German Word Sketches and prepositional phrases in German

Matej Durco

ICLTT, Vienna

Peter Durco

Univerzita Cyrila a Metoda, Trnava

2012-03-21

Abstract

Description of the current situation for German Word Sketches (with sketch grammar based on PoS-tags from RF-tagger) and their use for identifying and describing prepositional phrases

---

1. Previous work

- two attempts to define Word Sketches for German:
  - Ivanova et al. LREC, 2008

    Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case

    corpus: 100 mio. deWaC (part of a bigger web corpus)
  - Durco, Durco, Benko - experimental

    corpus: 20 mio. Gutenberg
- both based on `STTS`
- main finding:

  Tagset is too coarse-grained - not providing gender, number or case

  => not possible to distinguish between subject/object position of the noun.

2. RFTagger - New hope?

- RFTagger from Stuttgart [Schmid, Laws; COLING 2008]

  a successor of TreeTagger

building on top of STTS with more fine-grained PoS-tags

```
word      part of speech
Das       PRO.Dem.Subst.-3.Nom.Sg.Neut
ist       VFIN.Sein.3.Sg.Pres.Ind
ein       ART.Indef.Nom.Sg.Masc
Testsatz  N.Reg.Nom.Sg.Masc
.         SYM.Pun.Sent
```

RF-Tagger german tagset

3. Current sketch grammar for German

- further development of the [durco,durco2009] grammar
- makes use of the richer tagset produced by the RF-Tagger
- based on a generic combinatory matrix of the keyword and the collocate wrt to certain grammatical categories (especially Casus - for example: `Attr + Sub1Nom, Sub1Dat + Verb`)

  introduced in [P. Durco (2007): Zum Konzept eines zweisprachigen Kollokationswörterbuchs - Prinzipien der Erstellung]

  proposing to consider collocations as constituted (mostly) by word forms rather than lemmas (restricted syntactic paradigma)

  (cf. Kollokationen_Durco_SK.pdf starting on page 13 are proposed the templates for individual word classes)
- 78 rules (29 UNARY)
- now applied on `deTenTen` (2.3 bio words)

3.1. Datasets used

Table 1. Datasets within sketch engine, that various versions of GWS were applied to

name	size (~ mio. tokens)	tagset	Grammar	rules
dewac100	90	STTS	ivanova2008	12
gutenberg1	20	STTS	durco,durco2009	31
bigdewac20	17	RF, STTS	DE-RF v2a	86
bigdewac100	75	RF, STTS	DE-RF v2b	86
deTenTen	2.330	RF, STTS	DE-RF v3d	70
deTenTen50	50	RF, STTS	DE-RF v3.5	78

3.2. Naming convention

- the keyword is marked **X** and the collocate **Y**, eg:

```
    =SubstXNom+VerbY (subj_of)   => "1:Baum 2:wächst"
vs. =SubstYNom+VerbX (has_subj)  => "2:Baum 1:wächst"
```

- If the rule is matching in both directions (irrespective of the order of the collocation partners) the two parts are infixed with **+** sign:

```
=SubstXNom+VerbY => "Hier 2:befindet sich das 1:Haus", "Das 1:Haus 2:befindet sic
=PräpY SubstXDat => "2:auf der 1:Seite"
```

- Where possible also the grammatical relation is added in brackets:

```
=SubstYNom+VerbX (has_subj)
=VerbY + %s SubstXDat (pp_obj_of_dat)
```

3.3. Structure

- 78 Rules for (and grouped by) **nouns, verbs, adjectives, adverbs, adpositions**
- 29 **UNARY** relations checking individual grammatical information, eg:

```
= Casus_Dat
= Casus_Acc
= Ambiguous_Casus
```

- 9 **TRINARY** relations summarized into two groups: **Subst+Verb** and **Subst+Subst**

```
*SEPARATEPAGE Subst+Verb
*TRINARY
= VerbY + %s SubstXDat (pp_obj_of_dat)
```

`*TRINARY` directive allows to catch relations of three elements

- almost no ***DUAL** rules - just for the sake of order in the grammar (i.e. potential duals were split in two rules in respective word classes)
- defined gaps

```
define(`gap',` tag!="SYM.*" & word!="und" & word!="oder" ')
define(`gap_nN',`gap & tag!="N.*"')
```

Gaps are used in the rules to individually specify the size of the window to consider between the collocates, still respecting sentence/clause boundaries (`tag!="SYM.*"...`):

```
2:[adj_] [gap_nN]{0,4} 1:"N.*"
     => "die 2:alten knorrigen 1:Bäume"
1:"VFIN.Full.*" [gap_nV]* 2:"N.*Nom.*"
     => "1:fliegen dann auch ordentlich die 2:Fetzen"
```

3.4. Structure - Matrix

Table 2. combinatory matrix

	Subst	**Verb (Präp)**	**Präp**
Subst Nominativ		`SubstNom+Verb` `(has_subj/subj_of)`	

Subst Genitiv	`Subst+SubstGen`	`Verb+SubstGen`
		`(obj_of/has_obj_gen)`
		
		`Verb + %s SubstGen (!)`
		`(pp_obj_of/has_pp_obj_gen)`
			`Präp SubstGen`
			
			`SubstGen Präp`

Subst Dativ	`Subst+SubstDat`	`Verb+SubstDat`
		`(obj_of/has_obj_dat)`
		
		`Verb + %s SubstDat`
		`(pp_obj_of/has_pp_obj_dat)`
			`Präp SubstDat`
			
			`SubstDat Präp`

Subst Accusativ	`Subst+SubstAcc`	`Verb+SubstAcc`
		`(obj_of/has_obj_acc)`
		
		`Verb + %s SubstAcc`
		`(obj_of/has_pp_obj_acc)`
			`Präp SubstAcc`
			
			`SubstAcc Präp`

mostly "`DUAL`" (i.e. `Präp SubstGen` resolves to `PräpX SubstYGen, PräpY SubstXGen`)

3.5. Rules for prepositions

- before nouns

```
=PräpX SubstYDat  => "1:Auf dem weiteren 2:Weg"
```

- after verbs

```
=VerbY PräpX  => "2:beziehen sich nicht 1:auf Marketplace"
```

- between nouns and nouns (TRINARY)

```
=SubstY %s SubstX  => "neuartige 2:Interaktionen 3:zwischen 1:Mensch und Roboter"
```

```
=SubstX %s SubstY  => "die beiden ersten 1:Menschen 3:auf dem 2:Mond"
```

- between verbs and nouns (pp_obj) (TRINARY)

```
=VerbX %s SubstYGen (has_pp_obj_gen) => "1:begann 3:während des 2:Medizinstudiums
=VerbY %s SubstXAcc (pp_obj_of_acc) => "und 2:prallt 3:gegen einen 1:Baum."
```

- after adjective

```
=AdjY PräpX => "Seine Mutter war 2:stolz 1:auf ihn gewesen"
```

- before adverb

```
=PräpX AdvY => "1:Auf 2:einmal machte etwas Klick", "1:auf sich 2:selbst gestellt
```

4. Examples

- word sketch for "auf" in deTenTen with grammar v3d (cache)
- word sketch for "auf" in deTenTen (50) with grammar v3.5 (cache)
- auf/in diff (cache)

```
PräpX SubstYAcc => 1:auf diese 2:Weise
PräpX SubstYDat => 1:in dieser/einer/keinsten... 2:Weise
```

- Kopf

```
{Subst} im Kopf {Verb}
Kopf und Kragen
```

- lemma!

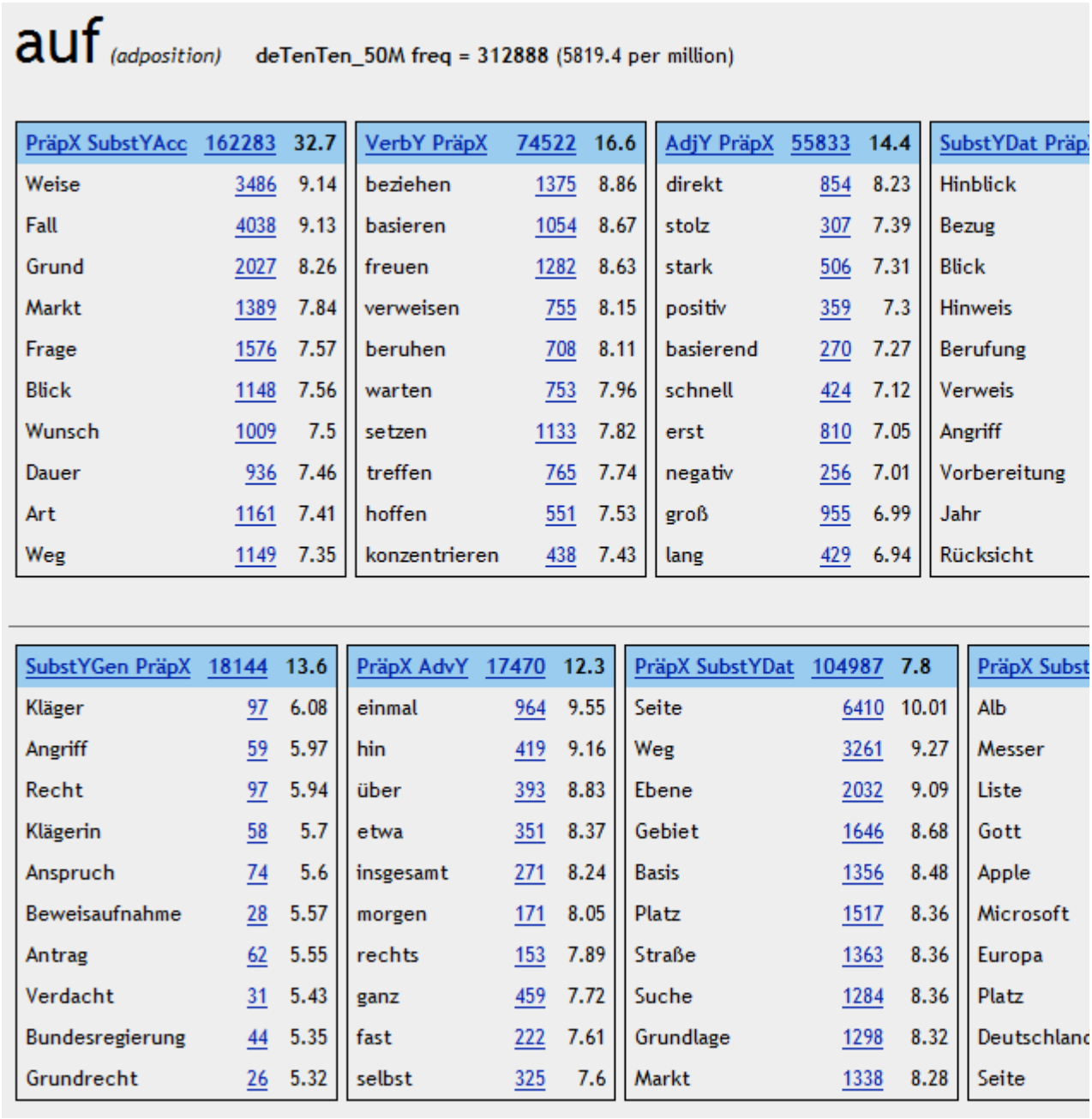
```
2:Stein der 1:Weisen (philosopher's stone)
2:auf diese 1:Weise (in this manner)
```

- binding!

```
=SubstYGen PräpX
  2:"N.*Gen.*" [gap_nN]{0,4} 1:"AP.*"
  => Der ->2:Antrag des !2:Klägers 1:auf Wiederaufnahme
```

4.1. Word Sketch for "auf"

Figure 1. Word Sketch for "auf" in deTenTen(50) with grammar version 3.5



auf *(adposition)* deTenTen_50M freq = 312888 (5819.4 per million)

PräpX SubstYAcc	162283	32.7
Weise	3486	9.14
Fall	4038	9.13
Grund	2027	8.26
Markt	1389	7.84
Frage	1576	7.57
Blick	1148	7.56
Wunsch	1009	7.5
Dauer	936	7.46
Art	1161	7.41
Weg	1149	7.35

VerbY PräpX	74522	16.6
beziehen	1375	8.86
basieren	1054	8.67
freuen	1282	8.63
verweisen	755	8.15
beruhen	708	8.11
warten	753	7.96
setzen	1133	7.82
treffen	765	7.74
hoffen	551	7.53
konzentrieren	438	7.43

AdjY PräpX	55833	14.4
direkt	854	8.23
stolz	307	7.39
stark	506	7.31
positiv	359	7.3
basierend	270	7.27
schnell	424	7.12
erst	810	7.05
negativ	256	7.01
groß	955	6.99
lang	429	6.94

SubstYDat Präp
Hinblick
Bezug
Blick
Hinweis
Berufung
Verweis
Angriff
Vorbereitung
Jahr
Rücksicht

SubstYGen PräpX	18144	13.6
Kläger	97	6.08
Angriff	59	5.97
Recht	97	5.94
Klägerin	58	5.7
Anspruch	74	5.6
Beweisaufnahme	28	5.57
Antrag	62	5.55
Verdacht	31	5.43
Bundesregierung	44	5.35
Grundrecht	26	5.32

PräpX AdvY	17470	12.3
einmal	964	9.55
hin	419	9.16
über	393	8.83
etwa	351	8.37
insgesamt	271	8.24
morgen	171	8.05
rechts	153	7.89
ganz	459	7.72
fast	222	7.61
selbst	325	7.6

PräpX SubstYDat	104987	7.8
Seite	6410	10.01
Weg	3261	9.27
Ebene	2032	9.09
Gebiet	1646	8.68
Basis	1356	8.48
Platz	1517	8.36
Straße	1363	8.36
Suche	1284	8.36
Grundlage	1298	8.32
Markt	1338	8.28

PräpX Subst
Alb
Messer
Liste
Gott
Apple
Microsoft
Europa
Platz
Deutschland
Seite

5. Related Projects

- DWDS Wortprofile

  [Geyken, A. (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In:Abel, Andrea/Zanin, Renata (Hg.). Korpora in Lehre und Forschung. Bozen/Bolzano, S. 115-137.]

  source: DWDS (~500 mio.) parsed with SynCoP [Didakowski (2008)]. planned new version on 1.8 bio. tokens

- signifikante Wortpaare als Webservice

  [Fritzinger, F., Kisselew, M., Heid, U., Madsack, A., Schmid, H. (2009). Werkzeuge zur Extraktion von signifikantenWortpaaren als Web Service. Vortrag: GSCL Symposium Sprachtechnologie und eHumanities, Duisburg, 26-27 Februar 2009]

  dependency parser FSPAR [Schiehlen (2003)]

- (to be) printed Kollokationenwörterbuch, Uni Basel (colloc.germa.unibas.ch)

sources: DWDS, CHTK, (DEREKO, Leipzig Corpora)

- [Peter Durco: project on slovak and german-slovak collocations dictionary]

  Example available at: http://vronk.net/wicol/index.php/Zeit

  sources: DWDS, DEREKO, Leipzig Corpora, (deTenTen)

- Wortverbindungsfelder (part of project Usuelle Worverbindungen at IDS-Mannheim)

  sources: DEREKO

- planned new project: Präpositionale Wortverbindungen (collocational PPs)

  Example: http://vronk.net/wicol/index.php/Auf

  sources: DWDS, DEREKO, deTenTen

6. Open issues

6.1. Quality of the tagging

- one approach: use "debugging"-grammar rules (queries)

```
=PräpY SubstXNom
  2:"APPR.*"  [gap_nN]{0,2} 1:"N.*Nom.*"
  => Gegengewichte aus/APPR.Dat/APPR Blei/N.Reg.Nom.Sg.Neut/NN
=Ambiguous_Casus
1:"N.*\*2.*" => "Seit Ende/N.Reg.*2.Sg.Neut/NN August",
                "um fast neun Prozent/N.Reg.*2.*3.Neut/NN"
```

- The double-tagged corpora allow for easy tagger-evaluation. With special queries (on both tagsets) we can find tag-disagreements.

  Here is a comparison of basic PoS-assignement (N, Adja, Vfin, V) between TreeTagger and RF-Tagger on the `BiGDeWaC20` and `deTenTen` corpus. It shows the ratio of disagreement between the two taggers:

	BigDeWaC20		deTenTen(50)	
type	**in RF, NOT in TT**	**in TT, NOT in RF**	**in RF, NOT in TT**	**in TT, NOT in RF**
Nouns [N*]	8,6 %	7,7 %	3,3 %	5,6 %
Adjectives [ADJ*]	13,7 %	21,5 %	9,5 %	9,7 %
finite Verb [Verb.fin.full] [VFIN]	11,2 %	38,9 %	10,3 %	15,3 %
Verb [V*]	1,9 %	16,7 %	2,1 %	3,7 %

  reasons mostly: hyphenation, compounds, special symbols

6.2. Postprocessing - Output

- lot of manual work

  => need for integration with the target (lexicographic?) environment

- requirement: bring together and compare different sources

  prototypical solution: use the webservice to get the data and transform and match data from different sources (example) This is similar to sketch diffs, but matching only on collocates (ignoring/inverting the gramrels structuring level) and allowing more than two datasets.

6.3. Other

- possible use of parsed text-data?
- in web corpora subcorpus definition by TLD (`.de, .at, .ch`)?
- the right size of the rule "window" (gap)?

- bookmark queries?
- bug? `*SEPARATEPAGE *TRINARY` relations (with brackets in name?) fail to show up:

  VerbY + gegen-i SubstXAcc (pp_obj_of_acc)

  (also when retrieving the sketches via the web services, the result is cut off at first such relation)

6.4. Planned changes in the grammar

- reduce `SubstY{casus} PräpX` to `SubstY PräpX`

  because the casus usually pertains to the previous relation

```
SubstYDat PräpX => "nach dem 2:Hinweis 1:auf"
SubstYAcc PräpX => "ohne einen 2:Hinweis 1:auf"
```

- + exclude verb in the gap `SubstY PräpX`

```
Im selben 2:Jahr !wurde 1:auf
```

- handle (remove or better restrict) `PräpX SubstYGen`

  because preposition usually binds the next noun

```
1:auf !2:Berlins ->2:Straßen
          1:auf !2:Messers ->2:Schneide
BUT: 1:wegen 2:Verstoßes gegen, 1:wegen versuchten 2:Mordes
```