Designing a noun guesser for part of speech tagging in Northern Sotho[1]

Ulrich Heid

Institut für maschinelle Sprachverarbeitung – Computerlinguistik
Universität Stuttgart, Azenbergstrasse 12, D - 70 174 Stuttgart, Germany, and
Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
Ulrich.Heid@ims.uni-stuttgart.de

Danie J. Prinsloo*

Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
danie.prinsloo@up.ac.za

Gertrud Faaß

Institut für maschinelle Sprachverarbeitung – Computerlinguistik
Universität Stuttgart, Azenbergstrasse 12, D - 70 174 Stuttgart, Germany, and
Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
gertrud.faasz@tuks.co.za

Elsabé Taljard

Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
elsabe.taljard@up.ac.za
* Corresponding author

In this article, we describe an element of a suite of computational tools for assigning word-class tags (as a preparation for part of speech (POS) tagging) to word forms in unrestricted Northern Sotho texts. POS-tagging is a step towards a linguistic analysis of the texts, which in turn allows for advanced data extraction.

The tool component that is described, identifies (and classifies) noun forms. Several types of linguistic knowledge are used to recognize nouns that are not contained in the noun lexicon of the system. These include the relationship between singular and plural noun prefixes, knowledge about noun derivation, and data about the co-occurrence of the candidate with concords, pronouns and adjectives in a local context.

Our implementation is a symbolic, voting-based process: together, all tests determine whether a candidate is a noun; accuracy on unseen test data is around 92%.

Introduction

Context and objective

The work discussed in this article is part of a larger effort, which has as its objective, the design of a reliable tool for word-class annotation of Northern Sotho texts. Word-class annotation (henceforth: POS-tagging) is an important step in the linguistic analysis and annotation of the corpus data of a language.[2]

The POS-tagging technology that we have chosen to implement for Northern Sotho (we use among others the RF-TreeTagger by Schmid and Laws, 2008), is dependent on lexical resources and on trainable statistical tools.

The lexical resource that is needed, is a list containing word forms and their word class description. In ambiguous cases, a word form may display several POS-tags. The lexicon may for example, state that *tsebe* is a noun of class 9 (N09), and also a verb (V). The algorithms underlying the statistical tools are designed to decide which of the POS-tags of an ambiguous item are appropriate in a given local context (cf. the ambiguity of closed-class items, such as *a*, *ka* and *go*), and possibly, to which word class an unknown item may also belong in a given local context.

In Northern Sotho, the majority of word categories are closed-class items, i.e. their forms can be listed. These include all function words, as well as adjectives. On the other hand, Northern Sotho nouns, verbs and to some extent, adverbs, are open-class items. Therefore, the lexical resource of the tagging tool will never be complete, as any new text to be processed may bring up nouns, verbs and adverbs which are not yet included in the lexicon. Nouns in particular, are created through highly productive morphological derivation processes, used for the formation of diminutives, augmentatives and deverbatives.

Consequently, providing a Northern Sotho POS-tagger with a list of closed-class items is not sufficient; more linguistic knowledge should be made available. In this article, we describe a guesser designed to identify nominal forms in Northern Sotho texts. The noun guesser is conceived as a step preceding the use of the abovementioned statistical tool: it dynamically provides lexical data for the tagger, so as to make the task of the statistical tagger somewhat easier. In this, the objective is similar to that of the verb guesser, as discussed in Prinsloo, Taljard, Heid and Faaß (2008). For an analysis of verbal extension sequencing, see Anderson and Kotzé (2008).

The noun guesser that is discussed in this article makes use of three types of knowledge to identify noun forms, i.e. knowledge of (i) noun class prefixes, (ii) nominal suffixes, and (iii) the local morpho-syntactic context of candidate words forms.

Relevant properties of Northern Sotho nouns

The grammar of Northern Sotho is described in detail in sources such as Lombard (1985), Van Wyk, Groenewald, Prinsloo, Kock and Taljard (1992) and Poulos and Louwrens (1994). Northern Sotho nouns are subdivided into nominal classes signalled by prefixes, and by class-specific concords and pronouns. Cf. Table 1 in this regard.

Table 1: Noun classes, concords and pronouns for Northern Sotho

CS = subject concord, CO = object concord,
CPOSS = possessive concord, CDEM = demonstratives,
PROEMP = emphatic pronouns, PROPOSS = possessive pronouns,
PROQUANT = quantitative pronouns.

Class #	Prefix	Example	CS	CO	CPOSS	CDEM	PROEMP/PROPOSS	PRO-QUANT
1	mo-	mosadi ‘woman’	o / a	mo	wa	yo	yena	yohle
2	ba-	basadi ‘women’	ba	ba	ba	ba	bona	bohle
1a	ø/N-	tate ‘father’	o	mo	wa	yo	yena	yohle
2b	bo+	botate ‘fathers’	ba	ba	ba	ba	bona	bohle
3	mo-	motse ‘village’	o	o	wa	wo	wona	wohle
4	me-	metse ‘villages’	e	e	ya	ye	yona	yohle
5	le-	lesogana ‘young man’	le	le	la	le	lona	lohle
6	ma-	masogana ‘young men’	a	a	a	a	ona	ohle
7	se-	selepe ‘axe’	se	se	sa	se	sona	sohle
8	di-	dilepe ‘axes’	di	di	tša	tše	tšona	tšohle
9	N-/ø-	nku ‘sheep’	e	e	ya	ye	yona	yohle

10	di+N-/ di+ø-	dinku 'sheep' (plural)	di	di	tša	tše	tšona	tšohle
11	-							
12	-							
13	-							
14	bo-	bogobe 'porridge'	bo	bo	bja	bjo	bjona	bjohle
6	ma-	magobe 'different kinds of porridge'	a	a	a	a	ona	ohle
15	go	go thuša 'to help'	go	go	ga	--	gona	gohle
16	fa-	fase 'below'	go	go	ga	fa	gona	gohle
17	go-	godimo 'above'	go	go	ga	--	gona	gohle
18	mo-	morago 'behind'	go	go	ga	mo	gona	gohle
	N-/ø-	ntle 'outside'	go	go	ga	--	gona	gohle
	ga-	gare 'inside'	go	go	ga	--	gona	gohle

Northern Sotho nouns are not equally distributed across these classes however. A frequency breakdown of nouns occurring four or more times in a 44,000-word test corpus (a doctoral dissertation by R.M. Thobakgale (2006), henceforth Thobakgale Test Corpus (TTC)), clearly indicates such an unequal distribution. Cf. Figures 1 and 2 below for a breakdown of the distribution of nouns in terms of tokens and types in the TTC.

Figure 1: Breakdown of nouns in terms of tokens

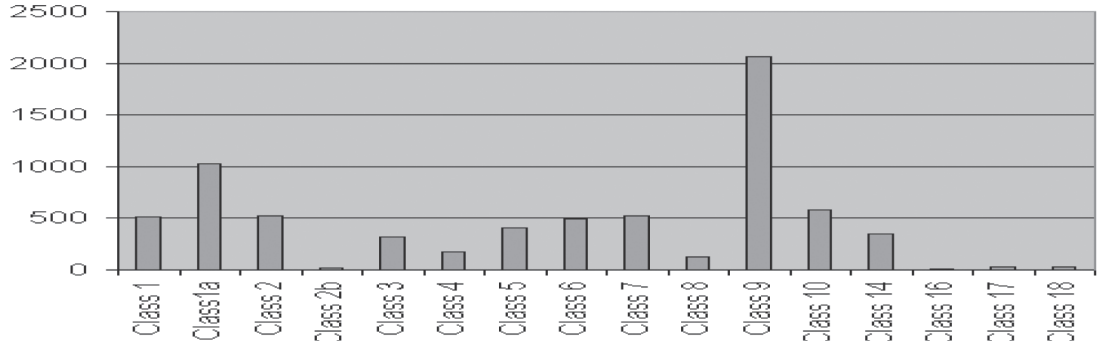
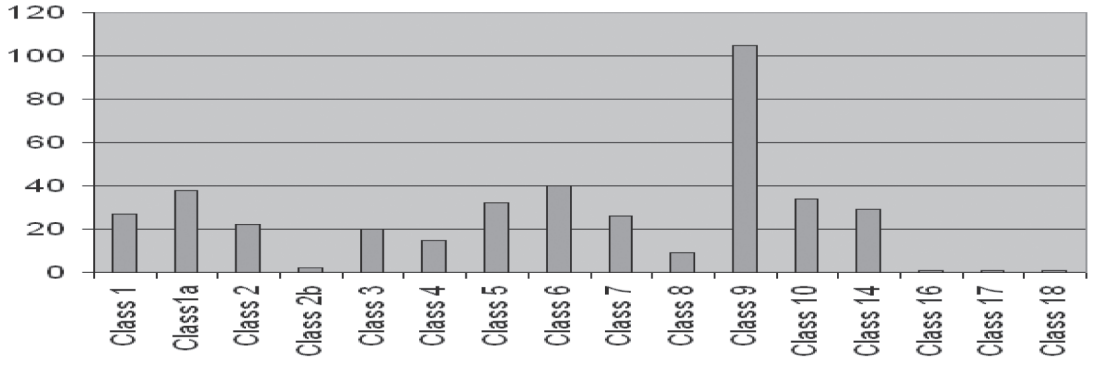


Figure 2: Breakdown of nouns in terms of types



In terms of tokens, 2068 nouns from class 9 were found in the TTC, representing 105 types, whereas only eight tokens for class 16 were found, which are eight occurrences of the same word, thus representing one type only. The noun classes 16, 17 and 18 as well as the *N*- and *ga*-classes (these five classes are termed 'locative classes') are unproductive classes. Each class contains a fixed number of nouns: class 16 has only two members; class 17 only one; class 18 has five members; the *N*-class has 11, and the *ga*-class has three. The number of locative nouns cannot be increased by means of nominalization. Therefore, it makes sense to treat the locative classes as closed classes, for the purpose of POS-tagging, i.e. by enumerating their members in the dictionary.

A tagset for Northern Sotho nominal forms

To provide a POS-tagged Northern Sotho corpus, a tagset has been designed that relies on distinctions made in traditional grammar and, at the same time, takes automatic assignability into consideration. A detailed description of this tagset can be found in Taljard, Faaß, Prinsloo and Heid (2008).

In this section, we only present the tags relevant for the annotation of nominal forms. Nouns, adjectives and concords are sub-classified according to two dimensions, i.e. (i) category and (ii) noun class. Nouns are simply tagged by a capital letter 'N', followed by the class number (#), e.g. *selepe* N07 'axe', *monna* N01 'man' and *motse* N03 'village'. Nominal concords, i.e. concords that are morphologically linked to nominal prefixes, are subject concords (CS#), object concords (CO#), possessive concords (CPOSS#), and demonstratives (CDEM#).[3] With regard to pronouns, three sub-sets are distinguished, i.e. emphatic pronouns (PROEMP#), possessive pronouns (PROPOSS#) and quantitative pronouns (PROQUANT#). Compare Table 1 in this regard.

A methodology for noun guessing

Preliminaries: the tagging process – statistical considerations

For the POS-tagging of Northern Sotho, we envisage the use of both symbolic and statistical linguistic knowledge. The POS-tagging tool chain (cf. Prinsloo & Heid, 2005) can be conceptualized as follows:

- The tagger uses a lexicon of *word form + POS-tag* pairs, the contents of which are applied statically to the text to be tagged, yielding a partially tagged text. As many forms are ambiguous, many of the tagged instances are ambiguously annotated.
- In a second step, we plan to apply the noun guesser (and a second guesser, for verbs) to reduce the number of unknown word forms, and to tag as many noun and verb forms as possible.
- A third step, which may turn out to be optional, could be a rule-based (partial) disambiguation of closed-class items.
- Finally, an instance of the RF-TreeTagger by Schmid and Laws (2008) is applied, which uses probabilistic knowledge about word/tag-associations and word/tag-sequences to disambiguate in context. Therefore, the (static) lexicons and the (dynamic) guessers are said to be symbolic pre-taggers that produce (possibly ambiguous) annotations.

The aim of this article is therefore not to give a full account of the POS-tagging process, but rather to focus on noun guessing as a supportive step towards POS-tagging. It should furthermore be noted that a well-described approach to POS-tagging in Northern Sotho was published by De Schryver and De Pauw (2007), in which they report that their (statistical) approach reaches an accuracy of 78.9% for unknown words. Their tagger, however, does not provide any information on noun class numbers. One can expect this percentage to be lower if noun class numbers are taken into account. By identifying nouns in a pre-processing step and adding them to a tagger lexicon used as input for a statistical tagger (cf. Faaß, Heid, Taljard & Prinsloo, 2009), an overall accuracy of more than 94% for POS-tagging is reached. An additional advantage of our tagger is that it provides information on noun class numbers, which can be used for further applications, such as parsing.

As of August 2008, the lexicon-based pre-tagging phase has been based on a dictionary of ca. 7300 word forms. It contains:

i. ca. 300 frequent personal names
ii. all closed-class items of Northern Sotho, with all possible tag options
iii. a list of ca. 3700 top-frequency verb forms, extracted from the University of Pretoria Sepedi Corpus[4]
iv. manually annotated list of the ca. 1000 most frequent word forms of Northern Sotho not covered by (i) to (iii).

There are good reasons to aim at a frequency-correlated coverage of any dictionary used for POS-tagging or for other NLP work on Northern Sotho, as the language is characterized by a particularly marked Zipfian distribution of words: the top 1000 word form types by frequency cover 77.5% of the lexical material of the PSC, and the top 10,000 types over 91%. A total of 5,957,553 tokens were taken into consideration in both instances. Similarly, the 57,000 word form lexicon compiled by Prinsloo for the creation of a Northern Sotho spell checker (cf. Prinsloo & Eiselen, 2005) has a lexical recall of around 98%, on arbitrarily selected Northern Sotho texts.

For lexicon-based pre-tagging, this implies that by projecting the lexical knowledge contained in a dictionary consisting of *word form + POS-tag* (alternatives), a considerable part of any text is already tagged, albeit ambiguously. In other words, once the issue of noun and verb form identification is solved, unseen Northern Sotho texts should no longer provide many unknown words, and the statistical tagger can then be used mainly for disambiguation.

Class prefixes and singular/plural relationships used as noun detectors

As mentioned above, we use different kinds of linguistic knowledge to identify text words as nouns. This knowledge includes a detailed analysis of the nominal prefix system and the singular/plural relationship, an account of nominal derivations and aspects of the local morphosyntactic context, as discussed below.

Taken at face value, it could be argued that nouns can be detected and identified by means of their class prefixes, since these prefixes, (cf. Table 1), are easily recognizable. The Northern Sotho linguist is however, familiar with the shortcomings of such a simplistic rule; these will briefly be summarized.

In the first instance, there is quite a substantial coincidental morphological overlap between the noun class prefixes and the initial syllables of other word categories. Compare for example: *sefofane* 'aeroplane' (N07, where *se-* = class prefix), as opposed to *sepela* 'walk' (verb), *sehlee* 'being greyish-yellow' (ideophone), *sekhwi* 'this particular one' (demonstrative), and *sengwe* 'another' (adjective). In this particular example, *sehlee*, *sekhwi* and *sengwe* are closed-class items; therefore the coincidental ambiguity of the string *se-* is most problematic for noun vs. verb distinction. This is a general trend.

Secondly, on another level, there is internal ambiguity within the noun class system, in that the same prefix signals different noun classes. Compare, in this regard, classes 1, 3 and 18 that all display the prefix *mo-*; classes 1a, 9 and the *N*-locative class, which also share the same prefix *N-*, and classes 8 and 10 which are both characterized by the prefix *di-*. It is thus clear that the most salient morphological feature of Northern Sotho nouns, i.e. the presence of a readily distinguishable class prefix, is inadequate on its own, for noun detection.

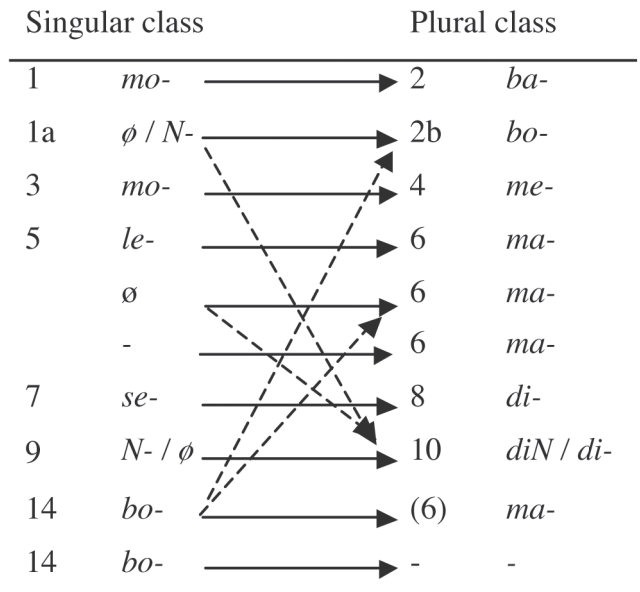
The first step towards successful noun guessing is to utilize a specific grammatical relation between classes, i.e. singularity versus plurality. From Table 1 above, it should be clear that the nouns belonging to classes 1 to 10 are arranged in pairs, with the unevenly numbered classes containing the singular forms, and their evenly numbered counterparts representing the plural forms. The obvious exception is class 14, where an evenly numbered class utilizes the class prefix of class 6 for the expression of some sense of plurality. The singular/plural pairing has direct implications for noun guessing. If it can be ascertained that a plural noun form exists in the corpus for a suspected singular noun form and vice versa, this could be a way to successfully guess nouns. For example, the word *selepe* 'axe' is a candidate for the tag N07 on the assumption that *se-* is the class prefix of class 7. Thus, if the same stem, i.e. *-lepe*, is found with the class prefix *di-* of class 8, i.e. *dilepe*, the noun candidate *selepe* can indeed be marked as N07.

The singular vs. plural relation is a detection option for nouns in classes 1 to 10, and to a lesser extent for class 14 nouns, since the singular/plural dichotomy is relevant for these classes. These relations are, however, not problem-free. In the following paragraphs problematic issues will be discussed briefly by moving through the noun class system:

- in terms of noun class prefix ambiguity
- in terms of the utilization of a variety of plural formation strategies
- with regard to morpho-phonological complexities and irregular forms.

A schematic illustration of the singular vs. plural relationship is reflected in Table 2 below.

Table 2: Singular vs. plural class pairings in Northern Sotho. (Key: solid lines indicate canonical relationships, dotted lines indicate secondary relationships.)



Prefixal ambiguity

The internal ambiguity between classes 1, 3 and 18 can be solved in two steps. As was mentioned above, class 18 contains only five members (*morago* 'behind', (*ka*) *moka* 'every', *moše* 'other side', *mošono* 'this side', *mošola* 'that side') with no possibility of adding to the nouns in this class. These are, therefore, tagged as NLOC in the system's dictionary.

As a second step, the ambiguity in terms of classes 1 and 3 can then be solved by looking for attested plural forms *mo-* : *ba-* and *mo-* : *me-*. For example, the item *monna* 'man' will be guessed as N01, since its plural counterpart *banna* (N02) occurs frequently in the corpus and no occurrences of **menna* are found. Likewise, *motse* 'village' will be correctly guessed as N03 since the plural form *metse* (N04) 'villages' occurs in the corpus, but not **batse*. The words *Modimo* (N01) 'God', and *modimo* (N01) 'ancestral spirit', or *modimo* (N03) 'a (foreign) god', however, are more problematic, since this homograph generates two plural forms *badimo* (N02) 'ancestral spirits' and *medimo* (N04) '(foreign) gods' (there is no plural form of *Modimo* (N01) 'God').

The three-way ambiguity between classes 1a, 9 and the *N*-locative class can be addressed in much the same way. First, the 11 nouns belonging to the *N*-class are entered into the lexicon, which leaves the two-way ambiguity between classes 1a and 9 to be solved. Again, class 1a nouns can only be distinguished from class 9 nouns on the basis of attested plural forms in classes 2b and 10 respectively. For example, *rangwane* 'paternal uncle' is tagged as N01a on the basis of the occurrence of *borangwane* 'paternal uncles' (N02b) and the non-occurrence of **dirangwane* in the corpus, and *tafola* 'table' as N09 on the basis of *ditafola* 'tables' (N10) versus the non-attested **botafola*. The fact that 90% of the class 9 nouns in the TTC, cf. Figures 1 and 2 above, have a ø-prefix, (and not *N-*, cf. Table 2), makes it even more challenging to successfully guess class 9 nouns because without a typical noun prefix, in principle, any word can be a class 9 candidate.

The ambiguity between classes 2b and 14 can be solved in terms of the attested singular vs. plural forms 1a vs. 6. The assumption is that a noun with prefix *bo-* could either be the plural of a class 1a noun, thus belonging to class 2b, or a singular noun, belonging to class 14. The ambiguity of the item *borangwane* 'paternal uncle', for example, is solved on the basis of the existence of *rangwane* (N01a) versus the non-occurrence of **marangwane* in the corpus. In the same way, the item *bolwetši* 'illness' would correctly be guessed as N14, based on the occurrence of *malwetši* 'different types of diseases' and the non-occurrence of **-lwetši*.

To distinguish between class 8 and class 10 nouns, which both carry the class prefix *di-*, the relation of these nouns to the singular classes 7 and 9 is utilized. This means that nouns occurring with the prefix *di-* such as *dilepe* 'axes' and *ditafola* 'tables', can correctly be tagged as N08 and N10 respectively on the basis of occurrences of *selepe* 'axe' (N07) and *tafola* 'table' (N09), and the non-occurrence of **lepe* and **setafola* in the corpus. Details of the implementation of these tests are given below.

Plural formation strategies

The singular vs. plural relationship is complicated by the use of different plural formation strategies. In some cases a substitution strategy is employed, where the singular prefix is substituted by the plural prefix, cf. *le-sogana* 'young man' > *ma-sogana* 'young men'. In other cases an additive strategy is used, thus the singular prefix is retained and the plural prefix affixed to the full form: *n-ko* (*n-* = singular prefix class 9) 'nose' > *di-n-ko* (*di-* = plural prefix class 10, *n-* = singular prefix class 9) 'noses'. Lastly, examples are found where a single form can utilize both strategies as alternatives, cf. *bo-thata* 'problem' > *ma-thata* or *ma-bo-thata* 'problems'. These facts all need to be taken into account when designing a noun guesser – it thus may have to verify several possible forms for each class.

Morphophonological phenomena

Apart from noun class ambiguity and the utilization of a variety of plural formation strategies, internal complexities with regard to plural formation come into play, due to certain morphophonological rules. Classes 1 and 2, and 3 and 4 are used in Tables 3 and 4 to illustrate the extent of these complexities. Figure 3 shows all these relationships as they present themselves for a noun guesser based on the singular/plural dichotomy.

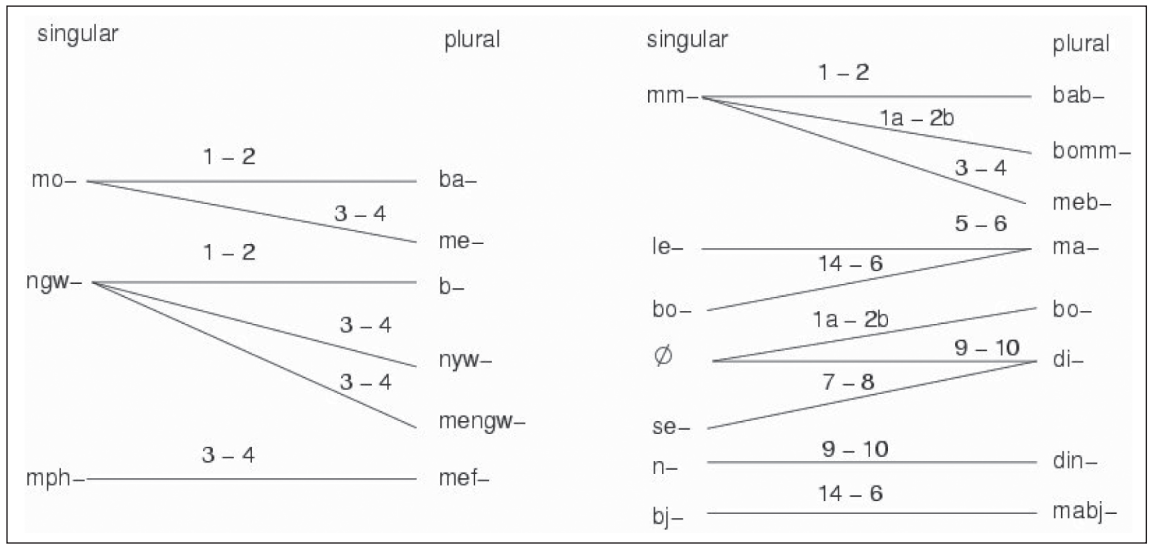
Table 3: Singular/plural relationships in classes 1 and 2

	Class 1	**Class 2**	**Plural formation strategy**
Canonical prefix	mo-	ba-	Replace mo- with ba-: mo-sadi 'woman' > ba-sadi 'women'
Variant 1	mm-	ba-b-	Replace mm- with bab-: mmadi 'reader' > babadi 'readers'
Variant 2	ngw-	b-	Replace ngw- with b-: ngwana 'child' > bana 'children'

Table 4: Singular/plural relationships in classes 3 and 4

	Class 3	**Class 4**	**Plural formation strategy**
Canonical prefix	mo-	me-	Replace mo- with me-: mo-ago 'building' > me-ago 'buildings'
Variant 1	mm-	me-b-	Replace mm- with meb-: mmele 'body' > mebele 'bodies'
Variant 2a	ngw-	nyw-	Replace ngw- with nyw-: ngwaga 'year' > nywaga 'years'
Variant 2b			Add me- to full form: ngwaga 'year' > me-ngwaga 'years'
Variant 3	mph-	me-f-	Replace mph- with mef-: mpholo 'poison' > mefolo 'poisons'

Figure 3: Relationships between singular and plural forms as marked by class prefixes and their morphophonological variants



As far as morphophonological complexities and irregular forms are concerned, the question is whether the number of nouns affected by a specific rule justifies the formulation and implementation of a dedicated rule, or whether nouns affected by morphophonological changes should simply be covered, in total, by the lexicon.

The morphophonological rule *mo+b- > mm-* serves as an example. This rule affects nouns in both classes 1 and 3 and implies that the singular forms of words such as *babuši* 'governors, rulers', *baboni* 'ones who see', *mebutla* 'hares', and *mebele* 'bodies', are not **mobuši*, **moboni*, **mobutla* and **mobele*, respectively, but rather *mmuši* (**mo-buši*), *mmoni* (**mo-boni*), *mmutla* (**mo-butla)* and *mmele* (**mo-bele*).

In this particular instance, the designer of the noun guesser is forced to formulate a general rule (instead of treatment in the lexicon), since the morphophonological rule *mo+b- > mm-* does not apply to a finite number of nouns, but is general and especially productive in the formation of deverbatives, i.e. nominalizations based on verbs in both classes. Whenever a nominalization is formed using a verb with a stem starting with the consonant *b-*, this rule will influence the morphology of the resulting deverbative. Also see Kotzé and Anderson (2005) for a detailed account of the morphological complexities of deverbatives, which they treat within a finite-state approach.

A number of words in class 5 occur without the prefix *le-*, e.g. *lapa* 'yard, household', *lokwa* 'net' while *(le)tšatši* 'day', *(le)ina* 'name', *(le)baka* 'time, reason', *(le)swiswi* 'darkness' and a few others can appear either with or without the prefix. A number of irregular forms or words that lack corresponding singular forms such as *meetse* 'water', *marega* 'winter', *maswi* 'milk' and *mare* 'spit' also exist, whereas a few nouns have two plural forms, e.g. *letšoba* 'flower' > *matšoba/maloba* 'flowers', *letsogo* 'arm' > and *matsogo/mabogo* 'arms'. Cases such as these represent a relatively small and restricted number of examples, and are thus treated in the lexicon.

Nominal suffixes (and rules for the formation of locatives) used as nominal detectors

The second fundamental component of the noun guesser focuses on the detection of the three nominal suffixes, i.e. locative ((*-e*)*ng*), augmentative/feminine (*-gadi*) and diminutive (*-ana*), and on rules for the formation of nouns that are derived from verbs. These features have not been implemented in the current version, but are discussed here with a view to future implementation. Compare the following examples by way of illustration:

(1a) Locative
dihlare 'trees' > *dihlareng* 'at/in/to/from the trees' (-(*e*)*ng* = locative suffix)

(1b) Diminutive
dinoka 'rivers' > *dinokana* 'small rivers' (*-ana* = diminutive suffix)
(1c) Augmentative/feminine
ditau 'lions' > *ditaugadi* 'lionesses' (*-gadi* = feminine suffix)

The examples given in (1) i.e. *dihlareng*, *dinokana* and *ditaugadi* will be ambiguously labelled N08:N10 as a first step, based on the presence of the assumed prefix *di-*. Matches will then be attempted for their possible corresponding forms in the singular classes 7 and 9, in terms of the singular vs. plural matching strategy described above, i.e. *sehlareng* (class 7) versus **hlareng* (class 9); **senokana* (class 7) versus *nokana* (class 9); and **setaugadi* (class 7) versus *taugadi* (class 9), respectively.

However, derived forms are often less frequent than their base forms, therefore it makes sense, in order to guess the noun status of e.g. *dihlareng*, *dinokana* and *ditaugadi*, to also check the corpus for the possible presence of their non-derived counterparts *dihlare*, *dinoka* and *ditau*. Thus the principles of the singular vs. plural relationship are applied to the derived forms, before applying the usual corpus test for plural candidates, and searching for *sehlare* and **hlare*, **senoka* and *noka*, **setau* and *tau*. The frequency distribution of base forms vs. noun derivates can be illustrated with PSC frequencies of *dihlare* (597) vs. *dihlareng* (33), or *sehlare* (491) vs. *sehlareng* (20).

Consider the following example from a 10,000 word excerpt from the TTC. The item *mererong* 'in/among the themes' and its non-locative singular *morero* 'theme' both occur, but *morerong* is not found. This means that an attempt to match the plural locative form *mererong* with its singular locative form *morerong* would not find any matches; but with the additional heuristic i.e. to also look for corresponding non-locative plural and singular forms, a reliable guesser result is obtained. The same procedure is followed with regard to the diminutive and augmentative suffixes.

As far as detecting deverbatives is concerned, frequency lists culled from the Northern Sotho corpora show that a significant percentage of nouns in Northern Sotho are derived from verbs. Consider the top five nominal derivations (frequency counts given in brackets) from the PSC in (2).

(2) *bophelo* 'life' (3 436) < *phela* '(to) live'
polelo 'language' (3 117) < *bolela* '(to)speak'
thuto 'lesson' (3 009) < *ruta* '(to) learn'
bohlokwa 'importance' (2 309) < *hloka* '(to) lack'
lerato 'love' (2 127) < *rata* '(to) love'

In principle, such nominalizations entail the prefixing of a nominal prefix to the verb, changing the verbal ending *-a* to *-o* or *-i* in most cases. Thus, a candidate noun should be tested against its nominal form minus its class prefix and with a replacement of its final vowel by the typical verbal ending *-a*, i.e. against the verbal stem used as a basis for deverbative formation. Compare the examples in (3):

(3) *moraloki* / *baraloki* 'player/s': search for *raloka* 'play'
moago / *meago* 'building/s': search for *aga* 'build'
lesego / *masego* 'laughter': search for *sega* 'laugh'
sediko / *didiko* 'circle/s': search for *dika* 'encircle'

In the case of classes 9 and 10, the (underlying) nasal *N-* in the class prefix causes phonological changes with regard to the initial consonant when deverbatives are formed in these classes. Compare the examples in (4) in this regard:

(4)

polelo 'language'	< *bolela* 'speak'	Rule: *N-* + *b-* > *p-*
thato 'will'	< *rata* 'love'	Rule: *N-* + *r-* > *th-*
temo 'field'	< *lema* 'plough'	Rule: *N-* + *l-* > *t-*
tšhilo 'grinding stone'	< *šila* 'grind'	Rule: *N-* + *š-* > *tšh-*, etc.

The respective morphonological rules therefore had to be specified in the noun guesser to enable it to detect the verbal stems from which deverbative nouns had been formed. For a full inventory of these changes, cf. Kotzé and Anderson (2005:66).

According to the principles set out above, the possibility of detecting deverbative nouns, obviously increases the recall of the noun guesser. In addition, the module for identifying deverbative nouns may use two sources of knowledge about verbs, i.e. either the corpus or the verbal entries of the system lexicon. As a by-product, it makes sense to collect the deverbative nouns and to include them in the lexicon of the system, possibly together with a feature reflecting their deverbative nature.

Syntactic environment: nominal concords and pronouns as noun detectors

The third fundamental component of the noun guesser rests on the assumption that a noun belonging to a specific noun class will, in most cases, have concords, pronouns and adjectives of the same class preceding and/or following it. Therefore, nouns should be studied in context, specifically in terms of the preceding and following words, as illustrated in Table 5 for *mareo* 'terms' and *mahlo* 'eyes', both nouns belonging to class 6.

Table 5: A selection of concordance lines for *mahlo* 'eyes' and *mareo* 'terms' from TTC

1	oletše ka ga mekgwa ye mene ya go bopa	**mareo**	ao gore Bapedi ba tle ba kgone go bolela
2	iponetše ka a ka. Ge a no go lebelela ka	**mahlo**	ale, O ka re o re: "Ke a go rata," ke ka fa
3	gamasome a mabedi maatleng a boraro;	**mahlo**	a ka ke a magolo ebile ke a mašweu go
4	o dipolelo tša go fapafapana di hlalošago	**mareo**	a mangwe a go fapana a go ithuta, go ak
5	(1972). Di ile tša ngwalwa ka pukung ya	**mareo**	le Mongwalo no. 4 (1988), matl. 71-72. (
6	antši dipuku tše tša Sepedi ke mananeo a	**mareo**	fela le diphetolo goba ditlhalošo tša
7	lwa ge a enwa. A rile melomo ye sehlee!	**mahlo**	a re hwibii! A laetša gore o lapile kudu, a.
8	polelo tše di fapafapanego di tšweletšago	**mareo**	makaleng a mangwe a thuto le go a diriša
9	o dikgopolo tše mpsha di fiwago maina/	**mareo**	ka gona. (6) 2. Gantši kgopolo ye e itšeg
10	- boloka, bea, fiša tšhelete. Moo re na le	**mareo**	a mararo a a lebanego le kgopolo ye e it
11	jalo. Mokgwa wo mongwe ke go šomiša	**mareo**	a a lego gona. Tlhalošo ya ona e a katolo
12	Thoba O ile a 149 šala moo a rapaletše,	**mahlo**	a tšwaletšwe, madi e le moetšana go tšwa
13	hloma bjang a makala a ba a se kgolwe	**mahlo**	a gagwe, se be se le ngatha. Ngatha ya k

Viewed from a grammatical angle, modifiers of the noun, which always appear in some concordial relation to the noun, can occur before or after the noun. Thus, in theory, analysis of the cotexts to the left of the presumed noun should be as informative as the cotexts to the right. Corpus analysis, however, clearly reveals that this is not the case in practice. Modifiers occur most frequently to the right of the noun. Concordial elements, such as subject concords, also appear to the right of the noun in most cases. Subject concords are part of the verb and verbs usually follow subject nouns, according to the dominant SV(O) order of the language.

In the context-based noun identification approach suggested here, an important issue is to determine an adequate size of the local co-text to be analysed. If the number of orthographical words preceding and/or following the noun candidate is too small, valuable evidence, which could support a noun guess, would be lost; if it is too big, the output will be noisy due to items incorrectly interpreted as supporting a given noun guess.

In the KeyWord in Context (KWIC) lines generated by WordSmith Tools (Scott, 1999) for *mareo* and *mahlo*, we started with a setting of 60 characters to the left and to the right of the keyword (Table 5 shows a smaller window). In line 10, supporting evidence for *mareo* as a class 6 noun is stacked in the positions KeyWord plus one and plus three (*a* = demonstrative class 6), in KW plus two (*mararo* = adjective class 6) and KW plus four (*a* = subject concord class 6). In line three, the most useful contextual evidence for *mahlo* being a noun in class 6 is found in the positions KW plus one, plus four and plus eight (*a* = demonstrative class 6), KW plus five (*magolo* = adjective class 6) and KW plus nine (*mašweu* = adjective class 6).

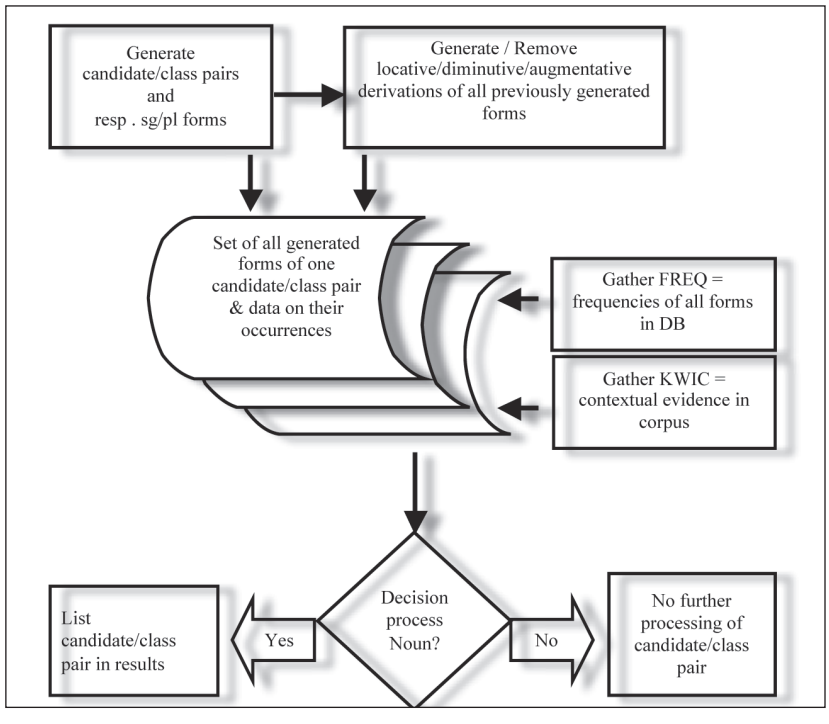
However, WordSmith's default setting of 60 characters turns out to be not ideal: morphosyntactic evidence for noun guessing gathered outside sentence boundaries are less useful. Furthermore, as was pointed out above, contexts to the right of the keyword provide support that is more reliable than the material appearing to the left: in line seven, the positions KW minus five and KW minus seven (i.e. 5 and 7 words to the left of the candidate *mahlo*) provide false evidence, since *a* is a subject concord of class 1 in both these positions. In fact, no supportive evidence for the nominal status of any of the candidates in Table 6 can be deduced from the co-text appearing to the left of the keywords.

As concordial morphemes are homographous between different noun classes, these morphemes cannot serve as unambiguous indicators in all cases. To counterbalance this effect, a particular sequencing of analysis steps has been adopted in our implementation of the noun guesser that will be discussed in the following section.

Implementation

The noun guessing software has been developed using the programming language Perl[5] on a standard PC with Linux as the operating system. Following the strategies described above, and as shown in Figure 4, the noun guesser is implemented by means of a modular structure. In the next step of the project, second level annotation (e.g. derivational forms) will be added to the guessed nouns, as detected information on such forms is only utilized during current analysis, but not stored for labelling. The guessing procedure will be illustrated by using the form *molomo* 'lip, mouth', a class 3 noun.

Figure 4: Flowchart of the noun guesser



There are three modules supporting the noun guessing decision process: class detection, i.e. the creation of a set of candidate/class pairs, generation of their appropriate singular and plural forms, and the generation of a number of typical derivations of both. For each noun candidate/class pair, a set of forms is hence generated.

The noun *molomo*, for example, is firstly identified as either a class 1 or a class 3 noun, based on its ambiguous prefix *mo-*. As both classes contain singular forms, the appropriate plural forms of classes 2 and 4 are generated and added to the candidate/class pairs, as in Table 6.

Table 6: The candidate/class pairs for *molomo* 'lip, mouth' with the generated plural forms

Candidate/Class pairs	*molomo*/1	*molomo*/3
Plural forms generated	*balomo*/2	*melomo*/4

Secondly, typical derived forms are generated by the second module and added to the set, as in Table 7. We opted to implement the locative first, for it is to be the most frequent derivation. An implementation of the other possible derivations, i.e. augmentative and diminutive derivations may follow at a later stage.

Table 7: The extended set, containing derived forms

Candidate/Class pairs	*molomo*/1	*molomo*/3
Locative derivation	*molomong*/1	*molomong*/3
Assumed plural forms	*balomo*/2	*melomo*/4
Locative derivation of plural forms	*balomong*/2	*melomong*/4

For all of the forms of each set, two features are examined: evidence in the form of the number of occurrences (FREQ), and contextual evidence, as evidenced by a KWIC search of the PSC. The data is collected and fed into a decision process that can be seen as a voting process: text fulfilment leads to points for a candidate pair. Points are added for any evidence that is found. However, if no evidence is found, points may also be subtracted.

Other tests might also add significant evidence (expressed in points for the candidate) towards a more general positive decision. For example, the status of a candidate, as a noun, is also verified by a module that will check if the candidate has the suffixes *-i* or *-o*. If the result is positive, this module will execute further tests to decide if the candidate is a deverbative. In the case of *molomo*, the hypothesis will be made that the noun might be derived from the verb *loma* 'bite'. As this verb is found in the system's lexicon, two additional points will be allocated to both the candidate/class pairs.

The extended set, reproduced in Table 7, is supplied to the second group of modules of which one retrieves data (FREQ) from a database containing the frequencies of all tokens contained in the PSC, and the second (KWIC) queries contextual data in the six million token PSC itself. This second module checks each form of the candidate/class pairs to ascertain whether it occurs in a predefined class dependent local context. To check the corpus for contexts suggesting that the candidate is a noun of a certain class, the module searches for the respective concords, pronouns or adjectives (= triggers) that may surround such a candidate. For example, the triggers for class 1 (and 1a), of which *molomo* is possibly a member, is the following set:

{*yo/yokhwi/yola/yono/yoo/youwe/yowe/mo/wa/o/a/yena/gagwe/yohle/mobe/mobjang/mobotse/mofsa/mogolo/mohubedu/mohwibidu/mokaaka/mokae/mokopana/mokoto/mongwe/monyane/mosehla/mošele/mosese/moso/mosootho/mošoro/moswa/mošweu/motala/motelele/mothata/motsothwa*}.

Evidence for a candidate/class pair will only be valid if at least one member of the set that is defined for the respective class is found either in the left (up to Keyword-2) or in the right (up to Keyword+3) context of the candidate. The second frequency value that we note is the number of contexts identified with such a query.

The candidate *molomo* is also identified as possibly belonging to class 3; hence context data is examined again, this time with the class 3 hypothesis. To collect evidence for a class 3 membership, the following set of triggers of class 3 is searched for:

{*wo/wokhi/wola/wono/woo/wouwe/wowe/o/wa/wona/wohle/mobe/mobjang/mobotse/mofsa/mogolo/mohubedu/mohwibidu/mokaaka/mokae/mokopana/mokoto/mongwe/monyane/mosehla/mošele/mosese/moso/mosootho/mošoro/moswa/mošweu/motala/motelele/mothata/motsothwa*}.

Note that there is an overlap between the trigger sets for the two classes. Such overlaps are common across indicator sets for either singular or plural classes, but not across the number dimension. However, as the aim of the guesser is to firstly determine the status of the candidate as a noun, and then, secondly, to identify all possible classes the noun might belong to, a rather generous selection strategy can be implemented. The results of the guesser are examined by a language expert who selects from several suggestions to avoid missing a possible identification. In other words, the implementation strategy aims at recall, assuming that false positives will be removed manually.

The resulting values (FREQ and KWIC) are added to the set of candidate pairs in order to supply the decision process with the necessary data. Table 8 shows the contents of the set after this addition:

Table 8: Word frequencies and KWIC frequencies of all forms generated for the candidate *molomo*

Candidate	*molomo*/1/1560/1029	*molomo*/3/1560/691
Assumed plural forms	*balomo*/2/0/0	*melomo*/4/567/306
Assumed locative forms 1	*molomong*/1/325/292	*molomong*/3/325/229
Assumed locative forms 2	*balomong*/2/0/0	*melomong*/4/106/76
Summary	*molomo*/1/1885/1321	*molomo*/3/2558/1302

After the data has been collected, the decision process examines the data and allocates points to each of the candidate pairs according to the following conditions:

- one point is added if FREQ > 0
- one point is subtracted if FREQ = 0

At this stage, collected points are compared. Concerning the candidate *molomo*, the candidate/class pair molomo/1 has zero points (+1-1+1-1) and is excluded from any further processing as not enough evidence was found to warrant further tests. The candidate/class pair *molomo*/3 on the other hand, was allocated four points (+1+1+1+1) and hence this candidate will be further processed.

Next, the relation between the two values, FREQ and KWIC, will be examined. The more contextual evidence that is available for a given noun class guess in relation to the overall amount of evidence for the form under analysis, the more clearly the evidence speaks in favour of the noun class guess. A candidate with a relation 4:1 will have one point, whereas a candidate with a relation 3:1 will have two points added. Candidates for which the relation FREQ/KWIC is less than 3:1 will be classified as not guessed, as not enough contextual evidence was found.

The candidate/class pair *molomo*/3 (2558/1302=1.96) easily fulfils the condition FREQ/KWIC = 3:1. The guesser therefore correctly reports *molomo* as belonging to noun class 3.

In other, less clear-cut examples, still no decision might be possible at this point. An additional test will then be run to check if the forms generated by the guesser are found in the system's lexicon. If this is the case, further points will be added and the comparison between the candidate/class pairs will be repeated.

Results and evaluation

Sample results

To evaluate the noun guesser, we randomly selected a sub-corpus of 10,000 words from the TTC. This corpus is not part of the six million word PSC. Eighty six point nine per cent of the types (65% of all tokens) of this test set are covered by the lexical resources of the POS-tagger. The remaining 604 types were first submitted to our verb guesser (cf. Prinsloo et al., 2008). Candidates that were possibly considered to be ambiguous between verbs and nouns were then analysed by the noun guesser. Consider *mareo* 'terms' in Table 9 as a typical example of a noun frequently used in Northern Sotho in both singular and plural forms. It is shown here with its form and noun class hypotheses and cumulative as well as contextual frequencies.

Table 9: Frequencies of *mareo* forms and their possible verbal base forms

Candidate	*mareo*/6/205/159	
Possible verb base form	*rea*/0	*rea*/0
Possible singular form 1	*lereo*/5/66/63	*boreo*/14/0/0
Summary	*mareo*/6/271/222	*mareo*/6/205/159

The test candidate *mareo* was first submitted to the verb guesser, which rendered no positive results. The word was then submitted to the noun guesser, and a first guess on the basis of the class prefix was made, i.e. class 6. This was followed by a search for matching singulars, which could be, as described above, from either class 5 or class 14, thus *lereo* or **boreo*, focusing firstly on occurrence frequency in the PSC. The fact that *lereo* occurs 66 times with zero occurrences for *boreo*, is conclusive evidence, based on frequency considerations, that *mareo* is indeed a noun in class 6, and thus a member of the singular/plural class pair 5/6.

In the case of *mareo*, the guesser could at least start from a single assumption, i.e. that it is a noun in class 6, the primary objective being to determine the nominal status of the word. For the task of guessing the noun status of an item, it is obviously not relevant if a class 6 noun is plural to class 5 or to class 14. The information about the exact noun class pair is however used for semi-automatic lexicon enlargement after a manual validation by a language expert.

In the case of *diretotumišo* 'praise poems' in Table 10, for example, an ambiguity regarding the possible class has to be assumed right from the start, since nouns displaying the class prefix *di-* can belong to either class 8 or 10.

Table 10: Frequencies of *diretotumišo* forms and their possible verbal base forms

Candidate	*diretotumišo*/8/154/101	*diretotumišo*/10/154/101
Possible verb base form	*retotumiša*/0	*retotumiša*/0
Possible singular form	*seretotumišo*/7/142/79	*retotumišo*/9/0/0
Summary	*diretotumišo*/8/296/180	--

As was the case with *mareo*, no positive result was obtained from the verb guessing procedure for *diretotumišo*. For this particular example, classes 8 and 10 were both identified as possible classes; however, class 10 cannot be considered, since the assumed singular class 9 form **retotumišo* rendered zero counts on both cumulative frequency and contextual frequency. In the end, both the total counts for the pair 7/8 were high (candidate and assumed singular form = 296), and the total for the KWIC frequency even exceeds 50% of the total frequency count (180/269 = 0.608, > 0.33), i.e. the necessary 3:1 relation is fulfilled.

Evaluation

In this section, we will describe one of the procedures used to evaluate the noun guesser. For a start, a second set of approximately 10,000 tokens taken from the TTC was submitted to the guessing process. The noun guesser was set to ignore the lexical resources of the tagger in order to also guess tokens for which the tags are known. Of the 5968 tokens (1575 types) that were guessed (tokens consisting of less than three letters were ignored), 1202 tokens (291 types) were nouns. However, six of these nouns begin with an irregular prefix; these items cannot be guessed by the proposed system as it identifies the possible classes solely by the prefix of the candidate.

Table 11 shows the result of the guessing process in detail. For each class, the second column ('Freq') indicates the number of nouns (correct manual assignment): there were, for example, 18 nouns of class 1 (line 1 of Table 11). In the third column, the guessed classes are listed; the fourth and fifth columns show the number of correct and incorrect guesses respectively, followed by the sixth column ('No guess'), informing us of the failure of the guessing process and the reasons for the failed guesses (column 7).

Some candidates were guessed correctly, but were also identified as belonging to another class, the class 5 noun *ganong* being a case in point. The candidate was correctly guessed as belonging inter alia to class 5, but also ambiguously, and incorrectly, classified as possibly belonging to class 9. We classify such a result as a false positive. Other nouns do indeed belong to two classes, for example, *lefehlo* 'wooden spoon', which also appears as a name, in which case it belongs to class N01a. We count such nouns as two nouns for the sake of simplifying the resulting statistics. In the case of *lefehlo*, due to its unambiguous class 5 prefix, the guesser cannot identify it as N01a (cf. Figure 3). There is also a high number of nouns categorized in the system's lexicon as 'N08 and N10', as it could not be decided clearly which of the classes are in use; the guesser assists in the disambiguation of these nouns (see below).

For the evaluation, we count all guesses that contain at least one of the two classes as a correct decision. The column headed 'Reason' in Table 11, informs us briefly of the reasons for the failed guesses. Its contents will be explained in more detail in the following paragraphs.

The numbers shown in Table 11 in parenthesis, indicate how many candidates begin with an irregular class prefix. These are ignored when calculating the success rate of the guesser, as its design does not allow for an identification of such items.

Table 11: The results of the guessing process

Class	Freq.	Guessed class	Correct guess FREQ	Wrong guess FREQ	No guess	Reason
N01	18	N01	18			
N01a	3 (+1)	N01a	3			
		N06		(1)		*mang* 'who': irregular prefix
N01a: N05	1 (+1)	N05	1		(1)	*lefehlo* 'wooden spoon': correctly identified as N5, not identified as class 1a (irregular prefix)
N02	9	N02	9			

Class	Freq.	Guessed class	Correct guess FREQ	Wrong guess FREQ	No guess	Reason
N03	27	N03	23			
		N01		3		*mmele* 'body', *moko* 'marrow', *moya* 'air': not enough evidence in KWIC context
					1	*molaetša* 'message': ending *-etša* interpreted as unambiguous verbal suffix, candidate excluded from noun guessing
N04	13	N04	13			
N05	36(+1)	N05	35			
		N02		(1)		*baka* 'reason': irregular prefix
		N05:N09	0.5			*ganong* 'mouth': false positive, as it is also (wrongly) identified as N09
N06	44(+3)	N06	41			
		N4		(2)		*meetse* 'water', *meno* 'teeth': irregular prefix
					(1)	*meetseng* 'in the water': irregular prefix
					3	*maabane* 'yesterday', *mabapi* 'with regard to', *mathomothomo* 'very beginning': not enough evidence in KWIC context
N07	16	N07	16			
					1	*sebjalebjale* 'modern times': not enough evidence in KWIC context
N08	4	N08	3			
		N10		1		*dibe* 'sins': more evidence for class 10 than for class 8
N08: N10	14	N08:N10	1			*dihlong* 'shame': significant evidence found for both candidate/class pairs
		N08	8			*diatla* 'hands', *dibjana* 'dishes', *dijo* 'food', *dillo* 'crying', *dilo* 'things', *direto* 'poems', *ditšhaba* 'nations', *ditseka* 'pieces': not enough evidence for N10
		N10	5			*dikgomo* 'cattle', *dingaka* 'doctors', *dintho* 'wounds', *diphiri* 'hyenas', *ditaba* 'news': not enough evidence for N08
N09	68		55			
		N05:N09	0.5			*thušano* 'help': false positive, as it is also (wrongly) identified as N05
		N01a		8		*kgoga* 'porridge made of pumpkin, dragonfly', *kgonthe* 'truth', *kwana* 'lamb', *nnete* 'truth', *nnoši* 'alone', *peapeanong* 'arrangement', *tsebe* 'ear', *tšwelopele* 'progess': not enough evidence in KWIC context
		N05		4		*khutšo* 'peace', *kwane* 'lamb', *thari* 'skin', *thipa* 'knife': not enough evidence in KWIC context
N10	9	N10	9			
N14	24	N14	15			
		N02b		1		*boomo* 'wilfulness': not enough evidence in KWIC context
		N02b:N14	7			*bobedi* 'second', *bošego* 'night', *bogologolo* 'in the past', *bohlokwa* 'precious, scarce', *bolele* 'slime, seaweeds, algae', *boleta* 'soft', *botse* 'beautiful': false positives
					1	*bodutu* 'loneliness': not enough evidence in KWIC context
sums	286 (+6)		263	17(+4)	6(+2)	
%	100		92	6	2	

Disambiguation of nouns classified as ambiguous so far

The dictionary contains a number of nouns ambiguously classified as N08:N10. One such example is *diatla* 'hands'. The guesser detected clear evidence for class 8, not only based on contextual evidence, but also based on the detection of the singular class 7 form *seatla,* which occurs significantly more frequently in the PSC than the assumed singular class 9 form *atla,* cf. Table 12. However, *atla* (316) is coincidentally a verb meaning '[to] kiss', which explains the low count concerning contextual evidence (33).

Table 12: Frequencies of *diatla* forms

Candidate	*diatla*/8/1624/533	*diatla*/10/1624/533
Possible singular form 1	*seatla*/7/1106/459	*atla*/9/316/33
Summary	*diatla*/8/2730/992	*diatla*/10/1940/566

Reasons for no guesses

Additionally to nouns that have irregular noun prefixes, there are other nouns as well that cannot be successfully identified with the described methods. Such nouns do not form a plural (or a singular), cf. *bodutu* 'loneliness'. The noun guesser did not find any evidence for such forms and therefore did not provide results (cf. Table 13).

Table 13: Frequencies of *bodutu* forms

Candidate	*bodutu*/14/509/71	*bodutu*/2b/509/68
Possible singular form 1	*madutu*/6/0/0	*dutu*/1a/0/0
Summary	*bodutu*/14/509/71	*bodutu*/2b/509/68

Coincidental morphological similarities between verbal suffixes and nominal endings

The N03 noun *molaetša* 'message' belongs to a small group of nouns ending in a string which is homographous with the verbal suffix *-etša*. Tokens with this suffix and with other typically verbal suffixes are excluded from the noun guessing to save processing time. As tokens with irregular affixes are usually listed in the lexicon, we do expect those cases to occur infrequently.

Insufficient evidence in KWIC context: data sparseness problem

There are 21 tokens for which there is insufficient evidence for the guesser to suggest any class, their correct class or all the classes that they belonged to: *bodutu* 'loneliness' (N14), *boomo* 'purpose' (N14), *kgoga* 'porridge made of pumpkin, dragonfly' (N09), *kgonthe* 'truth' (N09), *khutšo* 'rest, peace' (N09), *kwana* (N09) and *kwane* (N09), both meaning 'lamb' (N09), *maabane* 'yesterday' (N06), *mabapi* 'with regard to' (N06), *mathomothomo* 'very beginning' (N06), *mmele* 'body' (N03), *moko* 'marrow' (N03), *moya* 'air' (N03), *nnete* 'truth' (N09), *nnoši* 'alone' (N09), *peapeanong* 'in the arrangement/agreement' (N09), *sebjalebjale* 'modern' (N07), *thari* 'skin for carrying infants' (N09), *thipa* 'knife' (N09), *tsebe* 'ear' (N09), *tšwelopele* 'prosperity' (N09). The problem of data sparseness should become less severe in future, as efforts are under way to extend the Northern Sotho corpora.

False positives

Nine tokens *bobedi* (N14) 'second', *bošego* (N14) 'night', *bogologolo* (N14) 'past', *bohlokwa* (N14) 'precious, scarce', *bolele* (N14) 'slimes, seaweeds, algae', *boleta* (N14) 'soft', *botse* (N14) 'beautiful', all guessed as N02b:N14; *ganong* (N05) 'mouth', guessed as N05:N09; and *thušano* (N09) 'help', guessed as N05:N09, were

ambiguously guessed as belonging to other classes, as well as to the correctly guessed ones. At this stage, false positives are accepted, since it is important that all possible class guesses are presented to the prospective user, who will have to make a decision as to the correct class membership.

Conclusion

We have implemented a noun guesser for Northern Sotho. Our tests show that it can identify noun candidates with their noun class numbers with around 92% accuracy. This result is acceptable for a tool designed to prefer recall over precision; nevertheless, we consider the noun guesser to be a pre-processing tool in a semi-automatic chain where human intervention is needed to produce a completely correct output.

The noun guesser combines syntagmatic, i.e. context-based tests with paradigmatic ones, relying on relationships between singulars and plurals, or between morphologically related words. Thus we make use of several kinds of linguistic knowledge, combined in our voting-based decision algorithm. In fact, one objective of this work is also to assess to what extent one can utilize linguistic knowledge. This constellation may give rise to machine learning-based experiments in the future; for the time being, we rely on a symbolic approach and on manual cross-checking, to maximize the accuracy of our corpus annotation.

The noun guesser is part of a larger tool suite for the POS-tagging of Northern Sotho; it is similar in its objectives, though not in its technical realization, to our verb guesser (Prinsloo et al., 2008). Both cater for unknown open-class word forms, i.e. those items that tend to dramatically reduce the accuracy of the automatic part of speech taggers. In this sense, the noun guesser and the verb guesser can also be seen as dynamic lexical analysis (and acquisition) components to be combined with static lexical resources for tagging. They can also be used as lexical acquisition tools.

Future work will include more detailed evaluation, and the implementation of minor additions, such as the analysis of diminutives and augmentatives. Furthermore, we will run the complete tools suite on unrestricted unseen text, to assess the impact of the guesser on the overall tagging accuracy achievable with our tools on Northern Sotho texts.

Notes

1. Financial support from the National Research Foundation (NRF) for this project is gratefully acknowledged.

2. As described in Taljard et al. (2008), we tag on the orthographical level, therefore the term ‘word class’ (alternatively ‘parts of speech’) should be interpreted in its widest sense, referring to orthographic units, rather than linguistic words. In some cases, orthographic units are also linguistic words, but due to the disjunctive writing system of Northern Sotho, morphemes often appear as orthographic units, and are therefore tagged.

3. Based on the concordial relationship between nouns and demonstratives, the latter are categorized as concords. This decision might be revised in future.

4. The University of *Pretoria Sepedi Corpus* (PSC) is a collection of ca. six million running words of Northern Sotho, containing texts from different genres and domains. More than 300 texts (literary works, e.g. prose, poetry, novels) were scanned and a number of texts were electronically collected (mostly governmental publications) from the Internet at the University of Pretoria (cf. De Schryver & Prinsloo, 2000).

5. See http://www.perl.com/

References

Anderson, W.N. & Kotzé, A.E. 2008. Verbal extension sequencing: An examination from a computational perspective. *Literator* 29(1):43–64.

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

De Schryver, G-M. & Prinsloo, D.J. 2000. The compilation of electronic corpora, with special reference to the African languages. *Southern African Linguistics and Applied Language Studies* 18(1–4):89–106.

Faaß, G., Heid, U., Taljard, E. & Prinsloo, D.J. 2009. Part-of-Speech tagging of Northern Sotho: Disambiguating polysemous function words. Paper read at *EACL 2009 Workshop on Language Technologies for the African Languages*. Athens: Association for Computational Linguistics.

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.

Lombard, D.P. 1985. *Introduction to the grammar of Northern Sotho*. Pretoria: J.L. van Schaik.

Poulos, G. & Louwrens, L.J. 1994. *A linguistic analysis of Northern Sotho*. Pretoria: Via Afrika.

Prinsloo, D.J. & Eiselen, Roald. 2005. Improving a lexicon-based spelling checker for Sesotho sa Leboa. *South African Journal of African Languages* 25(1):11–24.

Prinsloo, D.J. & Heid, U. 2005. Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. Paper delivered at the Conference for Lesser Used Languages and Computer Linguistics, EURAC research, European Academy. Bolzano, Italy. 27–28 October 2005.

Prinsloo, D.J., Taljard, E., Heid, U. & Faaß, G. 2008. Designing a verb guesser for part of speech tagging in Northern Sotho. *Southern African Linguistics and Applied Language Studies* 26(2):185–196.

Schmid, H. & Laws, F. 2008. Estimation of conditional probabilities with decision trees and an application to Fine-Grained POS Tagging. COLING 2008. Manchester, Great Britain.

Scott, M. 1999. *WordSmith Tools version 3*. Oxford: Oxford University Press.

Taljard, E., Faaß, G., Prinsloo, D.J. & Heid, U. 2008. Designing a tagset for Part-of-Speech tagging of Northern Sotho corpora. *Literator* 29(1):111–137.

Thobakgale, R.M. 2006. Khuetso ya O.K. Matsepe go bangwadi ba Sepedi. Unpublished doctoral thesis. Pretoria: University of Pretoria.

Van Wyk, E.B., Groenewald, P.S., Prinsloo, D.J., Kock, J.H.M. & Taljard, E. 1992. *Northern Sotho for First-Years*. Pretoria: J.L. van Schaik.