Automatic Detection of Copulatives in Northern Sotho corpora

Gertrud Faaß and Elsabé Taljard

Universities of Hildesheim and Pretoria
5th international Conference on Bantu Languages
Paris, June 12th to 15th, 2013

June 14th, 2013

Project Background

Scientific e-Lexicography for Africa, SeLA

- Universites of Hildesheim,
  Pretoria, Stellenbosch, South Africa (UNISA), and Windhoek
- Prototype e-dictionaries
  for several of the South African National languages
  (June 2012 – May 2015)
- Several sub-projects: specifically: acquisition tools and data
- Our task: a corpus linguistic study of the NSO copulative:
  - Which of the described constellations exist in the available corpus?
  - What are the frequencies of occurrence?
  - Can we learn anything about typical complements?
- Theoretical background: "Lexicographic Function Theory"

"The Function Theory"

Main Development: Centlex in Aarhus (see URL in link list)

Central notion is the purpose ("function") of a dictionary, e.g.

- *I need to understand words, phrases or sentences* $\rightarrow$ reception
- *I need to generate words, phrases or sentences myself* $\rightarrow$ production

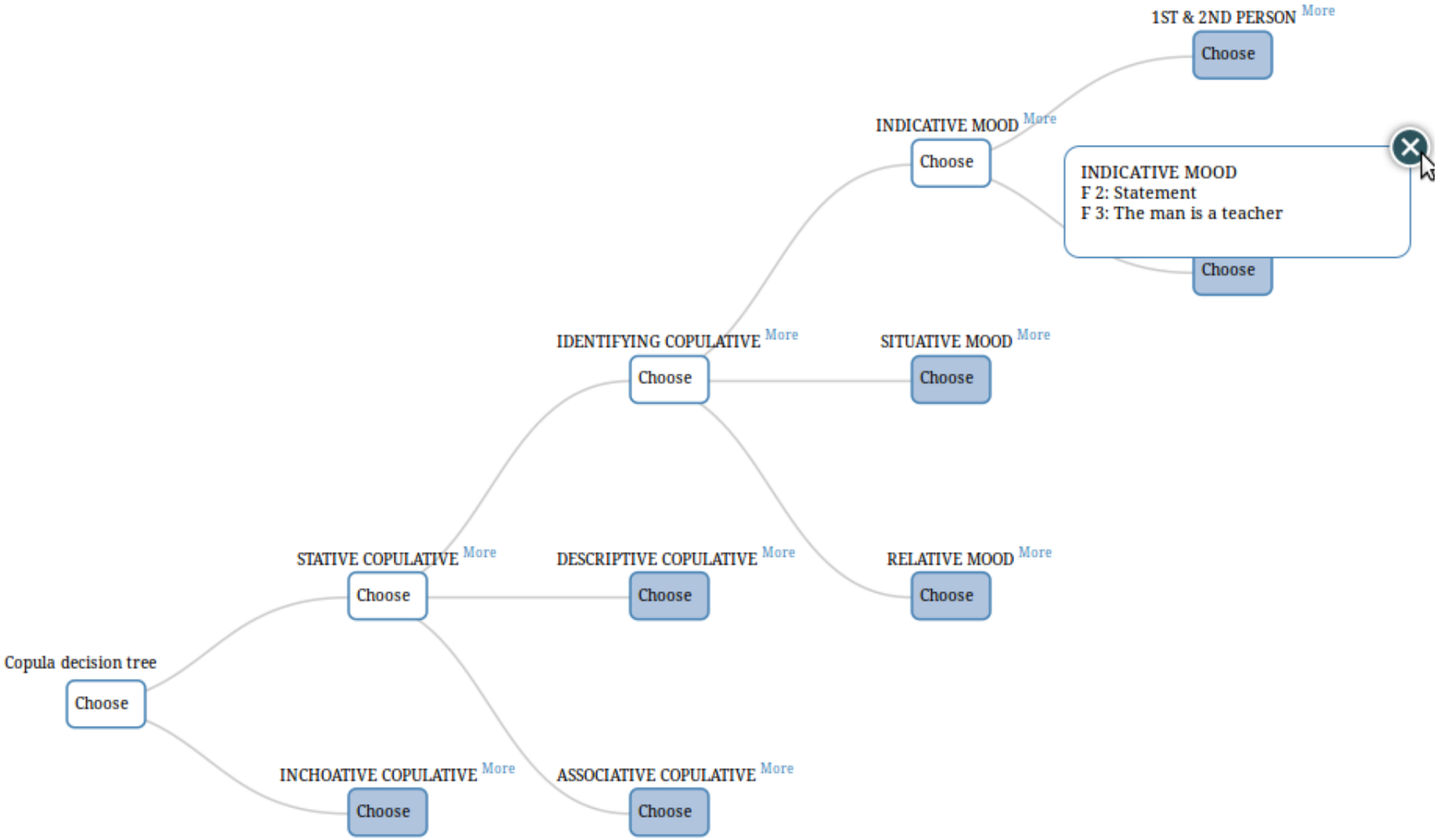
What do we need?

Production purposes

- A database containing all possible forms of NSO copulatives
- Add glosses, translations and examples (if possible, from corpora)
- Guide users in their text production, e.g. by means of a decision tree: Selection of appropriate copulatives

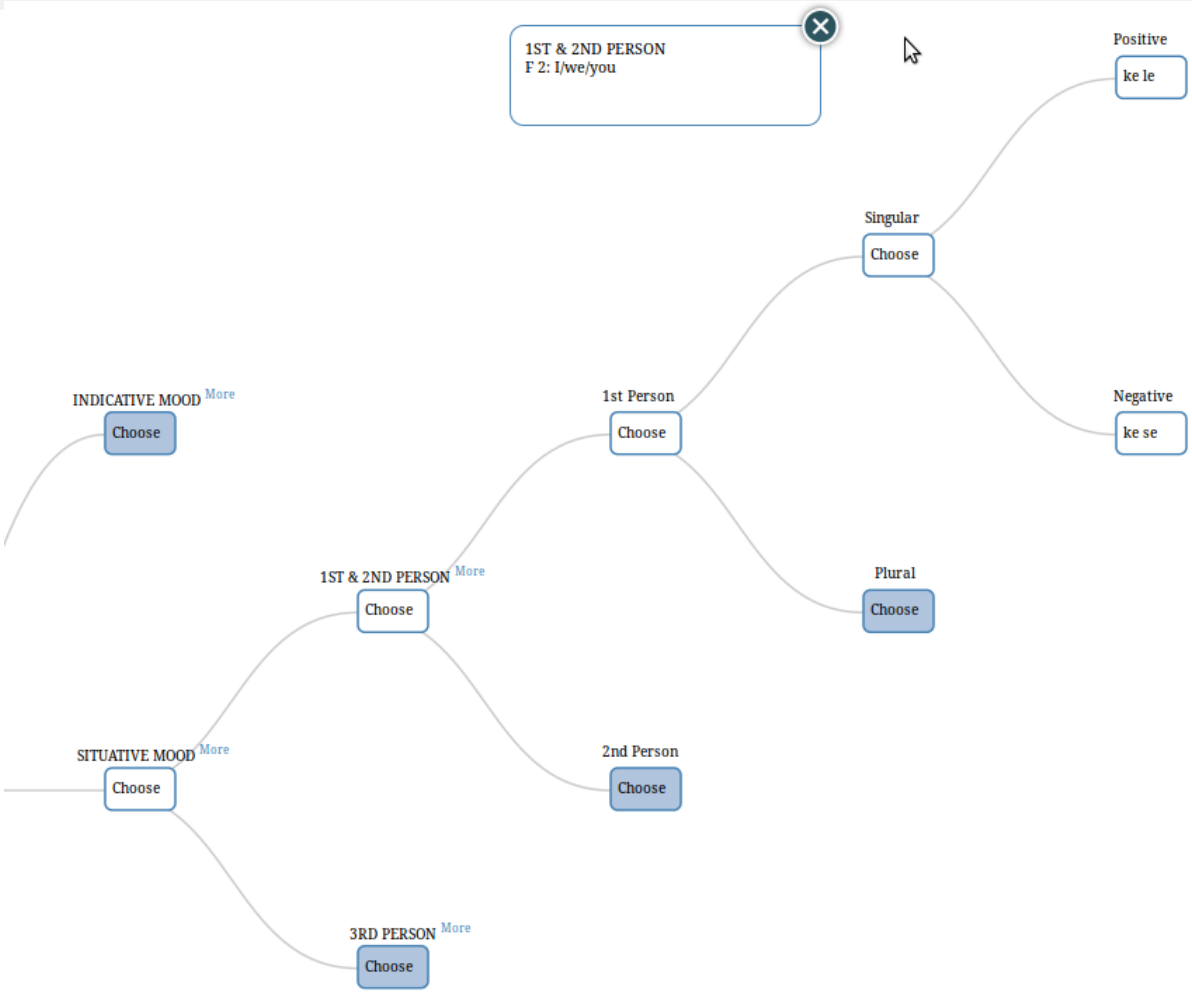
Example: Decision Tree

Production purposes: Experimental work by project team members



Example: Decision Tree

Production purposes: Experimental work by project team members



Open question: What to do for reception purpose?

Intro: What is a copula?

A simple account

- A copula links a subject with its complement(s)
- In **English**: "to be", i.e. *I am, you are, (s)he/it is, we are, ...*
- General: Possible verbal modifications[1]
  - person (1st/2nd/3rd)
  - number (sg/pl)
  - tense (non-past(present and future)/past)
  - aspect (simple/progressive/perfect/perfect progressive)
  - mood (indicative/imperative/emphatic/progressive/subjunctive)
- Leads to 3x2x3x4x5 = 360 possible constellations (of which a number are homographs)
- Polarity not specifically described
- *to have* (association) → not a copulative in English

[1] Origin of these definitions: Wikipedia: See list of urls on last slide.

Copula of Northern Sotho

For students

- **A Handbook of the Northern Sotho language**
  Ziervogel (1988:63): *"There are two kinds of copulatives, viz.*
  (a) the copulative of identification and
  (b) the copulative of description"
- Ziervogel does not refer to Lyons (1968), however Lyons had described these categories before:
  - Identifying copulative :
    Lyons (1968:389): *Apples are fruit*
    ('*sortal*')
  - Descriptive copulative :
    Lyons (1968:389): *Apples are sweet*
    ('*characterizing*')

Copula of Northern Sotho

For students

Northern Sotho for First-Years (Van Wyk et al. (1992:31)):

- The complement *"is always non-verbal ..."*
- *"There are three types* [...]
  identifying, descriptive and associative constructions"

The associative describes association, but also possession in the sense of "to be with" (e.g. another person, money, etc.)

O	*na*	*le*	*tšhelete*	*na?*
CSPERS_2sg	VCOP	PART_con	N09	PART_ques $.
You	are	with	money	hm?

Have you got money (with you)?

Copula of Northern Sotho

For scholars

A linguistic Analysis of Northern Sotho
(Poulos and Louwrens (1994:291 et seq.)):

(1) *The identifying copulative*

(2) *The descriptive copulative*

(3) *The assocative copulative*

(4) *The locational copulative*

N.B. The locational and descriptive copulas are morphologically identical; the distinction is based on the different nature of their complements.

Copula of Northern Sotho: Poulos and Louwrens' System

Modifications

- Copulative categories (identifying/descriptive/associative):
  - polarity (pos/neg)
  - 1st and 2nd person in singular and plural
  - classes (altogether 13)
  - present (principal/participial)
  - future (principal)
  - past (principal/participial)
  - potential, subjunctive, consecutive, habitual
  - infinitive, imperative
  - The descriptive[1] and the associative copulatives[2] have a "compound tense".
- No classification into tense/aspect/mood
- Poulos and Louwrens describe 1,328 possible constellations

[1] p. 311: *Diaparô di bê di le mêêtse – The clothes were wet.*

[2] p. 315: *Ke bê ke na le ntlô ka gê bê ke šoma ka maatla matšatšing ao – I had a house because I used to work hard in those days.*

Copula of Northern Sotho

For lexicographers

- **The Lemmatization of Copulatives in Northern Sotho**
  (Prinsloo (2002:28))
- *"two types of copulatives can be distinguished, namely static (in a state of rest) and dynamic (in motion or changing)"*
- *"copulatives express three different semantic relations between a subject and a complement, namely identification/equality, descriptive or associative"*
- Prinsloo estimates there are 2,040 different possible constellations, for the dynamic copulative only, including the potential forms which we have not included yet.

N.B. Lombard (1985:192 et seq.) describes the three categories: "identifying" and "descriptive" and "associative" in a similar way

Copula of Northern Sotho - Terminology

Differentiation between stative and inchoative

Lyons (1968:389):

- static copulative: *John has a book*
- dynamic copulative: *The book became valuable*

N.B.: In this presentation,
we refer to the static form of the copula as "stative"
and to the dynamic form as "inchoative"

Copula of Northern Sotho

For (computational) linguists

- **A morpho-syntactic description of Northern Sotho as a basis for an automated translation from Northern Sotho into English** (Faaß (2010:125 et seq.))
- An attempt to describe all possible constellations,
  with the exception of potential forms,
  relying on Prinsloo (2002)
  and Poulos and Louwrens (1994) –
  however, restricted for space reasons (similar to Poulos and Louwrens)

Copula of Northern Sotho

Constellations based on Faaß (2010:128, Table 3.30)

Copulative	Identifying		Descriptive		Associative	
Category	stative	inchoative	stative	inchoative	stative	inchoative
Tense						
pres	x	x	x	x	x	x
past	x	x	x	x	x	x
fut		x		x		x
Mood/Aspect						
indicative	x	x	x	x	x	x
situative	x	x	x	x	x	x
relative	x	x	x	x	x	x
consecutive		x		x		x
habitual		x		x		x
infinitive		x		x		x
imperative		x		x		x

Copula of Northern Sotho

Other categories

- person
- number (only for person)
- class
- polarity

  Our table currently contains 2,116 constellations

  (929 types; thus: many homographs!)

Reception? How to extract corpus examples?

- Very problematic from the start:
  - Faaß et al. (2009): many homographs (syncretism on the orthographic level): e.g. *a* is 8-ways ambiguous
  - Lombard (1985): the categories were mainly described on semantic and only partially on morpho-syntactic grounds
- No training data for statistical analysis (yet) available
  $\rightarrow$ manual inspection necessary

Searching corpora for copulatives

- *P*retoria *S*epedi *C*orpus, PSC (De Schryver and Prinsloo (2000))
- Current size: 8,007,653 tokens (including punctuation), sources/contents not defined exactly
- Part-of-speech tagged (cf. Taljard et al. (2008), Faaß et al. (2009)) and encoded in CorpusWorkBench (CWB, see link list)
- CWB allows for automated (offline) queries by means of scripts (e.g. perl) and macros

Copulative constellations in detail

Homography: *o tlo ba* as a case in point

- *o*:
  - subject concord of class 1, 3
  - subject concord of 2nd person singular
  - object concord of class 3
- *tlo*
  - future tense morpheme (exchangable with *tla*)
- *ba*
  - subject, object, and possessive concord of class 2
  - demonstrative of class 2
  - auxiliary and copulative verb stem

N.B. Heuristic taggers select the most frequent part of speech occuring in the training data → unreliable for such homographs while words with only one part of speech or with few differences in their distribution are easily identified and usually tagged correctly.

(cf. Faaß et al. (2009))

o tlo ba as a case in point

Excerpt of the overview of all constellations

no.	copulative	motion	tense	mood	polarity	pers/class
1	identifying	inchoative	future	indicative	positive	2nd pers.sg
2	identifying	inchoative	future	situative	positive	2nd pers.sg
3	descriptive	inchoative	future	indicative	positive	2nd pers.sg
4	descriptive	inchoative	future	indicative	positive	class 01
5	descriptive	inchoative	future	indicative	positive	class 03

Cases 1-2 $\rightarrow$ underspecification: indicative vs. situative
Cases 3-5 $\rightarrow$ homography *o* for 2nd.person.sg/classes 01 and 03
Cases 1-2/3-5 $\rightarrow$ underspecification: identifying/descriptive

o tlo ba may also precede a verb stem as part of a transitive future tense verb, where *ba* stands for an omitted or moved object noun of class 2

Task defition

- **Complements** should be identified:
  Identifying typical complements or complement types might help to differentiate not only verbs from copulative constellations,
  but underspecified constellations, too.
- **Subjects** should be identified:
  Identifying a copulative's subject will help to avoid disambiguation problems caused by homography (concordial agreement).

Nominal complements: Definition of a few constellations

- A typical noun chunk may consist of a noun alone
- This noun might be accompanied by
  - A demonstrative (possibly followed by an adjective)
  - An emphatic pronoun
  - A quantitative pronoun
  - A possessive concord follwed by a possessive pronoun or another noun chunk
- Each of the accompanying units or unit groups might appear alone as well
- ...

N.B. This is no exhaustive description,
see e.g. Faaß (2010:175 et seq.) for a (hopefully) complete overview

Corpus Queries

Method: Steps of the search procedure

- One macro for all:
  - nominal complements → constants (defined by their parts of speech)
  - copula → variables (defined as tokens)
- Execute the macros (= run the query)
  with each of the 929 copula types in CWB
  (making use of the perl interface)
- Extend the table of constellations with the frequencies of occurrences found in the corpus
- Generate a table containing all matches found in the corpus
  (copula and complement)

Results: a first attempt

Is it a copulative at all?

- We randomly chose 200 constellations found by the tool:
- Results:
  - 187 were correctly identified as copulatives
  - 13 incorrect: corpus errors, annotation problems, minor macro errors

→ Our general complement definitions (noun chunks) are correct!

Results: associative copulative

- 1,203 constellations (599 types) found with a nominal chunk as a complement
- Homography of single items (e.g. *a*, *o*, etc.)
- Underspecification (e.g. indicative/situative constellations)
- However, the forms do not occur in the other types of copulatives (identifying/descriptive)

Results for *o tlo ba*

- Frequency of occurrence in total: 364
- Frequency of occurrence followed by a noun chunk (as described above): 40
- Frequency of occurrence followed and preceded by a noun chunk: 33
- Manual inspection of the 33 sentences:
  - 17 identifying copulative:
  - 11 descriptive copulatives with complements of a specific type (see next slide)
  - Preceeding noun chunk is usually not the subject $\rightarrow$ we need a grammar
  - Following noun chunk is usually the object and the descriptives seem tot be distinguishable from the identifying by their morphosyntactic properties
  - 7 others: problems of corpus preparation/cases where semantics are not consistent with morphological features

Finding typical complements

Ongoing work

- Identify descriptive copulatives
  - When inspecting overall results, typical complements were identified: e.g. nouns with a locative ending (see *locational constellations*, cf. Poulos and Louwrens (1994) above)
  - Nouns and pronouns and demonstratives of class 14
- All the other homographous constellations found are currently assumed to be of an identifying character (verification outstanding)

Frequencies of occurrences

Data for copula: "ra ba le: number of matches: 2
ra ba le omiša phesente ye kgolwane ya maatla a monagano wa rena go feta ka fao go tlwaelegilego mme ka gona <ra ba le bokgoni> bja go phetha dilo tša senama ka mmele wa rena tšeo di sa kgonegego ka fase ga mabaka ao a tlwael
ra ba le olele ge a swarišitši mahlo a lephodisa bobi gore le se mo šetše ? Aowa tšeo re ka se di tsebe <ra ba le bohlatse> natšo . Se re se lemogago ka monna yo ke gore e be e le moretšhe wa leoma mahlwa a di bona . Tsie

----

Data for copula: "ra se be: number of matches: 0

# of occ.	constellations	types
0	618	350
1 - 99	1,212	477
100 - 999	210	81
1,000 - 4,999	69	18
$>$ 5,000	7	3
sums	2,116	929

Overall results

Still tentative

- Associative forms can be differentiated from the other copulatives easily and 216 such constellations are not homographous at all
- Differentiation between identifying and descriptive copulatives might be possible by complement definition of the descriptive forms (verification outstanding)
- Outstanding: Differentation between situative and identifying copulatives
- However, for lexicographic reception purposes:
  Distinguishing these constellations is not necessary for translation, rather worth a linguistic study

Future work

- Add potential forms of the copulatives to our table,
  make it an accessible database
- Examine the constellations not found in the corpus: too rarely used for complexity reasons
  or just described by linguists to fill the paradigm?
- From underspecification to specification $\rightarrow$
  Write a little grammar so that the homographs
  can be disambiguated – at least partially
- For lexicography: If typical complements are known, we can provide typical examples for text production
- General task: Work towards a cleaner corpus

References

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

Faaß et al.(2009). G. Faaß, U. Heid, E. Taljard, and D.J. Prinsloo. 2009. Part-of-Speech tagging in Northern Sotho: disambiguating polysemous function words. In *Proceedings of the EACL2009 Workshop on Language Technologies for African languages – AfLaT 2009*. 38 – 45. The 12th Conference of the European Chapter of the Association for Computational Linguistics; Mar 30 - April 3rd, 2009. Athens.

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

Louwrens (1991). L.J. Louwrens. 1991. *Aspects of the Northern Sotho Grammar*. Pretoria: via Afrika.

Poulos and Louwrens (1994). G. Poulos and L.J. Louwrens. 1994. *A Linguistic Analysis of Northern Sotho*. Pretoria: via Afrika.

Prinsloo (2000). D.J. Prinsloo. 2002. The Lemmatization of Copulatives in Northern Sotho. In *Lexikos 12*, 21 – 43. Stellenbosch: Buro van die WAT.

Taljard et al. (2008). E. Taljard, G. Faaß, U. Heid, and D.J. Prinsloo. 2008. On the development of a tagset for Northern Sotho with special reference to the issue of standardization. *Literator – special edition on Human Language Technology*, 29(1):111 – 137.

Van Wyk et al. (1992). E.B. Van Wyk, P.S. Groenewald, D.J. Prinsloo, J.H.M. Kock, and E. Taljard. 1992. *Northern Sotho for first years*. Pretoria:J.L. van Schaik.

Ziervogel (1998). D. Ziervogel. 3rd edition 1988. *A Handbook of the Northern Sotho Language*. Pretoria:J.L. van Schaik.

link list

Scientific e-Lexicography for Africa (SeLA):

http://www.uni-hildesheim.de/iwist-cl/projects/sela/

Permanent links from wikipedia:

(1) tense: http://en.wikipedia.org/w/index.php?title=Grammatical_tense&oldid=555002677

(2) aspect: http://en.wikipedia.org/w/index.php?title=Grammatical_aspect&oldid=551547626:

(3) mood: http://en.wikipedia.org/w/index.php?title=Grammatical_mood&oldid=542114879

CentLex:

http://bcom.au.dk/research/academicareas/centreforlexicography/research/

Corpus WorkBench:

http://sourceforge.net/projects/cwb/?source=directory