Universität Zürich[UZH]

Institut für Computerlinguistik

Incremental Morphosyntactic Disambiguation of Nouns in German-Language Law Texts

ESSLLI-13 Workshop on Extrinsic Parse Improvement (EPI)

Kyoko Sugisaki and Stefan Höfler

Background and Motivation

Aim:

To develop a German **style checker** for law texts

Task:

For the reliable detection of the violations of **syntax-related style rules**, existing parsers have to be adopted to the domain.

Current situation:

Lack of a large annotated corpus

Our approach:

Hybrid approach for the recognition of grammatical functions

Overview

- Introduction
- Morphosyntactic disambiguation of German nouns for the recognition of grammatical functions
- Evaluation
- Conclusion

Recognition of Grammatical Functions for German

The mapping of case markings and grammatical functions is straightforward (e.g. dative case marking = indirect object)

Challenge:

(1) Morphosyntactic ambiguity in German:

(A)

Das Amt erteilt die Bewilligung.

NOM or ACC NOM or ACC

(B)

Die Bewilligung erteilt das Amt.

NOM or ACC NOM or ACC

„The authority accords the permission“

Hard constraints

Soft constraints

(2) Relatively free word order

NOM >> ACC

ACC >> NOM

Case-feature Disambiguation for the Recognition of Grammatical Functions

Step 1: Hard Constraints

- ➢ Agreement, argument structures, voice, etc.
- ➔ Morphosyntactic ambiguity reduction of nouns
- ➔ Rule-based approach (Constraint Grammar)

How far can linguistically motivated hard constraints reduce morphosyntactic ambiguity before any soft constraint is applied?

Step 2: Soft Constraints

- ➢ Word order, definiteness etc.
- ➔ Morphosyntactic ambiguity resolution of nouns

Task: Morphosyntactic Disambiguation

Input: Outputs from Gertwol (morphological analyzer)

"Mitarbeitenden“: ‚*co-worker*‘

SELECT ➔ N(PART) POS SG ACC MASC

REMOVE

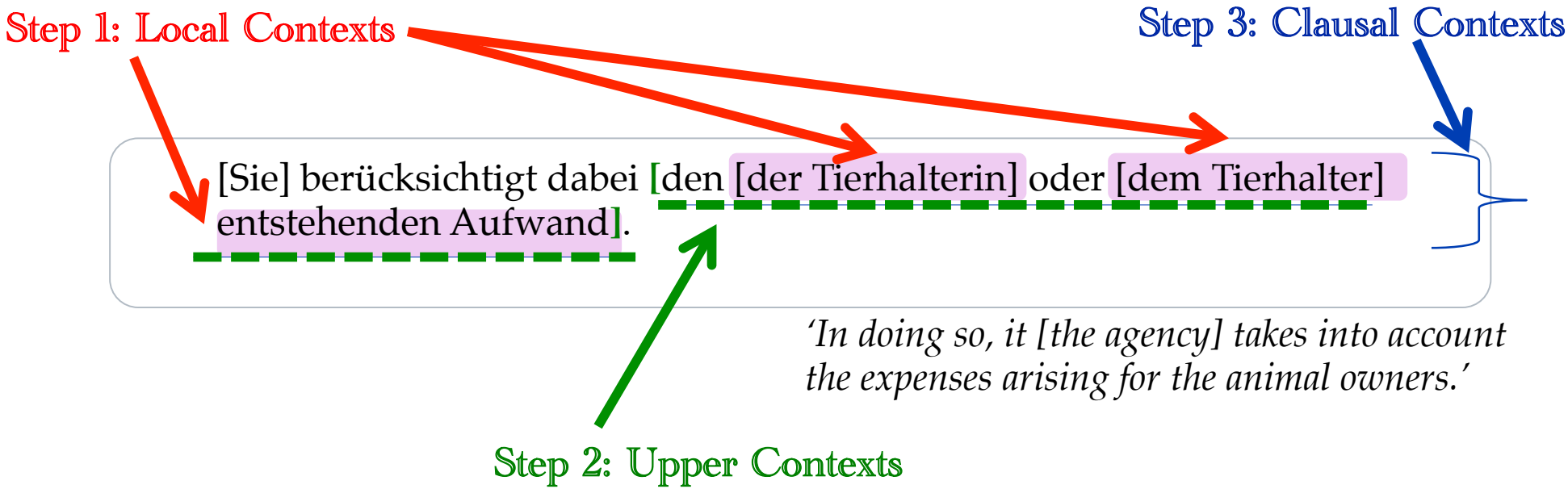
; N(PART) POS SG GEN MASC
; N(PART) POS SG GEN NEUTR
; N(PART) POS PL DAT
; N(PART) POS SG DAT MASC
; N(PART) POS SG DAT NEUTR
; N(PART) POS SG DAT FEM
; N(PART) POS SG GEN FEM
; N(PART) POS PL NOM
; N(PART) POS PL ACC
; N(PART) POS PL GEN

➔ **Optimal output: one morphosyntactic analysis per token**

Morphosyntactic Disambiguation of Nouns

Incremental 3-Step Disambiguation using hard constraints:

- Step 1: **Local** phrase-level feature **unification**
- Step 2: **Upper** phrase-level feature **unification**
- Step 3: **Clause**-level feature **unification**



Step 1: Local Phrase-level Feature Unification

Morphosyntactic feature unification in simple noun phrases:

- Agreement: number, gender and case

[Sie] berücksichtigt dabei [den [der Tierhalterin] oder [dem Tierhalter] entstehenden Aufwand].

'In doing so, it [the agency] takes into account the expenses arising for the animal owners.'

der: *'the'*
; ART DEF SG NOM MASC
ART DEF SG DAT FEM
ART DEF SG GEN FEM
; ART DEF PL GEN
; PRON DEM ...
; PRON RELAT ...

Tierhalterin: *'animal owner(fem)'*
; N FEM SG NOM
; N FEM SG ACC
N FEM SG DAT
N FEM SG GEN

Ambiguity reduction: 4 case features ➔ 2

dem: (*'the'*)
ART DEF SG DAT MASC
; ART DEF SG DAT NEUT
; PRON DEM ...
; PRON RELAT ...

Tierhalter: *'animal owner(masc)'*
; N MASC SG NOM
; N MASC SG ACC
N MASC SG DAT
; N MASC SG GEN

Ambiguity resolution: 4 case features ➔ 1 (DAT)

Step 2: Upper Phrase-level Feature Unification

Morphosyntactic feature unification in complex NPs und PPs:

- Agreement: NP coordination, participle phrases, prepositional phrases

[Sie] berücksichtigt dabei [den [der Tierhalterin] oder [dem Tierhalter] entstehenden Aufwand].

'In doing so, it [the agency] takes into account the expenses arising for the animal owners.'

Tierhalterin: *‚animal owner(fem)‘*
; N FEM SG NOM
; N FEM SG ACC
N FEM SG DAT
; N FEM SG GEN

Tierhalter: *‚animal owner(masc)‘*
; N MASC SG NOM
; N MASC SG ACC
N MASC SG DAT
; N MASC SG GEN

Ambiguity resolution: 2 case features ➔ 1 (DAT)

den: *‚the‘*
ART DEF SG ACC MASC
; ART DEF PL DAT
; PRON DEM ...
; PRON RELAT ...

Aufwand: *‚expense‘*
; N MASC SG NOM
; N MAC SG DAT
N MASC SG ACC

Ambiguity resolution: 4 case features ➔ 1 (ACC)

Step 3: Clause-level Feature Unification

Morphosyntactic feature unification of NPs in a clause:

- Subject-verb agreement, argument structure, voice, etc.
  - ➔ *Every clause has only one subject*
  - ➔ *Subject agrees with the finite verb*

[Sie] berücksichtigt dabei [den [der Tierhalterin] oder [dem Tierhalter] entstehenden Aufwand].

'In doing so, it [the agency] takes into account the expenses arising for the animal owners.'

Sie: *'it'*
PRON PERS SG3 NOM FEM
; PRON PERS SG3 ACC FEM
; PRON PERS PL3 NOM
; PRON PERS PL3 ACC

Aufwand: *'expense'*
; N MASC SG NOM
; N MAC SG DAT
N MASC SG ACC

Ambiguity resolution: 4 case features ➔ 1 (NOM/SG)

Evaluation: Test Data and Performance

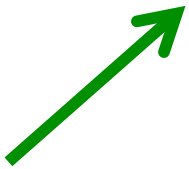
- Test data: 118 sentences (2,114 Tokens, incl. 655 nouns and pronouns) from the Swiss Legislation Corpus.
- Results:
  - (a) 96.30% (Recall): correct morphosyntactic analysis found by the system relative to the gold standard
  - (b) 67.60% (Precision): correct morphosyntactic analysis found by the system relative to the total number of system outputs

Evaluation (a)	Evaluation (b)
Tiehalterin: (NOM)	Tierhalterin: (NOM)
N FEM SG NOM	N FEM SG NOM“
N S FEM SG ACC	N S FEM SG ACC“
; N FEM SG DAT	; N FEM SG DAT
; S FEM SG GEN	; N FEM SG GEN

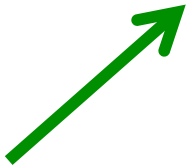
Data Analysis: 3-Step Disambiguation

System data: 239 sentences (4,789 tokens: 1,668 nouns/pronouns) from the Swiss Legislation Corpus.

Steps	1 Analysis/Token	2+ Analyses/Token
Input from Gertwol	148 (8.87%)	1,520 (91.13%)
After 1st step: Local phrase-level feature unification	387 (23.20%)	1,281 (76.80%)
After 2nd step: Upper phrase-level feature unification	917 (54.98%)	751 (45.02%)
After 3rd step: Clause-level feature unification	1,129 (67.69%)	539 (32.31%)



Completely disambiguated after 3-step disambiguation

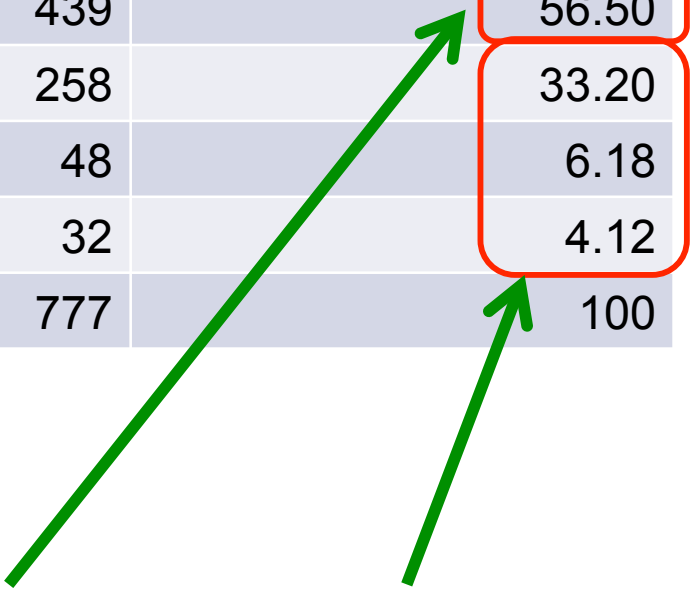


Not yet disambiguated after 3-step disambiguation

Data Analysis: Disambiguation of Case Features for GF Candidates

System data: 239 sentences(4,789 tokens; 777 GF-candidates) from the Swiss Legislation Corpus.

	Tokens	%
1 case feature/token	439	56.50
2 case features/token	258	33.20
3 case features/token	48	6.18
4 case features/token	32	4.12
Total: GF candidates	777	100



Completely disambiguated after 3 steps

Not yet disambiguated after 3 steps

Summary & Future Work:

Summary:

- Morphosyntactic disambiguation of nouns using hard constraints:
  - 91.12% ➔ 32.31% (in test data)
- Morphosyntactic disamiguation of case features in test data:
  - disambiguated: 56.50%
  - Two casus-features: 33.20%

Future work:

- Morphosyntactic disambiguation using soft constraints

Acknowledgement

We thank

The Swiss National Foundation, Switzerland
Prof. Dr. Michael Hess, Institute of Computational Linguistics, University of Zurich
Prof. Dr. Felix Uhlmann, Institute of Law, University of Zurich
Dr. Rebekka Bratschi, Swiss Federal Chancellery

for their support of our project.

Our project "*Automated Detection of Style Guide Violations in Legislative Drafts*":
http://www.cl.uzh.ch/research/maschinellestilpruefung/gesetzestextanalyse_en.html

Bibliography

- Bundesamt für Justiz (2007). Gesetztesgebungsleitfaden: Leitfaden für die Ausarbeitung von Erlassen des Bundes. Bern, 3 edition.
- Hansen-Schirra, S. and Neumann, S. (2004). Lin- guistische Verständlichmachung in der juristis- chen Realität. In Lerch, K. D., editor, *Die Sprache des Rechts: Recht verstehen: Verstandlichkeit, Missverstandlichkeit und Unverstandlichkeit von Recht,* volume 1. Walter de Gruyter, Berlin.
- Hinrichs, E. W. and Trushkina, J. S. (2004). Forg- ing Agreement: Morphological Disambiguation of Noun Phrases. *Research on Language and Computation,* 2(4):621–648.
- Höfler, S. and Sugisaki, K. (2012). From Drafting Guideline to Error Detection: Automating Style Checking for Legislative Texts. In *EACL 2012: Proceedings of the Second Workshop on Com- putational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering,* pages 9– 18. Association for Computational Linguistics.
- Höfler, S. and Piotrowski, M. (2011). Building Corpora for the Philological Study of Swiss Legal Texts. Journal for Language Technology and Computational Linguistics (JLCL), 26(2).
- Karlsson, F., Voutilainen, A., Heikkila¨, J., and Anttila, A., editors (1995). *Constraint Grammar: A Language-Independent System for Parsing Un- restricted Text.* Mouton de Gruyter, Berlin/New York.
- Mariikka, H. and Majorin, A. (1994). GERTWOL: ein System zur automatischen Wortformerkennung deutscher Wörter. Technical report, Lingsoft, Inc.
- Nussbaumer, M. (2009). Rhetorisch-Stilistische Eigenschaften der Sprache des Rechtswesens. In Rhetorik und Stilistik / Rhetoric and Stylistics: Ein Internationales Handbuch Historischer und Systematischer Forschung/An International Handbook of Historical and Systematic Research, volume 2, pages 2132–2150. Mouton de Gruyter, Berlin/New York.
- Regierungsrat des Kantons Zürich (2005). Richtlinien der Rechtssetzung.
- Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop.,* Dublin, Ireland.