Universität Zürich[UZH]

Institut für Computerlinguistik

Morphosyntaktische Desambiguierung von Nomen in Schweizer Gesetzestexten

Kolloquium in Computerlinguistik FS 2013

Kyoko Sugisaki

Hintergrund

SNF-Projekt

"Automated detection of styleguide violations in legislative drafts"

- Forschungsziel: maschinelle Stilprüfung von Gesetzestexten

Mein PhD-Projekt

- Ziel: maschinelle Überprüfung von syntaktischem Stil für Gesetztestexte

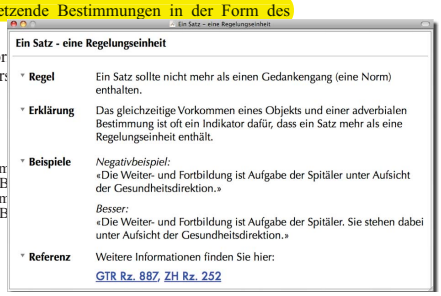
Art. 163 Form der Erlasse der Bundesversammlung

[1] Die Bundesversammlung erlässt rechtsetzende Bestimmungen in der Form des Bundesgesetzes oder der Verordnung.

[2] Die übrigen Erlasse ergehen in der For
beschluss, der dem Referendum nicht unter
bezeichnet.

83 Angenommen in der Volksabstimmung von
BRB vom 4. Febr. 2002 – AS **2002** 241; BB
84 Angenommen in der Volksabstimmung von
BRB vom 4. Febr. 2002 – AS **2002** 241; BB

50



Inhalt

Einleitung

Hauptteil

- Vorverarbeitung: Erkennung der topologischer Felder
- Desambiguierung morphosyntaktischer Ambiguität von Nomen
- Evaluation

Zusammenfassung und Ausblick

Style Guide: syntaxbezogene Regeln

Wortstellungsregel:

"Die Reihenfolge der Satzteile soll dem üblichen Sprachgebrauch entsprechen, also in der Regel Subjekt → Prädikat → Ergänzung (Dativ-, Akkusativ- und Genitivobjekte, andere Ergänzungen)."

Subjekt vorne

(A) Das Amt erteilt die Bewilligung.

Subjekt nach Objekt

(B) Die Bewilligung erteilt das Amt.

Herausforderung: Korrektes Erkennen grammatischer Funktionen

Morphosyntaktische Ambiguität:

Hard constraints

Soft constraints

Relativ freie Wortstellung

NOM >> AKK

AKK>>NOM

Belebtheit

ORGANIZATION = NOM

ORGANIZATION = NOM

Seite 5

Kasus-Feature Desambiguierung als Ansatz für die Erkennung der grammatischen Funktionen

Step 1: Hard Constraints

- Kongruenzen, Argumentsstrukturen, Diathese, etc.

➔ Morphosyntaktische Ambiguitätsreduktion von Nomen

Step 2: Soft Constraints

- Wortstellungen, Belebtheit, Definitheit, etc.

➔ Morphosyntaktische Ambiguitätsresolution von Nomen

Seite 6

Architektur

Step 1: Domänenspezifische Vorverarbeitung

- Textsegmentierung: maschinelles Erkennen von Kapitelgrenzen, Abschnittsgrenzen, Überschriften, Fussnoten, Satzgrenzen, ...
- Morphologische Analyse (Gertwol, ergänzt durch TreeTaggers Outputs)
- Morphosyntaktische Desambiguierung von Verben und Erkennung von topologischen Feldern (Constraint Grammar)
- **Morphosyntaktische Ambiguitätsreduktion von Nomen** (Constraint Grammar)
- Morphosyntaktische Ambiguitätsresolution von Nomen und Erkennen grammatischer Funktion

Step 2: Erkennen syntaktischer Stilverletzungen

- Suche nach Verletzungen des syntaktischen Stils

Ziel der morphosyntaktischen Desambiguierung

Input: Gertwols Outputs

"Mitarbeitenden“

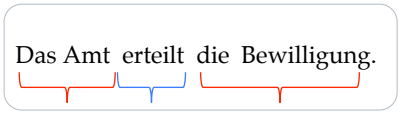
➔		"Mit#arbeit~end"	"S(PART) POS SG AKK MASK“
REMOVE!	;	"Mit#arbeit~end"	"S(PART) POS SG GEN MASK“
	;	"Mit#arbeit~end"	"S(PART) POS SG GEN NEUTR“
	;	"Mit#arbeit~end"	"S(PART) POS PL DAT“
	;	"Mit#arbeit~end"	"S(PART) POS SG DAT MASK“
	;	"Mit#arbeit~end"	"S(PART) POS SG DAT NEUTR“
	;	"Mit#arbeit~end"	"S(PART) POS SG DAT FEM“
	;	"Mit#arbeit~end"	"S(PART) POS SG GEN FEM“
	;	"Mit#arbeit~end"	"S(PART) POS PL NOM“
	;	"Mit#arbeit~end"	"S(PART) POS PL AKK“
	;	"Mit#arbeit~end"	"S(PART) POS PL GEN“

➔ Gewünschter Output: möglichst nur eine morphosyntaktische Analyse pro Token

Vorverarbeitung: Erkennen topologischer Felder

Topologisches Feldermodell:

- Syntaktische Distribution von Verben
- Satzstrukturen und Satztypen

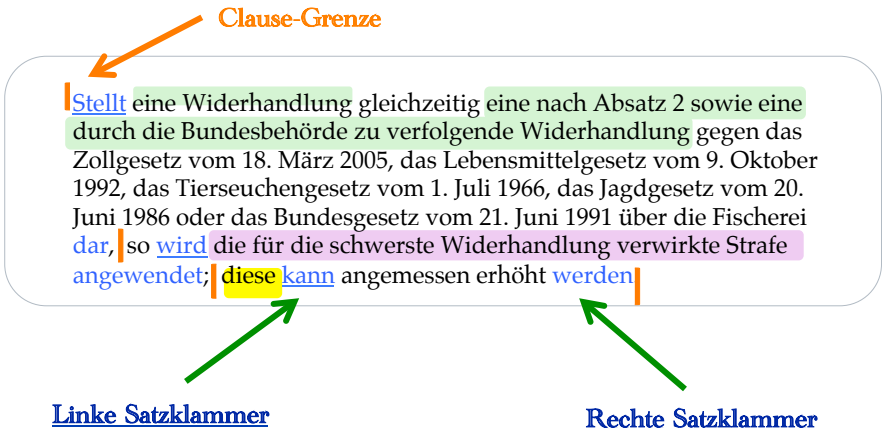


VORFELD LSK MITTELFELD

LSK = Linke Satzklammer
RSK = Rechte Satzklammer

VORFELD LSK MITTELFELD RSK (NACHFELD)

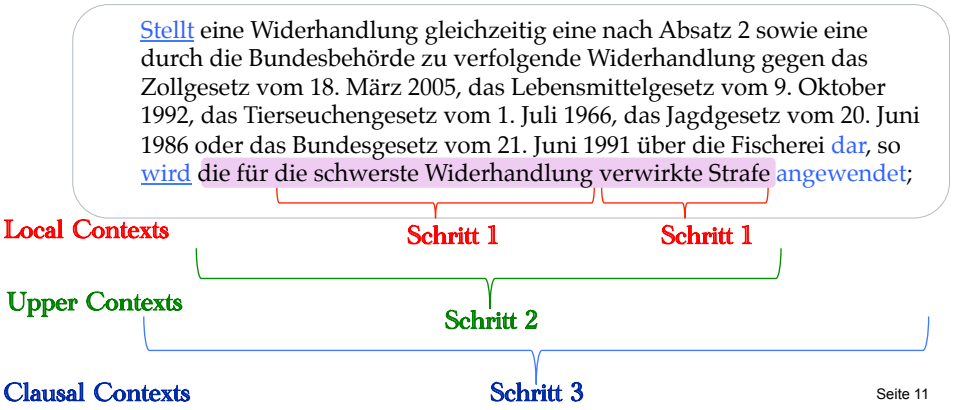
Topologische Felder als Clause-Grenze



Morphosyntaktische Desambiguierung von Nomen

Inkrementelle 3-Schritt Desambiguierung durch harte Constraints:

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



Seite 11

Schritt 1: Local phrase-level feature unification

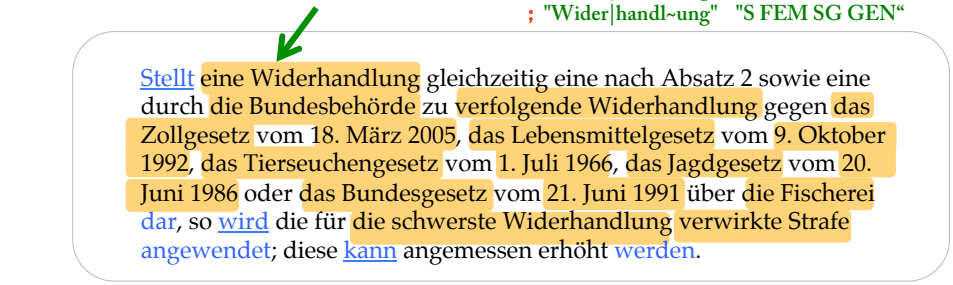
Morphosyntaktische Feature-Unifikation in einfachen NPs:

- Kongruenz: Numerus, Kasus und Genus
- <<REMOVE NOMEN + $$kasus IF (NOT -1 DET + $$kasus) (-1 DET); >>

Ambiguitätsreduzierung: 4 Kasus-Features ➔ 2 Kasus Features

```
eine:
"ein"    "ART INDEF SG NOM FEM"
"ein"    "ART INDEF SG AKK FEM"

Widerhandlung:
"Wider|handl~ung"   "S FEM SG NOM"
"Wider|handl~ung"   "S FEM SG AKK"
; "Wider|handl~ung"   "S FEM SG DAT"
; "Wider|handl~ung"   "S FEM SG GEN"
```

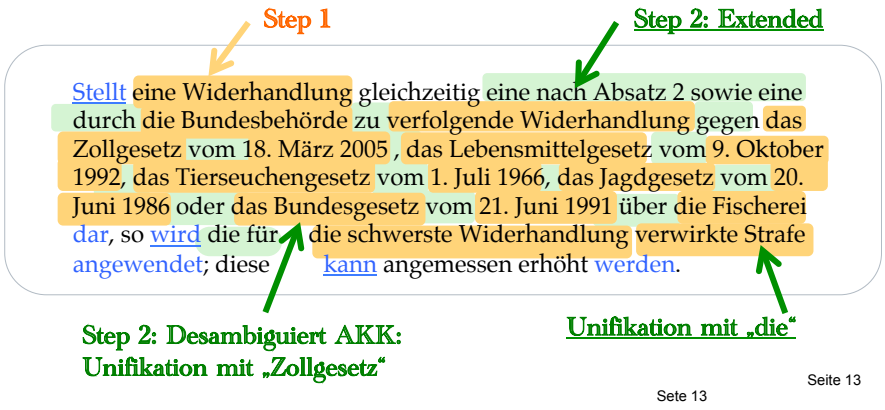


Seite 12

Schritt 2: Upper phrase-level feature unification

Morphosyntaktische Feature-Unification in complex NPs und PPs:

- Kongruenz: NP-Koordination, Partizip-Phrasen, Präpositionalphrase
- REMOVE (&COMPLEX-NP-R-EDGE) + $$kasus
  IF (NOT -1* (&COMPLEX-NP-L-EDGE) + $$kasus)
  (-1* (&COMPLEX-NP-L-EDGE) BARRIER field_boundaries);



Seite 13

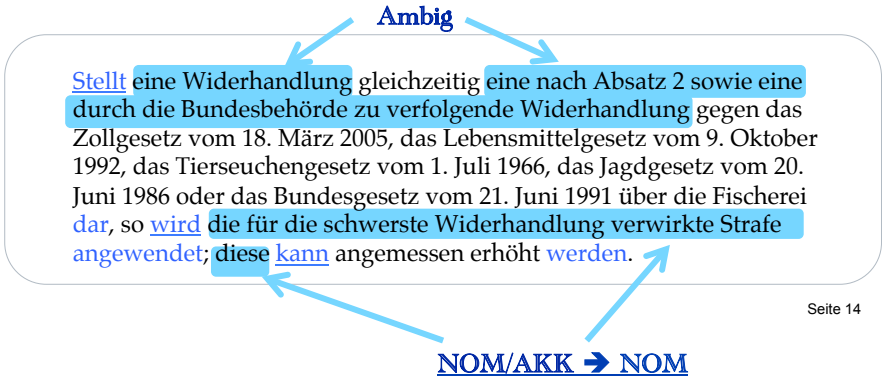
Sete 13

Schritt 3: Clause-level feature unification

Morphosyntaktische Feature-Unifikation von NPs im Clausal-Kontext:

- Subjekt-Verb Kongruenz, Argumentsstrukturen, Diathese, etc.

SELECT (NOM) IF (0 noun_GF) (-1* HS_head BARRIER felder_barrier) (1* (&PRED-ARG1) BARRIER felder_barrier);



Seite 14

Evaluation: Testdaten & Performanz

- Testdaten: 118 Sätze (2,114 Tokens, inkl. 655 Nomen and Pronomen) vom Swiss Legislation Corpus.
- Resultate:
  - (a) 96.30% korrekte morphosyntaktische Analyse: weder rausgekickt noch verloren gegangen
  - (b) 67.60% korrekter Systemoutput

Evaluation (a)

Widerhandlung: (NOM)
"Wider|handl~ung" "S FEM SG NOM"
"Wider|handl~ung" "S FEM SG AKK"
; "Wider|handl~ung" "S FEM SG DAT"
; "Wider|handl~ung" "S FEM SG GEN"

Evaluation (b)

Widerhandlung: (NOM)
"Wider|handl~ung" "S FEM SG NOM"
"Wider|handl~ung" "S FEM SG AKK"
; "Wider|handl~ung" "S FEM SG DAT"
; "Wider|handl~ung" "S FEM SG GEN"

Datenanalyse: 3-Schritt Desambiguierung

Systemdaten: 239 Sätze (4,789 tokens davon 1,668 Nomen/Pronomen) vom Swiss Legislation Corpus.

Schritte	1 Analyse/Token	2+ Analysen/Token
Eingabe	148 (8.87%)	1,520 (91.13%)
Nach 1. Schritt: Local phrase-level feature unification	387 (23.20%)	1,281 (76.80%)
Nach 2. Schritt: Upper phrase-level feature unification	917 (54.98%)	751 (45.02%)
Nach 3. Schritt: Clause-level feature unification	1,129 (67.69%)	539 (32.31%)

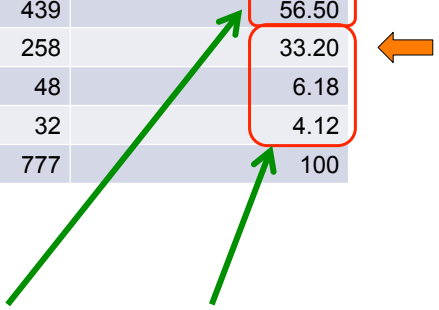
Fertig desambiguiert nach 3-Schritt Desambiguierung

Noch nicht desambiguiert nach 3-Schritt Desambiguierung

Datenanalyse: Desambiguierung von Kasus (GF-Kandidaten)

System Daten: 239 Sätze (4,789 tokens, davon 777 GF-Kandidaten) vom Swiss Legislation Corpus.

	Tokens	%
1 Kasus-Feature/Token	439	56.50
2 Kasus-Features/Token	258	33.20
3 Kasus-Features/Token	48	6.18
4 Kasus-Features/Token	32	4.12
Total: GF-Kandidaten	777	100



Fertig desambiguiert nach 3 Schritten

Noch nicht desambiguiert nach 3 Schritten

Seite 17

Fazit & Ausblick:

Fazit:

- Morphosyntaktische Desambiguierung von Nomen durch harte Constraints:
  - 91.12% ➔ 32.31% (in Testdaten)
- Morphosyntaktische Desambiguierung von Kasus im Testdaten:
  - Desambiguiert: 56.50%
  - Zwei Kasus-Features: 33.20%

Ausblick:

- Morphosyntaktische Desambiguierung von Nomen durch Soft Constraints

Seite 18

Bibliographie

- 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 *Pro- ceedings of the ACL SIGDAT-Workshop.,* Dublin, Ireland.