Universität Zürich[UZH]

Deutsches Seminar

„Linguistische Korpora und grammatische Variation“: Eine kumulative Dissertation

Doktorierendenkolloquium Linguistik

lic. phil. Simone Ueberwasser

„Linguistische Korpora und grammatische Variation: Möglichkeiten und Grenzen korpusbasierter Untersuchungen“ (Arbeitstitel).
Linguistik Deutsch
Fokus Korpuslinguistik, Varietätenlinguistik, Grammatik
Form Kumulative Dissertation
Zeitplan FS2012 – HS2014
Projekt „Variantengrammatik des Standarddeutschen“ (SNF, DFG, FWF)
Betreuung Prof. Dr. Christa Dürscheid
Zweitbetreuer Prof. Dr. Martin Volk

Inhaltsverzeichnis

1 Das Projekt „Variantengrammatik des Standarddeutschen“

- Tri-nationales Projekt (D, CH, AT)
- Sechs Doktorierende
- Ziel: Handbuch zur diatopischen Varianz in der Grammatik
- Standardsprache (kein Dialekt)
- Fokus: CH, AT, DE; Deutschsprachige Teile von Belgien, Luxemburg, Liechtenstein, Südtirol
- 2011 – 2014 (Verlängerung beantragt)
- Cf. www.variantengrammatik.net

Projektverlauf

Phase 1 Aufbau des Korpus

Phase 2 Analyse des Korpus

1. Corpus based
2. Corpus driven

Phase 3 Erstellen des Handbuchs (und einer Online-Ausgabe)

Parallel dazu: Dissertationen

Das Korpus

- Regionalteile von 69 Onlinezeitungen (z.B. Aargauer Zeitung, Basellandschaftliche Zeitung)
- über 640 Millionen (exakt: 643’502’344) Tokens
- Annotation:
  1. TreeTagger: PoS; lemma.
  2. RFTagger: PoS; morphosyntaktische Merkmale wie Kasus, Numerus, Tempus
  3. Morphisto: morphologische Analyse
  4. Strukturelle Einheiten (Autor, Quelle, Datum, Url, etc.).
  5. Syntaktische Einheiten (Satz, Stellungsfelder etc.).

Kritik

– Kann davon ausgegangen werden, dass ein Korpus eine diatopische Verteilung widerspiegelt, nur weil die Daten nach diesen Kriterien ausgesucht wurden?
– Wird allenfalls nicht die sprachliche Realität sondern die Versuchsanordnung widergespiegelt, wenn versucht wird, sehr feingliedrige Details aus dem Korpus zu extrahieren?

2 Kumulative Dissertation

2.1 Vorgaben DPL

Vorgaben des DPL:
– Vier Publikationen
– Davon mindestens eine peer-reviewed
– Eine Synopse, die erklärt, wie die Publikationen zusammenhängen

2.2 Geplante Publikationen

Geplante Publikationen

– Kernthemen: Korpuslinguistik, Varietätenlinguistik, Grammatik
– Geplante Publikationen:
  – «Non-standard data in Swiss text messages with a special focus on dialectal forms»(im Druck).
  – «The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages» (mit B. Ruef, im Druck).
  – „Corpus Annotations and the German Case System. Can Morphosyntactic Case Annotations meet the Needs of German Variational Grammar Research? A Case Study“ (Eingereicht).
  – „Einfluss von Schweizer Wortformen auf die Qualität der morphosyntaktischen Annotation“ (Arbeitstitel).
  – „Wird der Genitiv wirklich verdrängt? Eine Replik auf einen Aufsatz von Konstantin Niehaus“ (Arbeitstitel).
  – „Folgt die Realisierung des Genitiv(e)s mit *-es* oder *-s* einem diatopischen System?“ (Arbeitstitel).
– Mehr Informationen: www.ueberwasser.eu/index.php/ueber-uns

«Non-standard data in Swiss text messages with a special focus on dialectal forms» (peer-reviewed)

Abstract: The investigation in this paper is based on the Swiss SMS corpus. In a first step, the use of the two varieties *Swiss German Dialects* and *Standard Swiss German* are investigated as to who uses which variety and why informants decide to switch varieties. In a second part, the focus is on the spelling of the Swiss German dialects, a group of varieties without spelling norms. This lack of norms leads to an abundance of different spelling forms for the same signifié and accordingly to restrictions when working with the corpus. An additional layer of equivalent terms in Standard German will be created in the corpus to assist researchers.

«The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages» (peer-reviewed)

Abstract: The Swiss German dialect has no fixed orthography. Furthermore, text messages contain a lot of abbreviations and forms of code-switching. To enable corpus search and the application of computational linguistic methods (part-of-speech tagging, syntax parsing etc.) the Swiss SMS corpus was normalized by means of interlinear glossing. This paper describes the tool developed for this task and the practical experiences gained.

„Einfluss von Schweizer Wortformen auf die Qualität der morphosyntaktischen Annotation“

Ausgangslage: „Es kann nämlich sein, dass die Wortarten auch deshalb falsch annotiert werden, weil die Computerlinguisten, die den Tagger entwickelt haben, diesen nur an bundesdeutschen Zeitungen trainierten … Daraus folgt also, pointiert gesagt: Nicht nur in der Linguistik, auch in der Computerlinguistik muss der Tatsache Rechnung getragen werden, dass es mehrere Standardvarietäten des Deutschen gibt.“ (DÜRSCHEID, ELSPASS und ZIEGLER, i. Dr.)

Arbeitshypothese: In der Standardsprache ist die Dichte an nicht-bundesdeutschen Tokens pro Satz nicht gross genug, um den Taggern Probleme zu bereiten beim Erkennen der Wortarten, denn die verwendeten statistischen Methoden können Lücken in den Wortlisten überbrücken. Morphosyntaktische Merkmale hingegen folgen u.U. andern Regeln und werden deshalb falsch annotiert.

„Wird der Genitiv wirklich verdrängt? Eine Replik auf einen Aufsatz von Konstantin Niehaus“

Ausgangslage: Niehaus (noch nicht publiziert) untersuchte die veränderte Frequenz von Genitiven in den letzten 50 Jahren. Er findet einen massiven Rückgang des attributiven Genitivs, während meine eigenen Untersuchungen nur einen marginalen Rückgang zeigen.

Arbeitshypothese: Die unterschiedlichen angewandten Methoden führen zu den unterschiedlichen Resultaten, denn Niehaus vergleicht den Genitivgebrauch mit dessen Alternativen, während ich die Anzahl Genitive im Verhältnis zu den vorhandenen Nomen vergleiche.

„Folgt die Realisierung des Genitiv(e)s mit *-es* oder *-s* einem diatopischen System?“

Ausgangslage: Das ÖWB (2012) listet bei diversen Lemmata die Genitivbildung als *-(e)s* auf, macht aber keine Angaben, ob es sich dabei um regionale Varianten handelt. SZCZEPANIAK andererseits sieht vor allem phonologische Faktoren als Motivation für die jeweilige Variantenwahl (2010: 123).

Arbeitshypothese: Neben den phonologischen spielen auch diatopische Faktoren eine Rolle bei der Variantenwahl.

2.3 Synopse

Linguistische Korpora und grammatische Variation: Möglichkeiten und Grenzen korpusbasierter Untersuchungen

Linguistische Korpora und grammatische Variation
- Drei Aufsätze zur Korpuslinguistik
- Zwei Aufsätze zur Variantengrammatik
- Ein Aufsatz zu einer anderen Form von Variation (SMS)

Möglichkeiten und Grenzen korpusbasierter Untersuchungen
Grenzen:
- Einschränkungen beim Suchen nach Variation (-> Kasus-Aufsatz)
- Einschränkungen beim Suchen nach Kasus (-> Kasus-Aufsatz)
- Einschränkungen bei nicht-bundesdeutschen Varianten (-> Aufsatz CH-Wortformen)
- Einschränkungen bei nicht-standardisierter Graphie (->SMS-Aufsatz).

Möglichkeiten und Grenzen korpusbasierter Untersuchungen
Möglichkeiten:
- Finden von grammatischer Variation (-> Beide geplanten Genitiv-Aufsätze)
- Finden von Helvetismen (-> Aufsatz CH-Wortformen)
- Finden von soziolinguistischen Faktoren (->SMS-Aufsatz)
- Verbesserung der Möglichkeiten durch Standardisierung der Daten (->SMS-Aufsatz)

3 Corpus Annotations and the German Case System

Can Morphosyntactic Case Annotations meet the Needs of German Variational Grammar Research? A Case Study.

Fragestellungen:
- Welcher Kasus wird bestimmt?
- Spezielle Bedürfnisse in der Variantengrammatik?
- Wie zuverlässig sind morphosyntaktische Annotationen?
- Parser als eine Alternative?

Case Study: *wegen* mit Genitiv oder Dativ?

3.1 Welcher Kasus wird bestimmt?

Kasusträger

„Kasusträger sind komplexe Einheiten im Syntagma. Im syntaktischen Kontext ist der Kasusträger die ganze Phrase, die kasusmarkiert ist und innerhalb derer Kasuskongruenz zwischen Adjektiv, Determinans und Kernnomen besteht.“ (DÜRSCHEID, 1999: 2)

Beispiele:
- *der neue Mitarbeiter*
- *diese unfreundliche Frau*

Kasusform

„Den Kasusformen entsprechen die einzelnen Positionen in einem Paradigma.“ (DÜRSCHEID, 1999: 2)

Beispiele:
- *[der neue Mitarbeiter]$_{Nom/Sg}$*
- *[diese unfreundliche Frau]$_{Nom/Sg\ oder\ Akk/Sg\ ->\ Synkretismus}$*

> Entscheidungskriterium: Morphologie

Kasuskategorie

„Kasuskategorien sind abstrakte Einheiten der linguistischen Beschreibung. Sie bezeichnen Paradigmen, Klassen von Wortformen, die für einander einsetzbar sind.“ (DÜRSCHEID, 1999: 2)

Beispiele:
- *[Der neue Mitarbeiter]$_{Nom/Sg}$ hat heute angefangen.*
- *[Diese unfreundliche Frau]$_{Nom/Sg\ oder\ Akk/Sg}$*
  - *[Diese unfreundliche Frau]$_{Nom/Sg}$ hat heute angefangen.*
  - *[Diese unfreundliche Frau]$_{Akk/Sg}$ habe ich gestern gesehen.*

> Entscheidungskriterium: Kontext

Morphosyntaktischer Tagger (am Beispiel des RFTaggers, vgl. SCHMID und LAWS, 2008)
- ”The joint probability of the two sequences is defined as the product of context probabilities and lexical probabilities over all POS tags“ (SCHMID und LAWS, 2008: 1).
- ”The additional information may also help to disambiguate the (base) part of speech“: „*Ist das Realität?*“ (SCHMID und LAWS, 2008: 1).

> Kann mit diesen Mitteln nach sprachlicher Varianz gesucht werden?

3.2 Varietätengrammatik als spezielle Herausforderung

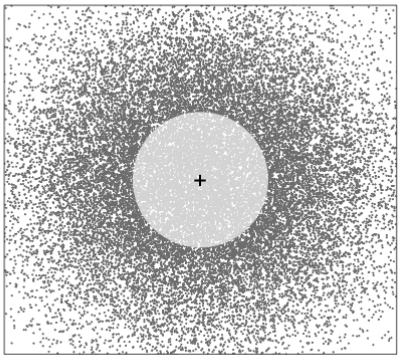


Abbildung 1: Kern der Sprache (und der Annotation) im Zentrum. Phänomene, die für die Varietätengrammatik interessant sind, liegen ausserhalb des Bereiches, für den der Tagger erstellt wurde. (Nach BELICA u. a., 2011).

Kern der Sprache vs. Peripherie der Sprache
- Können Varianten ausserhalb der Kernsprache gefunden werden?
- Sind quantitative Aussagen zu Kasus-Varianz möglich?
  - Unterscheidung Kasuskategorie/Kasusform
  - Erkennen und Ausschliessen von Synkretismen

> Der Zweifelsfälle-Duden akzeptiert nach der Präposition *wegen* nur den Genitiv als standardsprachlich (2011: 999). Wie gross ist der Anteil an abweichende Formen (d.h. Dative) im VG-Korpus? Um das korrekte Verhältnis zu berechnen, müssen Synkretismen ausgeschlossen werden können.

3.3 Zuverlässigkeit der Annotation

Fragestellungen:
- *Wegen* als Präposition oder als Postposition
- Annotation des TreeTaggers und des RFTaggers

> Disambiguierung: Von 162’140 Sätzen mit *wegen* sind deren 1’979 von den Taggern unterschiedlich annotiert.
>
> [lemma = "wegen" & pos = "APPR" & rfpos != "APPR.*"]

Von 100 zufälligen Belegen, die vom TreeTagger als *APPR* markiert werden, annotiert der RFTagger:

> Für die folgenden Untersuchungen werden nur Belege berücksichtigt, bei denen beide Tagger *wegen* als Präposition markiert haben.

	Anzahl	Beispiel
Postposition	41	„Vorher war auch eine Frau von Gesetz wegen ein Standesbeamter.“
Weg (Pl.Dat.)	31	„Bis das Wasser auf Wiesen und Wegen wieder verschwindet [...]“
von wegen	8	„Von wegen Boxenluder:“
Korrekt	20	„Wegen der möglichen Schliessung [...]“

Die Abfrage (*wegen* als Präposition)(kein Kasusträger)*(Kasusträger)

(1) a. [lemma = "wegen" & pos = "APPR" & rfpos = "APPR.*"]

b. [rfpos != "N.*(Gen|Dat).*" & rfpos != "ADJA.*(Gen|Dat).*" & rfpos != "ART.* (Gen|Dat).*" & rfpos != "PP.*(Gen|Dat).*" & rfpos != "PI.*(Gen|Dat).*" & rfpos != "PW.* (Gen|Dat).*"]*

c. [rfpos = "N.*(Gen).*" | rfpos = "ADJA.*(Gen).*" | rfpos = "ART.*(Gen).*" | rfpos = "PP.* (Gen).*" | rfpos = "PI.*(Gen).*" | rfpos = "PW.*(Gen).*"]

156’981 Belege im Genitiv und 51’603 Belege im Dativ

Tabelle 1: Manuelle Kontrolle von 100 zufällig ausgewählten Belegen je Kasus

	Tatsächlicher Kasus				
	Gen	Dat	Akk	Synkr.	Andere
Als Gen. annotiert	60	1		39	
Als Dat. annotiert	7	26	1	64	2

Zuverlässigkeit

Bei 86 von 200 Belegen (43%) wird der korrekte Kasus zugewiesen.

Andere

Fehler im Originaltext:

(2) *wegen geringem Fahrzeugaufkommens (sic.)*

Fehler beim Aufbereiten der Daten:

(3) *Der Liebe wegen nach Polling 1947 übernahm die damals schon selbstständige Frau das Wirtshaus ihrer Schwiegereltern.*

Falsche Annotation des Kasus: Zwei Typen von Fehlern

1. Zu viele Tokens zwischen den einzelnen Elementen des Kasusträgers (alle 8 falsch annotierten Genitive):

(4)

wegen	**seiner**	**öffentlich**	**ausgetragenen**
APPR.Gen	PRO.Poss.Attr.-.*.Sg.Fem	ADJD.Pos	ADJA.Pos.Acc.Pl.Fem

Differenzen
N.Reg.Acc.Pl.Fem

2. Kasusträger, die nur aus dem Nomen bestehen (der falsch annotierte Dativ):

(5) *Verbrechen wegen Karnickelfutter*

> Duden Genitivregel: Eine Nominalphrase kann nur dann im Genitiv stehen, wenn sie (i) mindestens ein adjektivisch flektiertes Wort und (ii) mindestens ein Wort mit *s*- oder *r*-Endung enthält. (DUDEN, 2011: 968)

wegen junger Hunde
* *wegen Hunde*

Synkretismus

103 Belege total
– 91 Feminin Singular
– 12 Feminin Plural

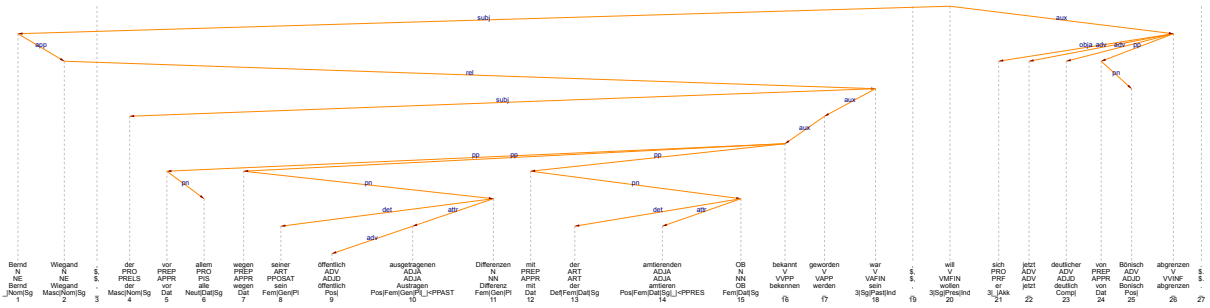
RFTagger: Zwischenbilanz

– Untersuchungen zum Kasus sind komplex (z. B. Genitivregel).
– Kasusannotationen sind oft fehlerhaft bei komplexen Konstruktionen. Gerade diese sind aber oft interessant.
– Synkretismen sind frequent. Der Tagger teilt immer einen Kasus zu, was eine quantitative Auswertung verfälscht oder verunmöglicht oder sehr aufwändig macht.

Mögliche Alternativen:
– Ausschluss von Synkretismen über Deklinationsklassen
– Dependenzparser

3.4 Parser als Alternative?

am Beispiel von ParZu (vgl. SENNRICH, SCHNEIDER und VOLK, 2009)



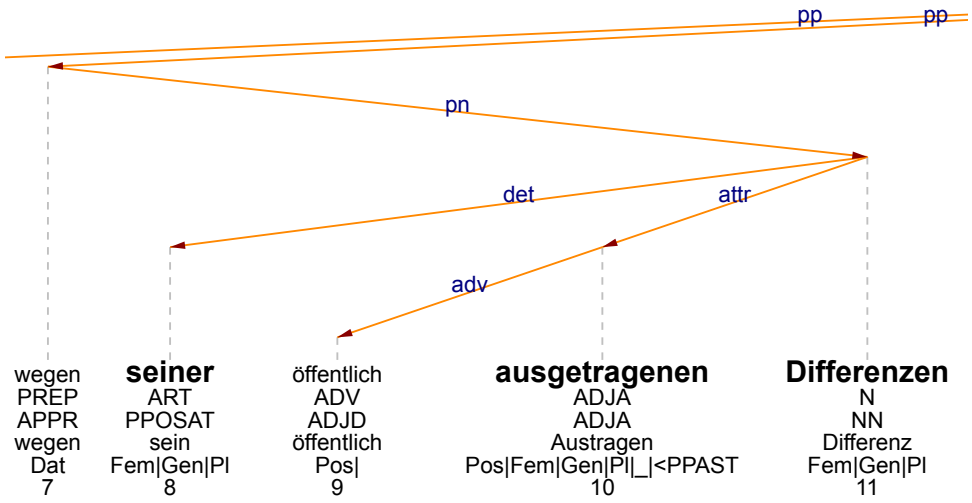


Tabelle 2: Manuelle Kontrolle von 96 Belegen, bei denen der Kasus manifestiert ist

	Tatsächlicher Kasus	
	Genitiv	Dativ
Als Gen. annotiert	58	1
Als Dat. annotiert	3	26

Weiterer Vorteil des Parsers

Richtige Annotation vorausgesetzt, untersucht man mit einem Parser immer den richtigen Kasusträger, auch bei komplexen Konstruktionen.

Synkretismen

- Auch der Parser weist immer einen Kasus zu.
- Quantitative Untersuchungen zu Varianz sind deswegen auch mit einem Parser schwierig.

3.5 Fazit

- Korpusbasierte Untersuchungen zu Varianz sind anspruchsvoll.
- Untersuchungen zum Kasusgebrauch sind auf drei Ebenen problematisch:
  - Korrektheit der Annotation

(*wegen* als Präposition)(kein Kasusträger)*(Kasusträger)

(1) a. [lemma = "wegen" & pos = "APPR" & rfpos = "APPR.*"]

b. [rfpos != "N.*(Gen|Dat).*" & rfpos != "ADJA.*(Gen|Dat).*" & rfpos != "ART.* (Gen|Dat).*" & rfpos != "PP.*(Gen|Dat).*" & rfpos != "PI.*(Gen|Dat).*" & rfpos != "PW.* (Gen|Dat).*"]*

c. [rfpos = "N.*(Gen).*" | rfpos = "ADJA.*(Gen).*" | rfpos = "ART.*(Gen).*" | rfpos = "PP.* (Gen).*" | rfpos = "PI.*(Gen).*" | rfpos = "PW.*(Gen).*"]

156'981 Belege im Genitiv und 51'603 Belege im Dativ

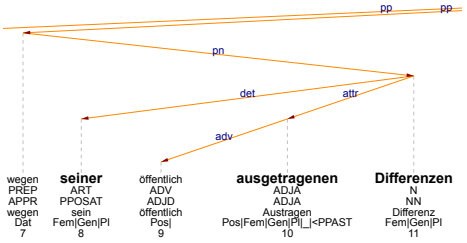


Abbildung 2: Vergleich der Komplexität und Präzision von Taggern und Parsern

  - Finden des Kasusträgers
  - Ausschluss von Synkretismen
- Parser liefern bessere Kasusannotationen als Tagger, das Problem der scheinbar zufällig zugeordneten Synkretismen besteht aber auch hier.

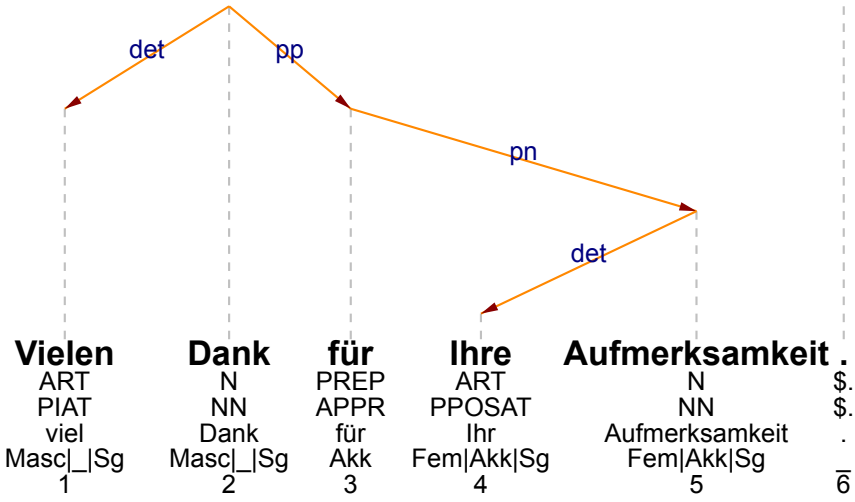
3.6 Konsequenz für das Projekt „Variantengrammatik“

- Genaues Festlegen des Untersuchungsgegenstandes
  - Kasusträger, die nur aus einem Nomen bestehen getrennt behandeln (Genitivregel)
  - *Wegen* als Prä- und als Postposition getrennt behandeln
- Abwägen einer Methodik, die zu einem möglichst grossen Recall und einer möglichst grossen Precision führt
  - Welcher Tagger unterscheidet besser zwischen Prä- und Postposition?
  - Wie kann der Kasusträger am besten gefunden werden?
- Ergänzende Suche von Belegen, die nicht berücksichtigt werden sollten. Subtrahieren der entsprechenden Werte.
  - Synkretismen aufgrund von Deklinationsklassen
  - *von wegen*

Mögliche Erweiterung: Parser

Probleme

- Zeitlicher Aufwand
- Möglichkeiten der Abfrage
  - In Graphiken kann nicht systematisch gesucht werden.
  - Prolog als Abfragesprache ist sehr komplex und nicht mit dem bestehenden Korpus kombinierbar.
  - Einbinden in bestehende CQP-Umgebung erlaubt keine Abfrage nach Dependenzen, verbessert jedoch tendenziell die Annotation der morphosyntaktischen Annotationen.



Literatur

BELICA, Cyril u. a. (2011): «The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls». In: *Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.-24.09.2009*. Hrsg. von Marek Konopka, Jacqueline Kubczak und Christian Mair. (= Korpuslinguistik und interdisziplinäre Persepektiven auf Sprache 1). Tübingen: Narr, 451–469.

DUDEN (2011): *Richtiges und gutes Deutsch. Wörterbuch der sprachlichen Zweifelsfälle*. 7., vollständig überarbeitete und erweiterte Auflage. (= Der Duden in zwölf Bänden 9). Mannheim: Bibliographisches Institut und F. A. Brockhaus.

DÜRSCHEID, Christa (1999): *Die verbalen Kasus des Deutschen*. Berlin, New York: de Gruyter.

DÜRSCHEID, Christa, Stephan ELSPASS und Arne ZIEGLER (i. Dr.): «Variantengrammatik des Standarddeutschen. Konzeption, methodische Fragen, Fallanalysen». In: *Wiener Arbeiten zur Linguistik (WAL)*.

ÖWB (2012): *Österreichisches Wörterbuch*. Hrsg. von Otto Back, Erich Benedikt und Karl Blüml. 42., neu bearbeitete Auflage. Wien: öbv.

RUEF, Beni und Simone and UEBERWASSER (2013): «The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages». In: *Non-standard Data Sources in Corpus-based Research*. Hrsg. von Marcos Zampieri und Sascha Diwersy. (= ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln 5). Maastricht: Shaker.

SCHMID, Helmut und Florian LAWS (2008): «Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging». In: *COLING 2008*. URL: *http://www.cis.uni-muenchen.de/~schmid/papers/Schmid-Laws.pdf*.

SENNRICH, Rico, Gerold SCHNEIDER und Martin VOLK (2009): «A New Hybrid Dependency Parser». In: *Proceedings of GSCL-Conference*.

SZCZEPANIAK, Renata (2010): «Während des Flug(e)s/des Ausflug(e)s? German Short and Long Genitive Endings between Norm and Variation». In: *Grammar between Norm and Variation*. Hrsg. von Alexandra N. Lenz und Albrecht Plewnia. (= VarioLingua 40). Frankfurt a. M. et al.: Peter Lang, 103–126.

UEBERWASSER, Simone (2013): «Non-standard data in Swiss text messages with a special focus on dialectal forms». In: *Non-standard Data Sources in Corpus-based Research*. Hrsg. von Marcos Zampieri und Sascha Diwersy. (= ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln 5). Maastricht: Shake.