http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

Applied Corpus Linguistics for Lexicography: Sepedi Negation as a Case in Point

Gertrud Faaß, *Department of Information Science and Natural Language Processing, University of Hildesheim, Germany (gertrud.faass@uni-hildesheim.de)*

Abstract: So far, Sepedi negations have been considered more from the point of view of lexicographical treatment. Theoretical works on Sepedi have been used for this purpose, setting as an objective a neat description of these negations in a (paper) dictionary. This paper is from a different perspective: instead of theoretical works, corpus linguistic methods are used: (1) a Sepedi corpus is examined on the basis of existing descriptions of the occurrences of a relevant verb, looking at its negated forms from a purely prescriptive point of view; (2) a "corpus-driven" strategy is employed, looking only for sequences of negation particles (or morphemes) in order to list occurring constructions, without taking into account the verbs occurring in them, apart from their endings. The approach in (2) is only intended to show a possible methodology to extend existing theories on occurring negations. We would also like to try to help lexicographers to establish a frequency-based order of entries of possible negation forms in their dictionaries by showing them the number of respective occurrences. As with all corpus linguistic work, however, we must regard corpus evidence not as representative, but as tendencies of language use that can be detected and described. This is especially true for Sepedi, for which only few and small corpora exist. This paper also describes the resources and tools used to create the necessary corpus and also how it was annotated with part of speech and lemmas. Exploring the quality of available Sepedi part-of-speech taggers concerning verbs, negation morphemes and subject concords may be a positive side result.

Keywords: AFRICAN LANGUAGES DICTIONARIES, CORPUS LINGUISTICS, NEGATION, SEPEDI, NORTHERN SOTHO, LEXICOGRAPHY, PART-OF-SPEECH TAGGING, CORPUS QUERY PROCESSING

Zusammenfassung: Eine korpuslinguistische Untersuchung der Sepedi-Negation für die Lexikographie. Bisher wurden Sepedi Negationen eher aus der Sicht der lexikographischen Behandlung betrachtet. Hierfür wurden theoretische Werke über Sepedi verwendet, wobei als Zielsetzung eine saubere Beschreibung dieser Negationen in einem (Papier-)Wörterbuch gesetzt wurde. Dieser Beitrag ist aus einer anderen Perspektive: statt theoretischer Werke werden korpuslinguistische Methoden eingesetzt: (1) ein Sepedi Korpus wird auf Basis bestehender Beschreibungen zu den Vorkommen eines einschlägigen Verbs untersucht und dabei seine negierten Formen aus rein präskriptiver Sicht betrachtet; (2) wird eine "corpus-driven"-Strategie eingesetzt, bei dem nur nach Sequenzen von Negationspartikeln (oder Morphemen) gesucht wird, um vorkommende Konstruktionen auflisten zu können, ohne dabei die dabei vorkommenden Verben —

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

abgesehen von ihrer Endung — zu berücksichtigen. Der Ansatz in (2) soll dabei nur eine mögliche Methodik aufzeigen, um bestehende Theorien über vorkommende Negationen erweitern zu können. Wir möchten auch versuchen, Lexikographen darin zu unterstützen, eine frequenzbasierte Reihenfolge der Einträge möglicher Negationsformen in ihren Wörterbüchern aufzustellen, in dem wir ihnen die Anzahl der jeweiligen Okkurrenzen aufzeigen. Wie bei allen korpuslinguistischen Arbeiten müssen wir jedoch Korpusevidenz nicht als repräsentativ ansehen, sondern als Tendenzen des Sprachgebrauchs, die festgestellt und beschrieben werden können. Dies gilt insbesondere für Sepedi, für das nur wenige und kleine Korpora existieren. Dieser Beitrag beschreibt außerdem die Ressourcen und Werkzeuge, die verwendet wurden, um das nötige Korpus zu erstellen und auch, wie dieses mit Wortart und Grundformen der Wörter angereichert wurde. Ein Nebenergebnis ist dabei die Untersuchung der Qualität von verfügbaren Taggern bzgl. Verben, Negationsmorphemen und Kongruenzpartikel

Stichwörter: WÖRTERBÜCHER AFRIKANISCHER SPRACHEN, KORPUSLINGUISTIK, NEGATIONEN, SEPEDI, NORD-SOTHO, LEXIKOGRAPHIE, TAGGING, BEARBEITUNG VON KORPUSANFRAGEN

1. Introduction

Negation is an important issue in language description because of the multiple forms in which it may occur. Attempts have been made to categorize negation alongside traditional linguistic fields like morphology (word level) and syntax (phrase level). Dahl (1979: 81), for example, initially distinguishes morphological and syntactical negation, describing syntactic negation as using "simple and double particles, negative auxiliaries and particle + dummy auxiliaries" while he sees negation on a morphological level as part of inflection. Yet he later adds (ibid: 83) that "in most cases" there is no such clear-cut description possible and that the distinction can only be made between negation morphemes as affixes (bound morphemes) or as particles (free morphemes).

This contribution investigates negation in the language Northern Sotho (ISO-code: NSO), also called Sepedi. The work is based on Prinsloo (2020), who describes the "lexicographic treatment of Sepedi negations" from the view of lexicography. In his article, he lists a number of forms that negation takes on, using theoretical work on Sepedi as knowledge base with the aim of properly describing Sepedi negation in a paper dictionary. Here, we will attempt to explore corpus data in a semi-automated way, utilizing existing Natural Language Processing (NLP) tools along the way.

There are two approaches to examine corpora: *corpus-based* and *corpus-driven* (Tognini-Bonelli 2001). The former starts with theoretical hypotheses about a language and investigates whether these are true, while the latter explores phenomena that are significant from a quantitative point of view (e.g. word sequences appearing frequently) to find new insights into the use of a language. In this paper, we start with a corpus-based approach by querying a cor-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

pus of pre-defined negated verb formations. Interestingly enough, we also come across new formations not defined in the literature.

NLP processing usually begins with tokens matching language units. As most NLP tools were initially developed for European languages, it is usually assumed that one token is either to be identified as a symbol (like punctuation) or equivalent to one word (which means: a free morpheme). For the South African indigenous languages, this assumption is however often not true. Sepedi, for example, utilizes the so-called "disjunctive writing system", that is bound morphemes are often written separately from the morpheme they belong to, hence negation affixes (bound morphemes) and negation particles (free morphemes) cannot be distinguished when tokenising. The tools therefore treat both as if they were particles.

Our approach is to work with language in use, so it is necessary to compile a corpus containing as many freely available sentences as possible. Before this corpus can be investigated properly, however, its tokens should be annotated with their respective part-of-speech (POS), so that queries can be performed on that level (e.g. to find all verbs). An annotation of lemmas (in the sense of a base form of each word) might also prove helpful, especially for languages with a rich morphology.

For the purpose of sentence collection, we make use of data available from the South African Centre for Digital Language Resources (SADiLaR[1]) and the Sepedi corpora collected by the CURL web collection machine located at Leipzig University (Goldhahn et al. 2016). For corpus annotation, NLP tools provided by SADiLaR and an own Sepedi tagging parameter file (Faaß et al. 2009), developed for the TreeTagger (Schmid 1994; Schmid 1995) are utilized. The corpus is encoded within the freely available IMS Open Corpus Query Workbench (OCWB) system (Evert and Hardie 2011), and the respective queries for the corpus are written as macros for the two purposes of a better documentation and reproducibility.

2. Aims

Corpus linguistics is not just a science in its own right; it can also be seen by other research fields as a helpful method that can be applied for their empiric research. In lexicography, utilizing corpus data is essential (see e.g. Faaß 2018).

In corpus linguistics, the language in use needs to be described and the resources utilized should contain utterances by as many speakers as possible, to at least get a grip on how certain linguistic phenomena appear in the living language. Frequencies of occurrences found in corpora assist in deciding which linguistic phenomena and/or word forms should preferably be included in a dictionary because a user of a dictionary — often a learner of the language — should find at least the most frequent word forms.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

Concerning Sepedi, there are often several ways to negate a verb, therefore the first aim of this contribution is finding the most frequent negation strategies of a selected verb[2] as a case in point with the aim of showing them in the right order in a dictionary entry. Secondly, Prinsloo (2020) describes the frequencies of occurrences of single negation morphemes; this paper will add morpheme (or particle) sequences appearing in the corpus thereby widening the issue to a morpho-syntactical description of negation strategies utilized. Although Prinsloo (ibid.) also dedicates a chapter to copulatives, this contribution focuses on full verbs.

3. Resources and their utilization

Starting from collecting a corpus of Sepedi sentences, we proceed with tokenising the texts and detecting sentence borders. Subsequently, the tokens are labelled (annotated) with POS utilizing two freely available taggers. A Sepedi lemmatiser is then used to add lemmas as additional labels to the tokens. The resulting corpus is encoded in the IMS Open CorpusWorkBench (OCWB, version 3.4.32) to ease its exploration. Lastly, the corpus is queried with the help of written macros to ensure reproducibility of results.

3.1 The SEPEDI2021 Corpus

The first task when collecting a corpus is to find as many utterances of the language as possible. Often, such resources are made available by repositories. The virtual language observatory hosted by the Common Language Resources and Technology Infrastructure, CLARIN[3] shows that there are data available from SADiLaR[4] and from the University of Leipzig. Additionally, Leipzig offers CURL (Goldhahn et al. 2016), an online web crawler tool into which URLs of web pages containing Sepedi text can be fed. After the fully automated crawling and pre-processing is completed, the resulting corpus is made available for download. CURL was already processed in 2017 generating a small Sepedi corpus, and again, with newer URLs collected by the author of this contribution in 2021.

We built our corpus using these available resources: SADiLaR's Sepedi Text corpus forms the biggest part of our corpus (Eiselen and Puttkammer 2014). In the Sepedi Speech corpus (De Vries et al. 2014), we change the boundary mark <orth> to <s>, that is count utterances as sentences in line with the other parts of our corpus. Adding the two CURL corpora, we compile a corpus of 154,204 sentences (2.7 million tokens), as shown in Table 1.

Name of the resource	*repository*	*no. of unique …*	*no. of tokens*
NCHLT Speech Corpus[5]	SADiLaR	56,284 utterances	238,905
NCHLT Text Corpus[6]	SADiLaR	83,614 sentences	2,224,593
NSO_Community 2017	Leipzig Wortschatz Project	4,746 sentences	113,392
NSO-Community 2021	Leipzig Wortschatz Project	9,560 sentences	178,005
Total		154,204 sentences	2,754,895
SEPEDI2021		81,274 sentences	2,327,390

Table 1: Parts of the SEPEDI2021 corpus

However, when using textual data from different sources, doublets must be expected. Therefore, we sort all sentences uniquely before further processing. We also manually delete sentences which have been collected even though they contain several words from other languages. The resulting SEPEDI-2021 corpus consists of 69,439 unique sentences (near doublets were not deleted). Lastly, we run a local tokeniser (its output is a one token per line format) that adds a number of sentence borders (some lines with "sentences" of the provided text corpora contained several sentences). We also change all occurrences of more than three dots into "…" (that is one token). Counting the output, we find that our final SEPEDI2021 corpus in total contains 2,327,390 tokens in 81,274 sentences.

3.2 Tagging the corpus with parts-of-speech (POS)

Unfortunately, the POS-tagger parameter file (Taljard et al. 2008, and Faaß et al. 2009) produced in 2009 for the rft-tagger (Schmid and Laws 2008) can only be used on 32-bit machines which have been replaced during the past decade with 64-bit machines. Using that parameter file, the RFT-Tagger achieved 94.16% precision (Faaß et al. 2009). The training material is no longer available, we therefore have to make use of the alternative and still usable TreeTagger parameter file reaching 92.46% precision (ibid.), using the label "tpos". The SADiLaR (NCHLT) tagger by Eiselen and Puttkammer (2014), claims a tagging precision of 96%, therefore we annotate its annotation as "npos" to the corpus, as well[7].

3.3 Lemmatising the corpus

Eiselen and Puttkammer (2014) also provide an NCHLT lemmatiser[8], a tool generating base forms for inflected word forms. However, applying this tool is different from similar ones: instead of directly utilizing it on a running text, one must provide the tool with a lowercase word list containing the words of the corpus (to save execution time, the list should be sorted uniquely beforehand). This way of processing is unusual — lemmatising like tagging is usually a process of looking at words in their context (seeing that many words are ambiguous and thus may have several lemmata to choose from). To annotate the lemmata which were identified in the corpus, it is necessary to write a tool using the output of the lemmatiser as an inventory. Examining the results of the lemmatiser, we note that many inflected words (especially verbs) of the corpus' word list were not lemmatised but remained in their original form. It seems that wherever there a lemma is unknown, the tool uses the word itself. Before adding these word forms, we change all characters to lower case to have at least an entry in the lemma field and to facilitate the formulation of queries at a later stage.

3.4 Corpus Annotation Overview/Encoding

The resulting corpus, ready to be encoded with the IMS Open CorpusWork-Bench (Evert and Hardie 2011), has a table format, containing the columns "word" for the original token, "tpos" (for tree-tagger POS), "npos" (for NCHLT POS and lemma (which might be the lower case version of the token). Note that the column titles of Table 2 will be utilized as attributes in the queries described from section 4 on.

Table 2 shows a small excerpt of the corpus, demonstrating the ambiguity of *a*, which according to the taggers in its first appearance is either a particle or a subject concord of noun class 1 (the npos annotation is correct). Both taggers annotate it as morpheme (of the present tense) in its second appearance (a table listing the tags utilized by both taggers can be found in the Appendix).

word	*tpos*	*npos*	*lemma*
<s>			
A	PART	CS01	a
pudi	N.09	V	pudi
re	CS.PERS	CSPERS	re
a	MORPH	MORPHPRES	a
feditše	V	V	feditše
.	$.	ZE	.
</s>			

Table 2: An example SEPEDI2021 sentence ready for encoding

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

4. Utilizing the Corpus Query Processor

4.1 General Information

With the IMS Open CorpusWorkBench (Evert and Hardie 2011), the tool Corpus Query Processor (CQP) is provided. This tool can easily be used to query corpus data not only on word level, but also on each of the annotation levels provided by the encoder by way of attribute-value constraints. The queries work on corpus positions (each containing one token and its annotations). A query for such a position is bound by "[]" (if not filled, all tokens in the corpus are found with this query). The query [word="a"] finds all occurrences of the token *a*, while a combination with the npos-attribute [(lemma="a") & (npos="CS01")] finds all upper and lower case occurrences of *a* being annotated by the NCHLT tagger as a subject concord of class 1.

In our corpus, the general query searching for the lemma *a* without any further constraint finds 67,623 occurrences. *a* is a highly ambiguous morpheme (see also Faaß et al. 2009). Table 3 shows the annotations and their frequencies as identified by the two taggers in the SEPEDI2021 corpus.

Annotation *NCHLT / TreeTagger*	*NCHLT tagger* *("npos")*	*TreeTagger* *("tpos")*
CPOSS06 / CPOSS.06	22,583	18,046
CS01 / CS.01	21,472	32,292
CS06 / CS.06	15,466	8,130
CD06 / CDEM06	4,335	3,605
MORPHPRES / MORPH	2,350	2,927
TENSE / MORPH	29	n.a.
PART / PART	451	1,972
CO06 / CO.06	440	628
RV (wrong tag)	152	—
CD01 / CDEM.01	146	0
RS (wrong tag)	112	—
VCOP / VCOP	54	5
QUE / QUE	33	0
Total	67,623	67,623

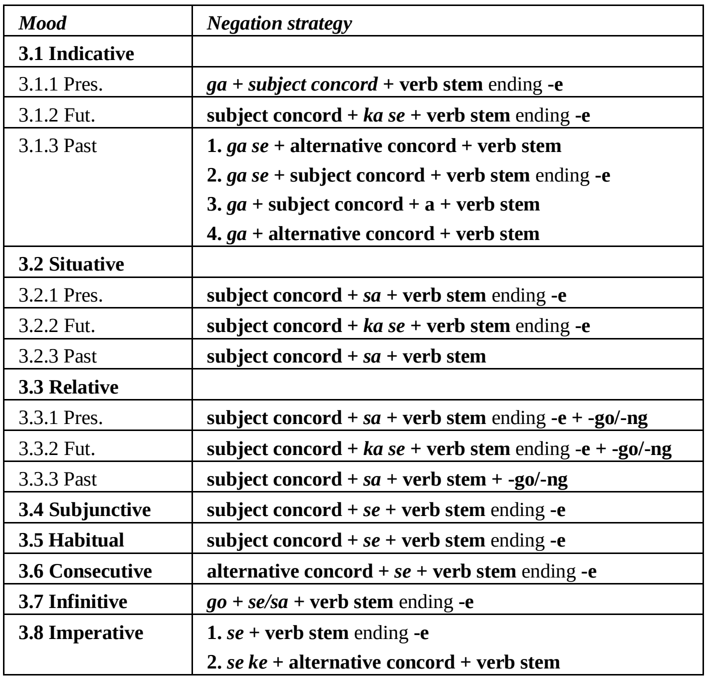
Table 3: Occurrences of *a* in the Sepedi2021 corpus with its npos- and tpos-annotations

CQP also allows for querying sequences of tokens, we can thus for example query the sequences of negation morphemes again on all available levels of annotation. Marking structural annotations like sentence borders in our queries ensures that we remain within a sentence when querying sequences. As an important advantage of this tool, we may make use of regular expressions (RegEx) when describing values to match. Using RegEx shortens the necessary processing time significantly. Therefore, we may describe the set of regular subject concords *ke, re, le, se, e, bo, go, o, ba, a, di,* in the compact regular expression '([klrs]?e | [bg]?o | b?a | di)'.

We formulate our queries at first on lemma level to be sure that incorrectly tagged items will still be found. However, as we would like to explore the tagging quality of the morphemes appearing in our structures, we will repeat the queries on npos and tpos level.

4.2 Developing macros for querying SEPEDI2021

In Table 1, Prinsloo (2020: 323) describes sequences of morphemes and conditions for a number of negation forms. We repeat parts of this table here as Table 4 that shows a productive perspective, in other words, how should a specific mood, tense and polarity be formulated? Working with corpus queries, we need to change to the receptive perspective: how should a sequence occurring in a corpus be interpreted?



Mood	*Negation strategy*
3.1 Indicative	
3.1.1 Pres.	***ga* + *subject concord* + verb stem** ending **-e**
3.1.2 Fut.	**subject concord + *ka se* + verb stem** ending **-e**
3.1.3 Past	**1. *ga se* + alternative concord + verb stem** **2. *ga se* + subject concord + verb stem** ending **-e** **3. *ga* + subject concord + a + verb stem** **4. *ga* + alternative concord + verb stem**
3.2 Situative	
3.2.1 Pres.	**subject concord + *sa* + verb stem** ending **-e**
3.2.2 Fut.	**subject concord + *ka se* + verb stem** ending **-e**
3.2.3 Past	**subject concord + *sa* + verb stem**
3.3 Relative	
3.3.1 Pres.	**subject concord + *sa* + verb stem** ending **-e + -go/-ng**
3.3.2 Fut.	**subject concord + *ka se* + verb stem** ending **-e + -go/-ng**
3.3.3 Past	**subject concord + *sa* + verb stem + -go/-ng**
3.4 Subjunctive	**subject concord + *se* + verb stem** ending **-e**
3.5 Habitual	**subject concord + *se* + verb stem** ending **-e**
3.6 Consecutive	**alternative concord + *se* + verb stem** ending **-e**
3.7 Infinitive	***go* + *se/sa* + verb stem** ending **-e**
3.8 Imperative	**1. *se* + verb stem** ending **-e** **2. *se ke* + alternative concord + verb stem**

Table 4: Mood and negation strategies (Prinsloo 2020: 323)

We utilize the descriptions of Table 4 for performing corpus queries, but there are challenges:

1. There are several identical sequences appearing in different moods (there is syncretism), see for example the negation of the future tense of Indicative (3.1.2) and Situative (3.2.2) or Subjunctive (3.4) and Habitual (3.5).
2. The infinitive (3.7) contains the highly ambiguous class prefix *go*. It would exceed the scope of this work to distinguish all infinitive class prefixes from the subject concords of class 15 and the locative classes. Therefore, we will count the syncretic cases where *go* appears separately.
3. The endings of verb stems are not described in a number of categories, we thus must use other, more detailed descriptions of the negation forms additionally to be able to precisely formulate our queries.
4. The available taggers do not differentiate between the different sets of subject and alternative concords; we must therefore also search our corpus on the levels of lemma or token, respectively even when trying to find tokens on "npos" or "tpos" level. Still, we will not be able to allocate some of our results to one specific mood (without a linguistic expert reading and interpreting all of the respective sentences).
5. Lastly, we need to be more precise in our descriptions, as we want to take transitive verbs into account that might be preceded by an object concord.

To solve issue 3 and 5, we rely on the morpheme sequences described in the PhD Dissertation of Faaß (2010), which was supervised by D.J. Prinsloo. These are based on the theoretical descriptions of Lombard et al. (1985), Louwrens (1991), and Poulos and Louwrens (1994) and thus identify more ways of negating. Exploring for example the negated future tense of the relative, Prinsloo (2020) provides one possibility: (1) subject concord + *ka* + *se* + verb stem (we presume that it ends in *-a*); using Faaß (2010) for comparison, there are two more strategies: (2) subject concord + *ka* + *se* + *tla/tlo* + verb stem ending in *-a* + *-go/-ng* and (3) subject concord + *ka* + *se* + *tlago/tlogo* + verb stem ending in *-a*. To find all possible forms, we will look for all of the described sequences, however, in our results, we will number them alongside the definitions of Prinsloo (2020) as shown in Table 4, so that a link to his article is established.

A typical token sequence describing the negation strategies for the indicative presence (ibid.) with the respective queries added, is described in Table 5.

Mood	*Negation strategy*	*CQP Queries*
3.1 Indicative		
3.1.1 Pres	***ga* + *subject concord* + *verb stem*** ending –e	**word:** [lemma ="ga"] [lemma="([klrs]?e|[bg]?o|b?a|di)"] [lemma="([lrs]?e|[bgm]?o|b?a|di)"]? [lemma=".+e"] **npos:** [(pos ="MNEG") & (lemma="ga")] [npos="CS.+"] [npos="OC.+"]? [(pos="V" & lemma=".+e"] **tpos:** [(pos ="MORPH") & (lemma="ga")] [npos="CS.+"] [npos="OC.+"]? [(pos="V" & lemma=".+e"]

Table 5: Transferring a negation description into CQP queries

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

We must be realistic: SEPEDI2021 is rather small and far from being representative of the language. We hence decide to only explore the most frequent verb in this corpus as a case in point, but in a reproducible way so that this exploration can be repeated on other OCWB-encoded corpora and with other verbs.

First, we produce a ranking list of all tokens tagged as V and find *feta* ((to) "pass"/"exceed"[9]) on a high-ranking position as the most frequent unambiguous verb form with 2,025 (npos)/2,026 (tpos) occurrences. Other verbs are more frequent, but at the same time ambiguous in terms of their POS, so there is a risk of them being incorrectly tagged. Taking all of feta's inflectional and derivational forms appearing in the corpus, we count their occurrences in all moods and tenses in order to get an overview of the forms that the verb appears in. A positive intermediate result is that all of them are annotated as "V" by both taggers.

The past form of the Indicative, for instance, is described by four different negation strategies in our Table 4 (Table 1 of Prinsloo 2020: 323). Therefore it might be of interest to lexicographers, which negation strategies are followed for this verb or for that matter any other verb. Since such queries are stored in text files as so-called "macros", they can be freely exchanged between researchers and re-used at any given time.

Future investigations regarding other verbs are made possible because our queries are furnished with a variable ($0 in the queries shown below) that will be replaced by the query processor with a regular expression describing any verb stem provided at the time of query.

Table 6 shows how a regular expression (RegEx) is built for the existing *fet-* forms for exemplification reasons.

Freq. of occ.	*Word form*	*Comments*	*Translation*	*Building a RegEx (ignore upper/lower case)*
		Indicative		
2,024	*feta*	active (-a)	pass, exceed	
85	*fete*	active (-e)	(must) pass, exceed	fet[ae]
27	*fetwa*	passive (-a)	is passed, exceeded	
5	*fetwe*	passive (-e)	(must) be passed, exceeded	fetw?[ae]
6	*fetana*	active reciprocal (-a)	pass, exceed each other	fet(w?[ae] | ana)
92	*fetile*	active perfect (-ile)	passed, exceeded	
6	*fetilwe*	passive perfect (-ile)	was passed, exceeded	fet(il | an)?w?[ae]
		Relative		
153	*fetago*	active (-go)	who/which pass(es), exceeds	
2	*fetang*	active (-ng)	who/which pass(es), exceeds	feta(go | ng)
81	*fetego*	active (-go)	who/which does not pass, exceed	
1	*feteng*	active (-ng)	who/which does not pass, exceed	fet[ae](go | ng)
7	*fetwago*	passive (-go)	who/which is passed, exceeded	
5	*fetwego*	passive (-ng)	who/which is not passed, exceeded	fetw?[ae](go | ng)
521	*fetilego*	active perfect (-go)	who/which passed, exceeded	
9	*fetileng*	active perfect (-ng)	who/which passed, exceeded	
9	*fetilwego*	passive relative (-go)	who/which was passed, exceeded	fet(il)?w?[ae](go | ng)
3,033				

Table 6: Generating regular expressions for the occurring word forms of *fet-*

We still need to deal with the problem of syncretism: 3.2.2 (future tense of the situative) in Prinsloo's table (2020: 323) is identical to 3.1.2 (future tense of the indicative). This shows that we cannot distinguish the two moods by only looking at the token sequences described. As the situative is often preceded by the conjunction *ge* ("when"), we first explore the typical distance between *ge* and a following verb and find a maximum of 3 tokens appearing, one of which may never be punctuation. We therefore add the condition that the token *ge* for the negated form of the future indicative may not appear in up to 3 tokens preceding the described sequence (while, for the situative, such an occurrence of *ge* is defined as obligatory). We are however conscious of the fact that the problem may only be partially solved.

An exemplifying macro is IND-FUT-NEG(1) in Figure 1. It describes the negated future tense of the indicative on lemma level (all of the indicative forms are summarized in Table 3.19 of Faaß (2010)). The variable $0 will be replaced by a verb form (without ending) when starting the macro (where the ending is pre-defined). Results of the query will first be written into a sub-corpus called _IND-FUT-NEG. These matches are then counted on a lemma level and the resulting table is written into a file called ind-fut-neg.csv (see Figure 2).

```
MACRO IND-FUT-NEG(1)
set MatchingStrategy longest;
show -cpos;
_IND-FUT-NEG =
[lemma!="ge"]{0,3}                  # no ge/Ge should precede the sequence
# future tense negative
[lemma="(b?a|[klrs]?e|[bg]?o|di)"]  # CS
[lemma="ka"]                        # MORPH_neg
[lemma="se"]                         # MORPH_neg
[lemma="([gbm]?o|b?a|[ls]?e|di)"]? # possible OC
[lemma="$0e"];                      # verb stem ending in -e
cat _IND-FUT-NEG;
count by word  > "/Users/faassg/corpora/sepedi2021/work/ind-fut-neg.csv";
;
```

Figure 1: Macro IND-FUT-NEG(1) finding all negated future tense indicative forms of a specific verb stem

The result of the macro IND-FUT-NEG(1), processed with the regular expression "fetw?" (the question mark stands for optionality of the previous character, thus we describe the active and the passive form of future tense) is shown in Table 7. It should be noted that the original .csv file contains matches with up to three tokens preceding the verb (described by [lemma!="ge"]{0,3}). We find 5 different sequences (types), one verb form in altogether 11 occurrences.

CS	*OC*	*NEG*	*NEG*	*VERB*	*freq*
le		*ka*	*se*	*fete*	6
e		*ka*	*se*	*fete*	3
a		*ka*	*se*	*fete*	1
di		*ka*	*se*	*fete*	1
				Total	11

Table 7: Summarized results of the Macro "IND-FUT-NEG", run with the verb stem expressed as "fet(il)?w?"

5. Results for *feta*

5.1 Results of the macros

In section 4 above, we developed a regular expression describing all indicative forms of the verb *fet-* appearing in the SEPEDI2021 corpus: fet(il|an)?w?[ae]. However, we need to delete the verbal endings [ae], as they are already described by the macro (see Table 1). The active reciprocal form *fetana* does not appear in the corpus with any other endings, it is thus only queried in the respective moods, tenses and polarities that display verbs ending in *-a*.

Table 8 shows our results. As mentioned above, occurrences of *go* in the indicative verbal phrases might be infinitives, therefore they are counted separately. The same applies to *se* in the subjunctive/habitual that was queried simultaneously. The morpheme *se* might be a negation morpheme instead of a subject or an object concord. To avoid counting identical forms twice for the subjunctive/habitual, all cases where *se* appears are counted as negated forms of the verb.

mood/tense	*macro run*	*type of sequences found*	*found verb forms fet-*	*freq*	*freq of go*	*Total*
3.1 Indicative						
3.1.1. Pres	IND-PRES-POS["fetw?"]	37	w?a	471	1,532	2,003
	IND-PRES-NEG["fetw?"]	7	w?e	16	0	16
3.1.2. Fut	IND-FUT-POS["fetw?"]	4	-a	10	0	10
	IND-FUT-NEG["fetw?"]	5	-e	5	0	5
3.1.3 Past	IND-PERF-POS["fet?"]	16	-ilw?e	126	n.a.	126
	IND-PERF-NEG["fetw?"]	2	-a	9	0	9
3.2 Situative						
3.2.1 Pres	SIT-PRES-POS["fetw?"]	15	-a	53	n.a.	53
	SIT-PRES-NEG["fetw?"]	0				0
3.2.2 Fut	SIT-FUT-POS["fetw?"]	0				0
	SIT-FUT-NEG["fetw?"]	0				0
3.2.3 Past	SIT-PERF-POS["fet(il)?w?"]	6	-ile	13	n.a.	13
	SIT-PERF-NEG["fet(il)?w?"]	0				0

3.3 Relative						
3.3.1 Pres	REL-PRES-POS["fetw?"]	24	-w?ago/ -ang	154	n.a.	154
	REL-PRES-NEG"fetw?"]	11	w?ego	78	n.a.	78
3.3.2 Fut	REL-FUT-POS["fetw?"]	0				0
	REL-FUT-NEG"fetw?"]	0				0
3.3.3 Past	REL-PERF-POS["fetilw?"]	0				0
	REL-PERF-NEG"fetw?"]	1	-a	8	n.a.	8
mood/tense	*macro run*	*type of sequences found*	*found verb forms fet-*	*freq*	*freq of se*	*SUM*
3.4 / 3.5 Subjunctive and Habitual	SUBJ-HABIT-POS["fetw?"]	11	-w?e	30	53	11
	SUBJ-HABIT-NEG["fetw?"]	10	-w?e		53	53
3.6 Consecutive	CONSEC-POS["fetw?"]	0				
	CONSEC-NEG["fetw?"]	0				
3.7 Infinitive	see column "go"					
3.8 Imperative	IMP["fet"] (pos and neg)	0				0
	sum					2,631

Table 8: Summarized results of the occurrences of forms of *fet-* in the corpus

5.2 Corpus data for developing dictionary entries for *fet-*

To find data usable for a theoretical dictionary entry for *feta* based on our (non-representative) corpus data, we explore the forms found and the sub-corpora generated by the macros in more detail. As these data are too extensive to be shown completely in a contribution of this kind, we attempt to summarize them here. Again, it must be stressed that any interpretation must be conscious of Sepedi syncretism and the non-representativeness of the corpus.

5.2.1 General data on the occurring word forms of *fet-*

fet- is not very frequently passivized (59 passive voice forms versus 2,968 active forms) and only one derivation, namely *fetana* ("pass, exceed each other") appears in the corpus. When forming the relative, the ending *-go* is clearly preferred (764 occurrences) while its alternative ending, *-ng*, only occurs 12 times.

5.2.2 Data on the occurrences of *fet-* in the different moods

As it is typical for most verbs, the use of the indicative seems to be decidedly preferred (2,175 occurrences), while the number of occurrences of the relative (240) is significantly higher than that of the situative (43). The situative/habitual appears a few times, but none of the other moods can be found in the corpus.

Concerning negation strategies, we find the following data:

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

1. The indicative in the perfect tense (3.1.3 in Table 4) allows for four ways of negation, in all 8 occurrences of this negated mood *feta* makes use of *ga* + alternative subject concord + verb stem.
2. The negated past tense of the situative was described by Prinsloo as subject concord + *sa* + verb stem. Faaß (2010), on the basis of Lombard et al. (1985: 149), describes three possible ways, of which one appears in the corpus: *a se a fetwa* (subject concord + *se* + alternative subject concord + verb ending in -*a*.

5.2.3 Data on the occurrences of *fet-* in the different tenses

The present tense dominates the occurrences (2,289) of *fet-*, while the past/ perfect tense appears far less frequently (154). The future tense seems to be rather irrelevant in this corpus (26).

5.2.4 Data on the polarity of *fet-*

As is the case for most verbs, its positive form appears by far more frequently: 2,352 positive sequences appear versus 170 negated sequences. However, for the relative, 86 negated sequences stand against 154 positives. Not counting the fact that the corpus is rather small, such data could lead to the suspicion of specific semantics of the verb in the relative — an aspect that could be further explored.

5.2.5 Data on the transitivity of *fet-*

We do not have a full overview of the transitivity of the verb *fet-* because we only check for occurrences of the object concord which stands for a known object in the discourse. An object usually occurs after the verbal structure. However, occurring object concords give a clear indication that the verb may appear in a transitive reading.

fet- appears in the corpus with and without object concords. In the indicative positive of the present tense, for example, there are 1,713 occurrences found without an object concord, of these, 1,535 show a preceding *go* (pointing to a possible infinitive). We find 7 occurrences where a subject concord is followed by *a* before the verb in its base (active and passive) form appears. This *a*, as stated above, may either be interpreted as a tense morpheme (long form of the present tense) or as an object concord. We find 13 more occurrences of *go* as the first element of the sequence, followed by a morpheme that could be an object concord. Lastly, we are left with 276 sequences in which the verb undoubtedly follows a sequence of subject and object concord.

5.3 Open issues for *fet-*

Altogether, 3,033 word forms of *feta* should have occurred as part of the predefined sequences, but only 2,522 were found. While syncretism certainly leads to finding doublettes, there are also some sequences found that do not appear as defined in books written for language learners.

Following these books, the relative, for example, should only appear as *fetilego,* namely with the verbal ending *-ego* when preceded by the negative morpheme *sa* in the negated perfect tense of the indicative. Preceded by *sa*, it does not occur at all in our corpus, with *ka se*, we also do not find any occurrences in SEPEDI 2021. Hence, all 539 occurrences of *fetilw?ego* are part of other sequences that we cannot define on the basis of the given literature. It would exceed the scope of this contribution attempting to interpret these cases. However, for possible future work in collaboration with linguists, Table 9 shows the forms and their preceding items as they occur in the corpus.

pos -2	***pos-1***	***found verb forms fet-***	***SUM***
ye	*e*	*-ilego*	222
wo	*o*	*-ilego*	84
tše	*di*	*-ilego*	50
yeo	*e*	*-ilego*	41
se	*se*	*-ilego*	26
le	*le*	*-ilego*	22
ao	*a*	*-ilego*	14
leo	*le*	*-ilego*	12
tšeo	*di*	*-ilego*	7
a	*a*	*-ilego*	6
bao	*ba*	*-ilego*	6
bjo	*bo*	*-ilego*	5
ye	*e*	*-ileng*	4
yeo	*e*	*-ilwego*	4
lebaka	*le*	*-ilego*	3
tše	*di*	*-ileng*	3
mmalwa	*ye*	*-ilego*	2
seo	*se*	*-ilego*	2
woo	*o*	*-ilego*	2
yeo	*e*	*-ileng*	2
3	*e*	*-ilego*	1
ao	*a*	*-ilwego*	1
bangwe	*ba*	*-ilego*	1
bao	*ba*	*-ilwego*	1
bja	*bao*	*-ilego*	1
bošegong	*bjo*	*-ilego*	1
6	*tše*	*-ilego*	1
e	*e*	*-ilego*	1
go	*go*	*-ilego*	1
go	*go*	*-ilwego*	1

go	*tše*	*-ilego*	1
kgoro	*ye*	*-ilego*	1
lekgolo	*e*	*-ilego*	1
mabaka	*a*	*-ilego*	1
mabakeng	*a*	*-ilego*	1
mengwaga	*ye*	*-ilego*	1
mo	*go*	*-ilego*	1
pedi	*tše*	*-ilego*	1
tse	*di*	*-ilego*	1
tše	*di*	*-ilwego*	1
tšeo	*di*	*-ilwego*	1
yo	*a*	*-ilego*	1
Total			539

Table 9: Occurrences of *fetilw?ego* in the corpus

6. Results for part-of-speech sequences

The macros are executed in two additional modified versions where sequences were queried on the basis of npos and tpos. Whenever the POS-set category included more than one item though it is explicitly specified in the definitions given, the item is named on lemma-level in these macros. For example, when a negative form has to contain the negative morpheme *ga*, the constraint is formulated [npos="MORPHNEG" & lemma="ga"] or [tpos="MORPH" & lemma="ga"], because other items like *se* are also classified as MORPH(NEG). If the POS category contains only one member, for example the present tense morpheme *a*, the constraint is defined on the POS level only ([npos="MORPHPRES"]).

mood/tense	***polarity***	***queried word form RegEx***	***verbal ending (def. in macro)***	***Total found for lemma***	***Total found for npos***	***Total found for tpos***
3.1 Indicative						
3.1.1. Pres	pos	["fetw?"]	-a	2,003	213	638
	neg	["fetw?"]	-e	16	0	14
3.1.2. Fut.	pos	["fetw?"]	-a	10	10	8
	neg	["fetw?"]	-e	5	0	1
3.1.3 Past	pos	["fet"]	(il|etš)w?e	126	102	108
	neg	["fetw?"]	-a, -e	9	0	1
3.2 Situative						
3.2.1 Pres	pos	["fetw?"]	-a	53	58	64
	neg	["fetw?"]	-e	0	0	0
3.2.2 Fut	pos	["fetw?"]	-a	0	5	3
	neg	["fetw?"]	-e	0	0	0
3.2.3 Past	pos	["fet"]	-(il|etš)w?e	13	20	16
	neg	["fet"]	-(il|etš)w?e	0	0	0

3.3 Relative						
3.3.1 Pres	pos	["fetw?"]	-a(go|ng)	154	145	148
	neg	["fetw?"]	-e(go|ng)	78	0	75
3.3.2 Fut	pos	["fetw?"]	-a(go|ng)?	0	0	0
	neg	["fetw?"]	-a(go|ng)?	0	0	0
3.3.3 Past	pos	["fetilw?"]	-a, -e	0	0	0
	neg	["fetw?"]	-a, -e	8	0	0
3.4 / 3.5 Subjunctive and Habitual	pos	["fetw?"]	-e	11	31	27
	neg	["fetw?"]	-e	53	0	11
3.6 Consecutive	pos	["fetw?"]	-a	0	0	0
	neg	["fetw?"]	-e	0	0	0
3.8 Imperative	pos	["fet"]	-a(ng)?	0	0	0
	neg	["fet"]	-e(ng)?	0	0	0
				2,631		

Table 10: Searching on npos and tpos-level for *fet-* in the corpus

In order to evaluate the taggers, in a first run, the respective verb forms of *fet-* were queried as well. Table 10, repeating the totals of Table 8 of the queries of the lemma level (for comparative reasons) shows the results. Note that in the case of the situative, all conjunctions were permitted to appear ([n/tpos="CONJ"]). There were more matches (*ge, gore, ebile* etc.) occurring than were found for the situative queried on lemma-level (only *ge*).

Comparing the number of occurrences found with and without a POS constraint, it is quite clear that the tagging quality especially of the ambiguous morphemes is still a problem. Here, the finely grained "MORPH" definitions of the NCHLT tagger seem especially problematic: to distinguish MORPHNEG, MORPHPRES and TENSE reduces the number of cases and leads to problems when training a heuristic tagger. For the TreeTagger, the tag "MORPH" was chosen for all of the abovementioned morphemes because they all appear in similar positions, that is within a similar context. Because of the more coarse-grained label, the tool can find more occurrences of this type of token in the training phase and thus its precision is enhanced.

As a computational linguist, one would need to dig deeper into this evaluation, however for the purpose of this article we can summarize that querying on lemma level without using POS constraints might be the better option — until such time that the tagging quality of the ambiguous items is enhanced.

Finally, we tried the same macros again, now without a constraint on the verb root. Table 11 (columns "npos" and "tpos") shows the numbers of occurrences of all sequences finalized with tokens annotated as verbs (with their verbal endings as defined above for each mood, tense, and polarity). Again, we must assume a number of doublettes (see for example the high number of subjunctives/habituals identified), caused by the syncretism explained above. Others will be tagged incorrectly — all in all, we can however get a general indication of which moods, tenses and polarities appear more frequently in the corpus than others. We know that the results do not seem sufficient for highly ambiguous items, so lastly, the macros were again defined on a lemma level — however now using [npos="V"] as the only constraint on their final element (adding the necessary endings as above). Results are shown in column "lemma+npos".

As the tagging results do not seem sufficient for highly ambiguous items, lastly, the macros were again defined on lemma level — however now using [npos="V"] as the constraint on their final element.

mood/tense	*polarity*	*verbal ending (def. in macro)*	*npos*	*tpos*	*lemma+npos*	*Totals*
3.1 Indicative						
3.1.1. Pres	pos	-a	49,072	64,053	138,909	
	neg	-e	0	1,720	2,819	
3.1.2. Fut.	pos	-a	8,244	7,097	8,583	
	neg	-e	0	47	1,162	
3.1.3 Past	pos	(il|etš)w?e	13,207	20,755	14,959	
	neg	-a, -e	0	201	1,384	
						167,816
3.2 Situative						
3.2.1 Pres	pos	-a	9,805	7,174	6,522	
	neg	-e	0	430	513	
3.2.2 Fut	pos	-a	1,420	802	115	
	neg	-e	0	1	74	
3.2.3 Past	pos	-(il|etš)w?e	2,654	2,526	1,194	
	neg	-(il|etš)w?e	0	155	155	
						8,573
3.3 Relative						
3.3.1 Pres	pos	-a(go|ng)	21,934	22,029	24,568	
	neg	-e(go|ng)	0	1,496	1,574	
3.3.2 Fut	pos	-a(go|ng)?	1,234	1,221	1,556	
	neg	-a(go|ng)?	0	2	72	
3.3.3 Past	pos	-a, -e	21,934	22,029	24,387	
	neg	-a, -e	0	0	201	
						52,358
3.4 / 3.5 Subjunctive and Habitual	pos	-e	38,691	48,866	38,691	
	neg	-e	0	760	2,224	
						40,915
3.6 Consecutive	pos	-a	0	0	0	
	neg	-e	0	0	0	
3.8 Imperative	pos	a(ng)?	0	0	1	
	neg	e(ng)?	0	0	1	

Table 11: Searching on npos and tpos-level for all tokens annotated as verbs in the corpus

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

6.1 Data on the occurrences of verbs in different moods

Like in the case of *fet-*, the indicative dominates the field with (reading column lemma+npos) 167,816 occurrences. The relative occurs with 52,358 occurrences, while the situative is again on rank three with 8,573 occurrences.

6.2 Data on the occurrences of verbs in different tenses

174,905 present tense sequences are found (some of which might however be infinitives), followed by past/perfect tense with 42,280 occurrences. The third rank is reserved for the future tense (11,562).

6.3 Data on the polarity of verbs in general

259,485 of all moods appeared in the positive, while 10,179 sequences were negated. The relation in the relative mood between the positive and the negative polarity does not seem significant.

7. Summary and possible future work

In this contribution, we attempted to gain some insights into how a Sepedi corpus can be compiled and annotated, and how it may assist a lexicographer with exploring a specific verb as it is used in the language. Corpus data will also assist when sorting negations of Sepedi verbs in a dictionary according to the frequencies they appear in.

We chose the verb *fet-* as a case in point because it is an unambiguous verb occurring frequently in our corpus. The majority of its occurrences could be assigned to pre-defined moods, tenses and polarities. We found that this verb has intransitive and transitive uses, that it occurs in the passive, but only one of the many possible derivations appeared in our corpus. In the case of the relative, speakers of the language seem to prefer the ending *-go* instead of *-ng* which would be available, too.

Given a bigger and more representative corpus, one could inter alia explore derivations of this and other verbs, however this corpus is at least a starting point.

In addition to the lack of resources, we find three main challenges when switching from a prescriptive to a receptive perspective:

1. Syncretism is certainly the biggest problem when analysing morphology and/or syntax of Sepedi sentences. Language experts together with computational linguists could in future work closely together exploring these constellations in more detail in an attempt to find more indicators in texts helping to disambiguate. In the longer term, we could even try to re-define

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

the modal system as it is always problematic — not only for learners of the language — to distinguish token sequences semantically when they are 100% identical.

2. For highly ambiguous bound morphemes, tagging corpora with POS should help with the disambiguation, but the tagging quality does still not seem sufficient for such items (maybe this is caused by inappropriate tag-sets, too). Here, newer technologies, possibly deep learning as already implemented for example by Schmid (2019) might be of help.
3. When comparing grammar books and corpus data, we find constellations which were not explained or described in standard grammars. It is therefore necessary to explore the living language further and to adapt the grammar books following a descriptive approach.

All results of this work are reproducible since the SEPEDI2021 corpus consists of freely available data, and since this corpus is annotated with freely available tools. In view of the fact that it is compiled from sources generated by others, it may not be forwarded to other researchers because of legal reasons. The corpus queries described here are stored in macros that the author shares freely on request by other non-commercial researchers.

8. Endnotes

1. URL: https://sadilar.org
2. It would go beyond the scope of this article to show negation strategies for all verbs (the corpus is too small for this), however the corpus queries developed here are written so that they are utilizable for other verbs, too.
3. See https://vlo.clarin.eu. The CLARIN VLO collects metadata about available resources and tools for language research.
4. See https://sadilar.org. SADiLaR offers its own repository, but also reports its resources to CLARIN.
5. See https://repo.sadilar.org/handle/20.500.12185/270?show=full for more details.
6. See https://repo.sadilar.org/handle/20.500.12185/330?show=full for more details.
7. The MBT tagger parameter file used for a demo show case tagger on the AFLAT pages by De Pauw and De Schryver (https://aflat.org/sothotag) is not available for download, and we did not find any other available taggers for Sepedi.
8. Available at https://repo.sadilar.org/handle/20.500.12185/326 though not mentioned in the SADiLaR list of Sepedi tools provided by the repository.
9. All translations in this paper are taken from the *Oxford School Dictionary: Northern Sotho and English*. Oxford University Press. 2007.

9. Bibliography

Dahl, Ö. 1979. Typology of Sentence Negation. *Linguistics* 17: 79-106.

De Vries, N., M. Davel, J. Badenhorst and W. Basson. 2014. A Smartphone-based ASR Data Collection Tool for Under-resourced Languages. *Speech Communication* 56(1): 119-131.

Eiselen, E. and M. Puttkammer. 2014. Developing Text Resources for Ten South African Languages. *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era. LREC, 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, May 26-31, 2014:* 3698-3703.

Evert, S. and A. Hardie. 2011. Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics 2011 Conference, ICC Birmingham, 20–22 July 2011.* Birmingham: University of Birmingham.

Faaß, G. 2010. *A Morphosyntactic Description of Northern Sotho as a Basis for an Automated Translation from Northen Sotho to English.* Ph.D. Dissertation. Pretoria, South Africa: University of Pretoria.

Faaß, G. 2018. Lexicography and Corpus Linguistics. Fuertes-Olivera, P. (Ed.). 2018. *The Routledge Handbook of Lexicography:* 123-137. Oxon, UK: Routledge.

Faaß, G., U. Heid, E. Taljard and D. Prinsloo. 2009. Part-of-Speech Tagging of Northern Sotho: Disambiguating Polysemous Function Words. *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages (AfLaT 2009), Athens, Greece, 31 March 2009:* 38-45.

Goldhahn, D., M. Sumalvico and U. Quasthoff. 2016. Corpus Collection for Under-resourced Languages with More than One Million Speakers. Soria, C. et al. 2016: *LREC 2016 Workshop: Collaboration and Computing for Under-resourced Languages: Towards an Alliance for Digital Language Diversity (CCURL 2016), Portorož, Slovenia, 23 May 2016:* 67-73.

Lombard, D., E. van Wyk and P. Mokgokong. 1985. *Introduction to the Grammar of Northern Sotho.* Pretoria: J.L. van Schaik.

Louwrens, L. 1991. *Aspects of Northern Sotho Grammar.* Pretoria: Via Afrika.

Poulos, G. and L. Louwrens. 1994. *A Linguistic Analysis of Northern Sotho.* Pretoria: Via Afrika.

Prinsloo, D.J. 2020. Lexicographic Treatment of Negation in Sepedi Paper Dictionaries. *Lexikos* 30: 321-345. doi: https://doi.org/10.5788/30-1-1610

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing. Manchester, UK.*

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland:* 47-50.

Schmid, H. 2019. Deep Learning-based Morphological Taggers and Lemmatizers for Annotating Historical Texts. *Proceedings of DATeCH,* May 2019, Brussels, Belgium.

Schmidt, H. and F. Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging. Scott, D. and H. Uszkoreit (Eds.). 2008. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 18–22 August 2008, Manchester, UK. Vol. 1:* 777-784. Manchester: COLING.

Tognini-Bonelli, E. 2001. Corpus Linguistics at Work. *Studies in Corpus Linguistics*. Amsterdam/ Philadelphia: John Benjamins.

Appendix: NCHLT and TreeTagger Tagsets

Morpheme	*NCHLT tagger**	*TreeTagger*
Verbs		
auxiliary	VAUX	VAUX
copulative	VCOP	VCOP
others	V	V
Nouns		
regular	N01a, N02b, N01-N10, N14, N16-N18, NLOC	N.01a, N.02b, N.01-N.10, N.14, N.LOC
name of place	—	NPP
abbreviation	—	ABBR
Pronouns		
emphatic	PROEMP01-PROEMP10, PROEMPLOC, PROEMPPERS	PRO.EMP.01-PRO.EMP.10, PRO.EMP.14, PRO.EMP.LOC, PRO.EMP.PERS
possessive	PROPOSS02-PROPOSS10, PROPOSS14, PROPOSSPERS	PRO.POSS.01-PRO.POSS.10, PRO.POSS.LOC, PRO.POSS.PERS
quantitative	PROQUANT01-PROQUANT10, PROQUANT14, PROQUANTLOC	PRO.QUANT.01-PRO.QUANT.10, PRO.QUANT.14-PRO.QUANT.15, PRO.QUANT.LOC
question word	QUE	QUE
Adverbs	ADV	ADV
Adjectives	ADJ01-ADJ10, ADJ14, ADJLOC	ADJ.01-ADJ.10, ADJ.14-ADJ15, ADJLOC
Morphemes		
negative	MNEG	MORPH
future	MORPHFUT	MORPH
? (always: *sa*)	MORPHPER	MORPH
potential (.**ka*)	MORPHPOT	MORPH
present tense (*w?a*)	MORPHPRES	MORPH
infinitive (*go)*	INF	MORPH
aspectual prefix (*no*)	ASP	MORPH
tense marker	TENSE	—
Concords		
subject	CS01-CS10, CS14-CS15, CSINDEF, CSLOC, CSNEUT, CSPERS	CS.01-CS10, CS.14-CS.15, CS.INDEF, CS.LOC, CS.NEUT, CS.PERS
object	CO01-CO10, CO14, COPERS	CO.01-CO.10, CO.14-CO.15, CO.LOC, CO.PERS
possessive	CPOSS01-CPOSS10, CPOSS14-CPOSS17, CPOSSLOC	CPOSS.01-CPOSS.10, CPOSS.14-CPOSS.15, CPOSS.LOC
demonstrative	CD01-CD10 CD14-CD18 CDLOC	CDEM.01-CDEM.10, CDEM.14, CDEM.COP, CDEM.LOC
Conjunctions	CONJ	CONJ
Particles		
question	PARTQUE	PART
others	PART	PART

Interjections	INT	INT
Enumeratives	ENUM	ENUM
Ideophones	IDEO	IDEO
Numerals	RS	NUM
Ordinals	RS	ORD
Punctuation		
.?	ZE	
!	ZE!	
,;-:	ZM	
left brackets/quotes	ZPL	
right brackets/quotes	ZPR	
.?!,;:		$.
brackets, quotes		$"
/\\-%&		$-
Others		
Abbreviation of *Morena, Mna.* (=Mister, Mr.)	RO	ABBR
guess: foreign language material, however a number of Sepedi names (N01A and NPP) are tagged as RV	RV	—

* A full description of the NCHLT tagset could not be found, hence only the categories appearing in the corpus are described by the author in this table.