ondersoek en navorsing is nog nodig. Die hoofdoel van hierdie artikel is om die vorming van naamwoord + naamwoord-samestellings op 'n rekenaarmatige morfologiese wyse te ondersoek om sodoende te verstaan hoe hierdie proses ten beste geformaliseer, gemodelleer en gevolglik geïmplementeer kan word. In die besonder word berig oor 'n empiriese studie wat op 'n versameling Setswana naamwoord + naamwoordsamestellings gebaseer is. Die rekenaarmatige morfologiese analise van hierdie samestellings het linguistiese afwykings van die standaard morfologiese reëls wat vir die vorming van naamwoorde en deverbatiewe geld, aan die lig gebring. Voorbeelde, rekenaarresultate en 'n bespreking van die hoofbevindings is ook ingesluit.

1. Introduction

Setswana, as a semi-disjunctively written, agglutinating South African language, is characterised by a complex morphology. Serious technological development of Setswana therefore presupposes the availability of a computational morphological analyser (Pretorius *et al.*, 2005). In achieving the long term goal of developing a broad coverage computational morphological analyser for Setswana it is necessary to ensure that, ideally all and only valid Setswana words are analysed correctly. Among others this means that all productive word formation processes should be modelled accurately.

Setswana pronouns and some Setswana particles are class bound and are considered closed (morphologically unproductive) word classes. The rest of the particles can be listed. Setswana pronouns and particles may therefore be explicitly included in a computational morphological analyser. Adverbs, interjections and idiophones could also be listed. On the other hand, nouns and verbs constitute open (productive) classes, posing specific computational challenges in terms of the rich agglutinating morphology of Setswana. Further complexity is added to Setswana morphology by productive processes such as compounding. Compounding, the process of combining two or more autonomous words to form a new one, is particularly productive in the coining of new words, particularly terminology (Krüger, 2006:41), which is becoming increasingly important. However, the analysis of compounds in Setswana has received relatively little attention in the research literature on Setswana linguistics. Indeed, Krüger (2006:41) states that the analysis of Setswana compounds "is an area of morphology that is sadly

neglected and much investigation and research still have to be done".

In the development of a computational morphological analyser for Setswana it is not only necessary to model the formation processes of nouns and verbs accurately, but special attention also needs to be given to compounds. This suggests the following research question: "How should the morphological analysis of compounds be formalised or modelled and implemented so that compounding in Setswana can be handled as accurately as possible by the morphological analyser under development?"

The article is structured as follows: Section 1 consists of an introduction and a general contextualisation as well as the formulation of the research problem. Section 2 briefly introduces and gives an overview of compounding as a general word formation process. In section 3 compounding in Setswana, and specifically noun + noun compounding, is discussed. Section 4 is devoted to a brief outline of the finite-state modelling and implementation of Setswana noun and deverbative noun morphology, using the Xerox finite-state toolkit. Furthermore, certain methodological issues such as the empirical approach followed, iterated noun and deverbative analysis as the basic computational assumption, and the idea of judiciously relaxing certain morphophonological rules are addressed. Section 5 discusses the investigation procedure, the compilation of development data, and the results obtained by applying the Setswana morphological analyser to this data. Section 6 contains a discussion of the results and subsequent findings, insights, and questions that may be relevant for the accurate modelling and implementation of compounding in the Setswana morphological analyser. The final section draws conclusions and mentions possible future work.

2. Compounding as word formation process

A rather simplistic definition of a *compound* is "a word formed from two or more units that are themselves words, for example blackboard from black and board" (Matthews, 1997). As mentioned before, Krüger (2006:41) states that "compounding is a process whereby two (or more) autonomous words, mostly members of a word group, are combined to form a compound". Benczes (2006:8) defines the English compound as "a word that is made up of two or more elements, the first of which is either a word or a phrase, the second of which is a word". That is, compound = 2 or more elements, and in particular, compound = word/phrase + word.

Research into the formation of new words in the United States over a 50 year period indicates that the most productive word formation pattern (68%) for new words is that of compounding, with 90% of these compounds being nouns (Algeo, 1991). Algeo's reason for the high number of nominal compounds is that there are more entities to name than events or qualities. Another reason for the growing relevance of, among others, compounding as a word formation process is the so-called intellectualisation of indigenous languages. It is indeed a tendency in many countries that the creation of new terminologies is based on the deliberate and conscious use of word-formation patterns or methods such as borrowing, compounding, derivation, loan translation or calquing, semantic shift, blending, clipping, et cetera (Finlayson & Madiba, 2002).

Benczes (2006:2) indicates that "noun-noun compounds have been at the forefront of linguistic analysis for a number of well-founded reasons". Benczes further indicates that they form the largest group of compounds in English and that children learn to produce this type of compound the earliest, from around the age of two.

In a semantic context Benczes (2006:2) indicates that

> the most traditional and pervasive classification of compounds in linguistic literature is based on the work of Leonard Bloomfield (1933), who suggested that compounds fall into two main groups, namely endocentric and exocentric. In the endocentric group the compound is the hyponym of the head element, for example *apple tree* is a kind of tree. In the case of exocentric or 'headless' constructions, however, the compound is not a hyponym of the head element.

The meaning of compounds can be either metaphorical or metonymical. Indeed, "the most remarkable about these compounds is the diversity of semantic relationships that can exist between the two components on the one hand, and between the individual elements and the compound as a whole on the other" (Benczes, 2006:2).

Within the endocentric group a further distinction is made between right centred and left centred compounds. This refers to the position of the semantically prominent word or "head element" in the compound. Examples such as apple tree, walking stick, running shoes, bathing costume, et cetera suggest that compounds in English are right centred. In Afrikaans examples such as *bolplant* (*bulbous plant*), *voordeur* (*front door*) and *drukspyker* (*thumb nail*) suggest that Afrikaans compounds are usually also right centred (Combrinck, 1990:63-65). On the other hand, Setswana compounds are mainly

left centred since they conform to the structure of word groups in which the subject and qualificative in these word groups appears at the beginning of the phrase. This will become clear from examples in subsequent sections. Similarly, in Northern Sotho compounds are usually left centred since compound stems in Northern Sotho can also be formed as a result of the combination of two different words (word + word) or of the combination of a word and a word group (word + word group) (Laas, 1974:17).

3. Compounding in Setswana

Setswana compounds very often originate from an underlying word group structure, i.e. compound = word + word group. Krüger (2006: 293) defines a *word group* as "a combination of two or more words, constituents, components bound together semantically and structurally to form a cohesive unit". By the processes of reduction (of words), replacement and elision (of morphemes) compounds are subsequently formed.

Word groups have lexical as well as functional components. The lexical component includes the individual words and word groups in so far as they belong to word classes and word group classes. The functional component includes the various functions of the lexical components in their mutual semantic-syntactic relations to each other. The functional classes include subject, object, antecedent, qualificative/qualifier, introductory member, complement, descriptive, predicate, head member and appositional member and coordinate member (Krüger, 2006). In the description of word groups in Setswana, word classes and functional classes are the two points of reference in the identification of the underlying structure. This is illustrated by the following example:

Mosadi	*o tla reka*	*dijo* (individual words)
The woman	*will buy*	*food*
noun	verb	noun (word classes)
subject	predicate	object (syntactic functions)

Setswana compounds are nouns, and the following underlying word group structures may be observed in them:

- **Possessive groups**

This is a nominal group. On the function class level it consists of an antecedent followed by a qualificative group. On the word class level

it consists of a noun followed by a possessive relation (particle) word group. The first noun is the antecedent and it is qualified by a possessive group. The possessive particle is reduced, for example:

ditselanko (air passages/nasal airways)	-	*ditsela tsa nko (roads of the nose)*
boemadikepe (harbour)	-	*boema ba dikepe (standing place of the ships)*
bokgabosedumedi (religious decoration)	-	*bokgabo ba sedumedi (decoration of religion)*
bolwetsisukiri (diabetes)	-	*bolwetsi ba sukiri (illness of sugar)*
botlholemadi (blood poisoning)	-	*botlhole ba madi (poison of the blood)*

- **Infinitive groups**

This group originates from an infinitive verbal group. On the function class level it consists of a predicate followed by an object. On the word class level it consists of an infinitive verb followed by a noun. The infinitive verb is nominalised to the preferred/relevant noun class.

Examples:

modulasetulo (chairperson)	-	*go dula setulo (to sit on the chair)*
moepapitso (convenor)	-	*go epa pitso (to convene a meeting)*
molomatsebe (informer/informant)	-	*bokgabo ba sedumedi (decoration of religion)*
bolwetsisukiri (diabetes)	-	*go loma tsebe (to bite (some-one's) ear)*

- **Adjective relative groups**

This is a nominal group. On the function class level it consists of an antecedent followed by a qualificative group. On the word class level it consists of a noun followed by a qualificative relation (particle) word group. The qualificative particle is reduced, for example:

moririmosweu (gray haired) - *moriri o mosweu (hair that is grey/white)*

moepapitso (convenor) - *meno a masweu (teeth that are white)*

- **Verbal relative groups**

This is a nominal group. On the function class level it consists of an antecedent followed by a qualificative group. On the word class level it consists of a noun followed by a qualificative relation (particle) word group. The qualificative particle is reduced, for example:

sekepekgogi (tugboat) - *sekepe se se gogang (ship that pulls)*

leotwanakgogi (pulley) - *leotwana le le gogang (wheel that pulls)*

- **Associative (infinitive) groups**

This group originated from two infinitive verbs combined by the associative relation (particle) word *le*. The verbs are nominalised and the associative particle is reduced, for example:

temothuo (agriculture) - *go lema le go rua (the ploughing and the keeping)*

- **Miscellaneous group**

In this group the identification of an underlying word group is not obvious. In the example, *sefikantswe* (*lefika* + *lentswe*) (*stone mound*), the first noun is shifted from the *le-* class to the *se-* class while the prefix of the second noun is omitted. It also seems as if the components appear in the wrong order. Should the compound have originated from a possessive group, it would have been *lentswem-afika* from the group *lentswe la mafika* (*mound of stones*).

There are also compounds which seem to have developed from a simple juxtaposition. Compounds indicating place names or locations and the direction of the wind could be mentioned here. Consider the following:

aferikaborwa (aferika + borwa) (South Africa)

bokonebophirima (bokone + bophirima) (North-West)

bokonebotlhaba (bokone + botlhaba) (North-East)

Although compounds very often originate from an underlying word group, this is not visible in the surface form. The surface form essentially exhibits the compound = word + word structure. Therefore compounds in Setswana may be viewed as combinations of various word classes. Cole (1955:117) distinguishes between six types of compounds, viz.

- **noun + noun**

kgogonoka (*African coot*); *leebarope* (*rock pigeon*); *nogapotsane* (*large unidentified snake – said to bleat like a goat – probably mythical*)

- **noun + qualificative**

monnamogolo (*old man*); *mmemogolo* (*grandmother*); *tautona* (*king*, *chief*)

- **noun + adverb**

motshegare (*midday*, *during the day*); *bosigogare* (*midnight*); *tswelelopele* (*progress*, *advancement*)

- **verb stem + noun**

molalathakadu (*large hole* or *burrow*); *leaparapelo* (*pericardium*); *bophirimatsatsi* (*west*); *molomatsebe* (*secret informer*)

- **verb stem + adverb**

motlapele (*pioneer*); *motsalwapele* (*first-born*)

- **miscellaneous compounds**

modulaesi (hermit, *one who sits alone*); *moswaoeme* (*a dead tree – still standing*)

In the examples given by Cole in the miscellaneous class a nominal member is included in all instances. The verb stem + adverb type is the only one that includes a constituent that is not nominal. Since deverbatives rankshifted to the category of nouns they are included as nouns.

Regarding the frequency of occurrence of the various compound types, Ungerer (1983:108) observes that 91% (1 225 of 1 348) of the Zulu compounds in his data are compounds that involve at least one nominal constituent. According to Landsberg (1987:266-267) approximately 70% of compounds in Northern Sotho consist of noun + noun, verb + noun and noun + verb combinations. Of these, noun +

noun combinations constitute 63%. Furthermore, noun + adjective compounds form 12%, and noun-locative noun compounds 6% of the nouns in Landsberg's (1987) data. Therefore, Setswana, as a member of the Sotho language family, should display a similar trend. This is the reason for focusing on noun + noun compounds in this exploratory study.

In the sections above compounds were discussed in terms of so-called words. However, for the purposes of (computational) morphological analysis it makes sense to also consider the morphological structure of compounds since the modelling, as discussed in the next section, takes place at the morphophonological level. It should be noted that a detailed discussion of Setswana morphology falls outside the scope of this article. The interested reader is referred to Krüger (2006). Nevertheless, certain central aspects are mentioned briefly. Krüger (1994:18) indicates that Setswana words (including compound words) may include grammatical morphemes (prefixes and suffixes), root and stem morphemes. The essential difference between a root and a stem is as follows: A *root* is the semantic core of a word, it does not include a grammatical morpheme, it has no word correlate and it is dependent like the prefixes and suffixes (grammatical morphemes). A *stem*, on the other hand, has a word correlate and may include one or more grammatical morphemes (Krüger, 1994:18; Posthumus, 1994:30 & Pretorius, 2000:55-57). In Setswana four types of stems are distinguished, namely compound stems as well as simple, complex and reduplicated stems (Pretorius & Berg, 2005). Examples of compound stems are *modulasetulo* (*chair person*), *khudutlôu* (*big tortoise*) and *leebarope* (*rock pigeon*).

Different approaches exist with reference to the morphological analysis of Setswana words (Pretorius & Berg, 2005). The hierarchical analysis entails the systematic and repeated identification of meaningful relationships between components based on word-formation processes. Hierarchical arrangement applies to the paradigmatic view and is therefore not a linear arrangement technique. On the other hand, the grammatical (neutral) analysis of the noun does not take into account the mutual relationships between the morphemes. The grammatical analysis does not include any hierarchical intermediate levels and the syntagmatic structure is not considered. Examples are *monna* (*man*) (consists of *mo-* + *-nna*) and *ditlhare* (*trees*) (consists of *di-* + *-tlhare*).

The linear approach inherent to finite-state computational morphological analysis (see, for example, Beesley & Karttunen, 2003) is closely related to grammatical analysis, but may be extended by

further levels of parsing to a representation that is equivalent to the two-dimensional hierarchical morpheme structure.

4. Finite-state computational approach to noun + noun compounding

Finite-state methods are well known as a preferred and state-of-the-art approach to computational morphology. Numerous software packages and modelling and implementation toolkits are available, of which the Xerox integrated set of finite-state software tools for creating finite-state networks is arguably among the most sophisticated and mature (Beesley & Karttunen, 2003). The main challenges of computational morphology are the accurate modelling and implementation of the morphotactics and the morphophonological alternation rules that characterise word formation in a specific language. These processes and rules are modelled and implemented with an appropriate high-level declarative language for specifying morphotactics and lexicons (*lexc*) and the powerful Xerox regular expression engine for writing rules (*xfst*). The resulting finite-state networks are finally composed into one network that constitutes the morphological analyser (see, e.g. Pretorius & Bosch, 2002).

The development of a morphological analyser requires that all and only the valid words of Setswana be recognised and analysed. This means that not only should the morphotactics and the alternation rules be accurately and completely implemented, but all and only the (basic) verb and noun roots from which Setswana words are formed should be included in the morphological analyser.

As practical design and implementation approach we systematically include the morphotactics and alternation rules by word category. The current prototype covers the Setswana noun (Pretorius *et al.*, 2005), as well as Setswana deverbatives, including verbal suffixes. The reflexive and certain conjunctively written object concords will be incorporated as a next step. Although parts of the verb morphology have been included, a discussion thereof falls outside the scope of this article. Regarding the noun and verb roots the current prototype contains 574 noun roots and 255 verb roots. These lists are expanded as the process of development and testing proceeds. For the purposes of this article it is important to note that the current prototype is sufficiently complete in order to undertake preliminary exploratory work regarding noun + noun compounds. However, the

mining of corpora and the acquisition of any available noun and verb root lists constitute important future work.

Our methodology is as follows: The basic computational assumption is that compounding is like stringing words together, i.e. compound = word + word, which implies that the existing noun and deverbative computational morphological analysis may be iterated. A data-driven approach is followed in the sense that empirical data in the form of available Setswana compounds are analysed in order to investigate deviations, if any, from the basic computational assumption, to gain linguistically informed insights; and to draw plausible conclusions. For this purpose certain rules that govern noun and deverbative morphology are systematically relaxed, as described in the subsequent procedure.

5. Investigation procedure, data and results

5.1 Procedure

- The starting point is
  - a Setswana morphological analyser prototype that is able to analyse non-compound nouns and deverbatives; and
  - the basic (simplifying) computational assumption that compounds are just words strung together or juxtaposed, i.e. compound = word + word.
- Compile a linguistically sound and sufficiently large data set of Setswana compounds.
- Apply the analyser to this data.
- Study the results, i.e. the compounds that are successfully analysed and the ones that are not. Attempt to find trends and understand why the analyses failed.
- Adapt the analyser by judiciously relaxing certain morphophonological rules for exploratory purposes and obtain frequencies of observed phenomena and trends. The results obtained in step 4 above suggested the following:
  - Allow for the absence of the class prefix in the second constituent of the compound,
    e.g. *leebarope* (*rock pigeon*) – *leeba* + *marope*.

- Allow for *-a* as deverbative suffix in the absence of the passive suffix *-iw* or *-w*,
  e.g. *moepapitso* (*convenor*) – *go epa pitso*.

- Allow for *-i* as deverbative suffix instead of *-o* in class 9-10,
  e.g. *maatlakgogedi* (*magnetic force*) – *maatla + go goga.*

5.2 Compilation of data

In the compilation of the data several sources were employed. Compounds were collected from, among others, Landsberg's study on Northern Sotho (Sepedi), textbooks, dictionaries and terminology lists supplied to us by the Department of Arts and Culture. For practical reasons not all sources were utilised to the full. The lists were then proofread by a mother-tongue speaker. One as yet untapped future source of compounds is large untagged language corpora. However, for the purposes of this exploratory investigation the list of approximately 700 Setswana compounds were considered adequate. These compounds consisted of two constituents although, in principle it is possible to apply the procedure to compounds with more constituents.

5.3 Morphological analysis statistics

The original list of compounds was utilised as follows: As a first step the lexical items or single words constituting the compounds were separately considered, sorted and all duplicates were removed. This yielded 737 single words of which 96% (i.e. 705) were successfully analysed by means of the existing analyser prototype. All compounds involving any of the 32 failures were removed from the original compound list, leaving 689 compounds. This was done to prevent single word failures from influencing the compound analysis statistics. Of these 689 compounds 91% (i.e. 630) were successfully analysed. In the case of multiple analyses the linguistically correct one was retained. The failures (9%) were due to Setswana compound morphotactics and morphophonological alternation rules not yet implemented. These 630 correct analyses formed the basis of our statistics in tables 1 and 2 below and the findings in section 6.

Table 1: Morphological statistical information

Compound information	Total number (%)
Noun + noun compounds	630
Lexical items occurring in compounds	705
Unique noun roots	321
Unique verb roots	174
Noun root + noun root	293 (47%)
Noun root + verb root	146 (23%)
Verb root + noun root	90 (14%)
Verb root + verb root	101 (16%)
Repeated applicative extension	4
Causative extension	39
Applicative extension	24
Reciprocal extension	4
Passive extension	23
Causative + passive	3
Applicative + passive	1
Causative + reciprocal	1

Table 2: Deviations from rules

Deviation	Number of compounds
no class prefix in second constituent	48
class 1-2	1
class 3-4	1
class 5-6	42
class 7-8	4

no class prefix in second constituent of noun root + noun root compounds	38
deviation from deverbative formation rules	99
deverbative suffix -*a* without passive in both constituents	2
deverbative suffix -*a* without passive in first constituent	50
deverbative suffix -*a* without passive in second constituent	17
deverbative suffix -*i* with class 9-10 (impersonal)	31

6. Discussion of results

The notational conventions used in the computational morphological analyses of compounds given in this section are as follows:

`+`: Concatenation of morphemes

`NPreX`: Noun prefix of class `X`

`[xyz]`: Root xyz

`DeverbSuf`: Deverbative suffix

`Appl`: Applicative extension

`Caus`: Causative extension

`[NoPrefix]+[xyz[ClassX]]`: Noun prefix of class `X` to root xyz absent

`DeverbSuf[RELAXED]`: *-a* appears as deverbative suffix in the absence of the passive extension

`DeverbSuf[RELAXEDending]`: irregular ending *-i* occurs instead of regular *-o*.

6.1 Regular compound = word + word formation

484 compounds (77%) in the data set satisfy the basic computational assumption.

Examples:

bukalesika (*family register*)

`Npre9+[buka]+NPre5+[sika]:Compound`

Notice that the first word is analysed as `Npre9+[buka]` and the second as `NPre5+[sika]`, i.e. the analysis has the form word + word.

botlholemadi (*blood poisoning*)

NPre14+[tlhole]+NPre6+[di]:Compound

kagomatlapa (*stone masonry*)

NPre9+[ag]+DeverbSuf+NPre6+[tlapa]:Compound

bukathalelo (*drawing book*)

NPre9+[buka]+NPre9+[thal]+Appl+DeverbSuf:
Compound

bukaditiragalo (*log book*)

NPre9+[buka]+NPre9+[diragal]+DeverbSuf:Compound

katlegokgwebo (*wealth*)

NPre9+[atleg]+DeverbSuf+NPre9+[kgweb]+DeverbSuf:
Compound

bokgabosedumedi (*religious art*)

NPre14+[kgab]+DeverbSuf+NPre7+[dumel]+DeverbSuf:
Compound

6.2 Absence of noun prefix

Cole (1955:118) contends that "[i]n the formation of compounds consisting of noun plus noun, other than those having reduplicated stems, the prefix of the second noun is omitted". However, as can be seen in Table 2 only 38 of the 293 noun root + noun root compounds support Cole's (1955:118) contention. If the first constituent is allowed to have either a noun root or a verb root, only 48 out of 383 second constituents does not have a noun prefix. This raises a question as to the general applicability of Cole's (1955:118) statement. It is noted that in our data this elision occurs mainly with class 5-6 noun roots (87,5%).

Examples:

leebarope (*rock pigeon*)

NPre5+[eba]+[NoPrefix]+[rope[Class5]:Compound

bojangwatle (*sea grass*)

NPre14+[jang]+[NoPrefix]+[watle[Class5]]:
Compound

borohuba (*hand-brace*)

NPre9+[boro]+[NoPrefix]+[huba[Class7]]:Compound

kgabisotlapa (*stone mosaic*)

NPre9+[kgab]+Caus+DeverbSuf+[NoPrefix]+[tlapa
[Class5]]:Compound

6.3 Deviation from deverbative formation rules

The data set contains of 337 compounds of which at least one of the constituents is a deverbative noun. In 238 cases these compounds satisfy the basic computational assumption. Examples are *kago-matlapa* (*stone building*) – *kago ya matlapa*, *kagodikgwa* (*building of forests – department of forestry*), *kekisontwa* (*mocking of fighting/ war – mock fight*).

In these examples the first constituents are deverbatives, and they are all the antecedent of a possessive group. However, a significant number, namely 99 compounds in the data set, violate the basic computational assumption in the sense that one or both of the constituents differ from their single word representations.

6.3.1 *-a* as deverbative suffix only when preceded by passive extension *-iw* or *-w*

Krüger (2006:111) indicates that "the suffix *-a* usually occurs when personal nouns are formed from passive stems …". In the context of single words he maintains that it is possible for personal deverbatives like *modisa* (*shepherd*) and *morena* (*master*/*lord*) to take an *-a* ending in limited cases. The deverbative suffix is irregular in these cases.

Regarding compounds specifically, we found 52 first constituents where *-a* appears as deverbative suffix in the absence of the extension. A possible explanation could be that these compounds originate from infinitive groups where the infinitive prefix was replaced by a noun class prefix as done in the formation of deverbatives. The (original) *-a* ending was retained.

Examples:

moepapitso (*convenor*)

```
NPre1+[ep]+DeverbSuf[RELAXED]+NPre9+[bits]+Deve
rbSuf:Compound
```

morekisadibuka (*book seller*)

```
NPre1+[rek]+Caus+DeverbSuf[RELAXED]+NPre10+[buk
a]:Compound
```

6.3.2 Class 5 deverbatives normally ending in *-o*

The data set contains 165 compounds with deverbative as second constituent that conform to this rule. Examples are *bukaditiragalo* (*diary*) – *buka ya ditiragalo*, *bukapoloko* (*savings book*) – *buka ya*

poloko, *bukathoto* (*stock register*) – *buka ya (di)thoto*). However, an irregular ending *-i* occurs in 31 cases.

Examples:

maatlakgogedi (*magnetism*)

```
NPre6+[atla]+NPre9+[gog]+Appl+DeverbSuf[RELAXED
ending]:Compound
```

sekepekgogi (*tug boat*)

```
NPre7+[kepe]+NPre9+[gog]+DeverbSuf[RELAXED
ending]:Compound
```

According to Setswana textbook morphology the *-i* ending should not appear in the deverbative as non-compound stem. In Northern Sotho, however, Kotze and Anderson (2005:62) make provision for class 9-10 deverbatives to take an *-i* ending.

Compounds with irregular endings might have developed from verbal relative groups, for example, *sekepe se se gogang* (*ship that pulls*), and *leotwana le le gogang* (*wheel that pulls*), *borwana e e atolosang*, (*drill that rooms*), *borwana e e fetlhang* (*drill that enlarges*), *galase e e godisang* (*glass that enlarges*), or possessive groups with infinitive possessors, for example, *sekepe sa go goga* (*ship to pull with*), *borwana ya go atolosa* (*drill with which to room*).

6.4 Consequences of relaxed rules

Our empirical investigation does not confirm the statement of Cole (1955) concerning the absence of prefixes in compounds. Results also suggest that deverbative endings in compounds may not adhere to the same formation patterns as those occurring in non-compound deverbatives. Furthermore, there is a question as to the status of deverbatives in cases where they are the first constituent of a compound and originate from an infinitive group such as *modula-setulo* (*chair person*). These questions require further investigation and form part of future work.

6.5 Miscellaneous observations and questions

In the course of our experiments a number of other issues emerged that warrant further investigation, both linguistically and computationally. We outline three pertinent ones:

- **Stems that seem to occur in the reverse order**

sefikantswe (*cairn, mound of stones*) – *lefika + ntswe*
The class prefix is *se-* instead of *le-*. The word group *lentswe la mafika* (*mound of stones*) conform better to the grammatical rules. Another such example found in our data is *molapokeledi* (*river flow*), which might have been taken directly from Afrikaans, as the correct form in Setswana seems to be *kelelo ya molapo* or *kelelomolapo*.

- **Roots appearing in dialectical form**

In a compound such as *modirikhudugi* (*migrant labourer*), *fuduga* (*migrate*) was spelt as *huduga* (*f/h*) in the original underlying word. The *h-* in *huduga* differs from the *f-* in *fuduga* when the verb is nominalised to class 9. It is important to pay attention to the appearance of words in dialectical form.

- **Lack of nasalisation**

Although the regular form *sekepekgogi* (*tug boat*) appears in the literature, the unnasalised form *sekepegogi* (the *g* in *-gogi*) also occurs in the data set. Another example where nasalisation does not occur is *kgapetlaomi* (the *o* in *-omi*) (*dry ice*). On the other hand, unexpected nasalisation also occurs. Cole (1955:119) indicates that in examples with relative stems such as *pelokgale* (*fierceness*) and *pelotshula* (*evil heartedness*) it is not clear why the initial consonant of the second stem is strengthened and speculates that it might be in sympathetic nasalisation due to the class of the first stem.

7. Conclusion and future work

Reiterating, the research question was: “How should the morphological analysis of compounds be formalised or modelled and implemented so that compounding in Setswana can be handled as accurately as possible by the morphological analyser under development?” In order to address this question for noun + noun compounds an empirical study was conducted.

From a computational perspective the experiments confirmed the accuracy of the noun and deverbative morphology as modelled and implemented in the current Setswana analyser prototype. However, when applying the basic computational assumption (compound = word + word) to the analysis of compounds it transpired that certain word formation rules do not strictly apply in compounding. This suggested the relaxation of these rules in order to gain insight into the processes that govern the formation of noun + noun compounds.

The extension of the empirical study to include larger data sets as well as various other types of compounds forms part of future work.

A noteworthy outcome is the refutation of Cole's claim that "[i]n the formation of compounds consisting of noun plus noun, other than those having reduplicated stems, the prefix of the second noun is omitted". Our experiment showed that for a significant number of compounds this is not the case. Another insight gained from this study is that the rules governing deverbative morphology are not adequate for a complete description of the formation of noun + noun compounds in which the second or both constituents are deverbatives. A deeper linguistic analysis, which would facilitate a more accurate computational model, forms part of future work. From a linguistic perspective the status of the constituents of compounds warrants further investigation. Can a stem be identified for each part of the compound in all cases or is only the root of the constituents of the underlying word group identifiable? In conclusion, it is hoped that this investigation will contribute towards a better understanding and a computational modelling of compound formation in Setswana.

8. Acknowledgement

This material is based on work supported by the National Research Foundation of South Africa under grant number FA2004042900039. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

List of references

ALGEO, J. 1991. Fifty years among the new words: a dictionary of neologisms, 1941-1991. Cambridge: Cambridge University Press.

BEESLEY, K.R. & KARTTUNEN, L. 2003. Finite state morphology. Stanford: CSLI.

BENCZES, R. 2006. Creative compounding in English. Amsterdam: Benjamins.

COLE, D.T. 1955. An introduction of Tswana grammar. Johannesburg: Longman.

COMBRINCK, J.G.H. 1990. Afrikaanse morfologie. Pretoria: Academica.

FINLAYSON, R. & MADIBA, M. 2002. The intellectualisation of the indigenous languages of South Africa: challenges and prospects. *Current issues in language planning*, 3(1):40-61.

KOTZÉ, P.M. & ANDERSON, W.N. 2005. A computational morphological analyser for Northern Sotho deverbative nouns: applying Xerox finite-state software to traditional grammar. *South African journal of African languages*, 25(1):59-70.

KRÜGER, C.J.H. 1994. Notes on morphology with special reference to Tswana. *South African journal of African languages*, 14(1):15-23.

KRÜGER, C.J.H. 2006. Introduction to the morphology of Setswana. München: Lincom Europa.
LAAS, J.A.M. 1974. Woordanalise in Suid-Sotho. Potchefstroom: PU vir CHO. (M.A.-verhandeling.)
LANDSBERG, G.A. 1987. Komposita in Noord-Sotho. Potchefstroom: PU vir CHO. (Ph.D.-proefskrif.)
MATTHEWS, P.H. 1997. Oxford concise dictionary of linguistics. Oxford: Oxford University Press.
POSTHUMUS, L.C. 1994. Word-based versus root-based morphology in the African languages. *South African journal of African languages,* 14(1):28-36.
PRETORIUS, L. & BOSCH, S.E. 2002. Finite state computational morphology: treatment of the Zulu noun. *South African computer journal*, 28:30-38.
PRETORIUS, R., VILJOEN, B. & PRETORIUS, L. 2005. A finite-state morphological analysis of Setswana nouns. *South African journal of African languages*, 25(1):48-58.
PRETORIUS, R.S. & BERG, A.S. 2005. The morphological analysis of Setswana nouns. *Tydskrif vir taalonderrig*, 39(2):274-291.
PRETORIUS, W.J. 2000. Die identifisering en beskrywing van die begrippe stam en wortel in die Afrikatale, met besondere verwysing na die Sothotale. *Tydskrif vir taalonderrig*, 34(1):51-62.
UNGERER, H.J. 1983. Komposita in Zulu. Potchefstroom: PU vir CHO. (Ph.D.-proefskrif.)

Key concepts:

compounds
computational morphological analysis
Setswana

Kernbegrippe:

rekenaarmatige morfologiese analise
saamgestelde woorde
Setswana