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.

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

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.

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

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

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

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.