Universität Zürich
Deutsches Seminar

**Seminar „Korpuslinguistik“**
Herbstsemester 2012
Dozentin: Prof. Dr. Christa Dürscheid

**Seminararbeit**

# Korpuslinguistische Untersuchungen von Variationsphänomenen im Zeitungskorpus „Variantengrammatik des Standarddeutschen“

Ein Analysebericht

Bettina Rimensberger

Abgabedatum: 18.9.2013

# Inhaltsverzeichnis

# 1 Einleitung

Das Deutsche ist eine plurizentrische Sprache, bei der verschiedene nationale und regionale Varietäten nebeneinander existieren. Diese Varietäten unterscheiden sich durch phonetische, orthographische, lexikalische, aber auch durch grammatische Merkmale, die sich auf morphologischer und syntaktischer Ebene manifestieren können. Während Variationsphänomene in den erstgenannten Variationsbereichen augen- bzw. ohrenfälliger und schon gut dokumentiert sind (vgl. etwa zur lexikalischen und phraseologischen Variation das „Variantenwörterbuch des Standarddeutschen“ von Ammon et al. 2004 und zur phonetischen Variation den „Atlas zur Aussprache des Schriftdeutschen“ von König 1989 und das Nachfolgeprojekt „Deutsch heute“ am Institut für deutsche Sprache in Mannheim), stellt die Erforschung der nationalen und regionalen Variation in der Grammatik der deutschen Standardsprache immer noch ein Desiderat der Grammatikographie dar. An dieser Stelle setzt das trinationale Projekt „Variantengrammatik des Standarddeutschen“ an. Auf der Basis eines Korpus, das gemäss eines Regionenschlüssels Artikel aus den Regionalteilen von online Zeitungen aus dem zusammenhängenden deutschen Sprachraum enthält, sollen die grammatische Variabilität im geschriebenen Gebrauchsstandard (der Begriff wird von Ammon 1995, 88 eingeführt und steht „in semantischer Opposition zum Terminus *kodifizierter Standard*“) der regionalen Presse untersucht und die gewonnenen Erkenntnisse in einem geplanten Handbuch „Variantengrammatik“ dokumentiert werden (vgl. zur Anlage und den Zielsetzungen des Projekts „Variantengrammatik des Standarddeutschen“ sowie zum Forschungsstand zur grammatischen Variation Dürscheid et al. 2011). Als Kriterium dafür, ob ein bestimmtes Variationsphänomen zur Standardsprache eines Landes oder einer Region gerechnet werden kann, oder ob es nur zur schriftlichen standardnahen Umgangssprache zählt (vgl. zur Unterscheidung von *Standardsprache*, *Gebrauchsstandard* und *Umgangssprache* Dürscheid/Giger 2010, 167f.), wird seine Frequenz im Korpus gelten. Das Projekt möchte nicht nur bekannte Variationsphänomene korpusbasiert überprüfen, sondern erhofft sich auch, im Korpus neue Variationsmuster erkennen zu können (*corpus-driven*). Darüber hinaus verspricht es sich einen sprachpolitischen Einfluss dahingehend ausüben zu können, dass standardsprachliche Varianten nicht mehr, wie bisher oft, als minderwertige stigmatisiert, sondern als gleichwertige akzeptiert werden.

Bis zum jetzigen Zeitpunkt sind 84 online Zeitungen aus Deutschland, Österreich, der

Schweiz, Liechtenstein, Südtirol, Belgien und Luxemburg aus den Jahren 2012 und Anfang 2013 erfasst, die in 17 Sektoren eingeteilt werden.

Das Ziel dieser Arbeit liegt darin, in dieses Korpus „einzutauchen“, bekannte Variationsphänomene zu untersuchen (vornehmlich *corpus-based*) und bei dieser Gelegenheit herauszufinden, welche korpuslinguistischen Untersuchungen zur diatopischen Variation unter den gegebenen Umständen des Korpus gut möglich sind und wo die Analyse mit (annotations-)technischen Schwierigkeiten konfrontiert wird. Dieses Vorhaben bestimmt die Gliederung dieser Arbeit. Nach der Beschreibung des methodischen Vorgehens bei der Arbeit am Korpus (Kapitel 2) werden in Kapitel 3 ausgewählte Phänomenbereiche mit den entsprechenden Korpusabfragen präsentiert, die sich korpusbasiert mehr oder weniger problemlos recherchieren lassen und zu aussagekräftigen Ergebnissen führen. In Kapitel 4 hingegen werden die Auswirkungen der technischen Schwierigkeiten, die ich auf drei verschiedenen Ebenen lokalisiere (Crawling, Lemmatisierung und Annotation), anhand von Beispielabfragen von grammatischen Variationsphänomenen illustriert. Diese Evaluation verfolgt das Ziel, auf kritische Punkte aufmerksam zu machen, das Bewusstsein für die diagnostizierten Problemfelder zu schärfen und Konsequenzen für künftige Korpusabfragen aufzuzeigen.

Die Auswahl der in der Arbeit untersuchten Variationsphänomene hat exemplarischen Charakter. Es wird darauf geachtet, dass die verschiedenen sprachlichen Bereiche (Lexik, Wortbildung, Flexion, Rektion, Reflexivität von Verben und Syntax) berücksichtigt werden. Dadurch soll es möglich sein, die an einem spezifischen Beispiel durchgeführten Analysen auf ähnliche Phänomene zu übertragen. Als Inspirationsquelle dienten einerseits die Überblicksdarstellungen in Dürscheid/Hefti 2006, Dürscheid et al. 2011 und Elspaß 2010, aber auch die Datenbank zum Projekt „Variantengrammatik des Standarddeutschen“ (https://www-gewi.uni-graz.at/variantengrammatik/ <15.9.2013>) und projektinterne Dokumente, die in der Bibliographie unter Kapitel 6.2 aufgeführt sind. Schwerpunktmässig werden in dieser Arbeit Phänomene untersucht, bei denen die Schweizer Variante vom bundesdeutschen Standard, mit dem die morpho-syntaktischen Tagger trainiert wurden, abweicht.

Die Seminararbeit versteht sich auch als Fortführung meiner Präsentation im Rahmen des Seminars „Korpuslinguistik“, in der Variationsphänomene am Pilotkorpus untersucht wurden. Insbesondere knüpft die vorliegende Arbeit an jenen Punkten an, die damals mit den Möglichkeiten einer RegEx Abfrage in einem nicht-annotierten Korpus nicht abschliessend geklärt werden konnten.

## 2 Methodisches Vorgehen

Alle Abfragen in dieser Arbeit wurden im Korpus „Variantengrammatik des Standarddeutschen“ auf dem Stand des MarchCrawls (http://corpora.semtracks.org/marchcrawl/ <15.9.2013>) mit dem Query Mode CQP Syntax durchgeführt. Dem MarchCrawl wurde gegenüber dem SummerCrawl der Vorzug gegeben, da sich letzterer im Untersuchungszeitraum (Juni bis September 2013) verändert hätte.

Grundsätzlich wurde bei den Abfragen nach dem Sechs-Punkte-Programm von Belica et al. (2010, 467) vorgegangen, d.h. es wurde zuerst ein möglichst grosser Recall angestrebt, mit dem alle Belege für das gesuchte Phänomen gefunden werden sollten (keine *false negatives*). Auf dieser Basis wurden Filter gesucht, durch die *false positives* ausgeschlossen werden konnten. Dieser Prozess wiederholte sich so lange, bis die präzisest mögliche Abfrage gefunden wurde. Falls die Abfragesyntax nicht soweit optimiert werden konnte, dass nur relevante Belege gefunden wurden, musste die Trefferliste manuell ausgezählt werden. Dies war bei den meisten untersuchten Phänomenen notwendig. Gemäss dem Flussdiagramm im projektinternen Dokument „Mustervorgehen im Projekt *Variantengrammatik des Standarddeutschen*“ (Ueberwasser 2013b, 2) wurden Fehlerraten von 5 Prozent als tolerabel erachtet.

Der Prozess des Herantastens an die präziseste Abfrage kann als *trial and error* bezeichnet werden. Es galt, alle Resultate kritisch zu hinterfragen und sicherheitshalber falls immer möglich mit einer zweiten Abfrage zu überprüfen, ob keine *false negatives* durch's Netz gefallen sind (vgl. Belica et al. 2010, 461). Insbesondere stellte es sich als ratsam heraus, zu vergleichen, ob die Abfrage über das Lemma gleich viele Treffer hervorbrachte wie jene über Word. Gelegentlich lohnte sich auch eine Disambiguierung der beiden Tagger (vgl. Kapitel 4.4). Um die in Kapitel 4 skizzierten Limitierungen des Korpus umgehen zu können, mussten für jeden spezifischen Fall Lösungsstrategien erarbeitet werden.

Zur Überprüfung der Signifikanz von länderspezifischen Verteilungen der Variationsphänomene wurde der Chi-Quadrat-Test eingesetzt, wobei nur die drei Hauptregionen Deutschland, Österreich und Schweiz berücksichtigt wurden, da in den kleineren Regionen (Belgien, Luxemburg, Liechtenstein und Südtirol) die absoluten Zahlen zu gering waren (häufig kleiner als 5). Da die Gesamtzahl der Tokens der verschiedenen Länder nicht ausgewogen ist, werden, wenn nicht ein Chi-Quadrat-Test, der die unterschiedlichen Grössen der Teilkorpora berücksichtigt, durchgeführt werden kann, die relativen

Frequenzen in der Einheit pMW (Instanzen pro Million Wörter, auf eine Nachkommastelle gerundet) angegeben.

Die Begriffe *Helvetismus, Austriazismus* und *Teutonismus* verwende ich im Sinn von absolutem Helvetismus, Austriazismus, Teutonismus und die Begriffe *Frequenzhelvetismus, -austriazismus* und *-teutonismus* im Sinn von Walter Haas (2000, 100) für die relativen Varianten, die in einem Land besonders häufig, aber nicht ausschliesslich dort, vorkommen.

# 3 Kommentierte Beispielabfragen im Korpus „Variantengrammatik des Standarddeutschen“

In diesem Kapitel werden diejenigen Variationsphänomene präsentiert und kommentiert, die sich mehr oder weniger problemlos im Korpus „Variantengrammatik des Standarddeutschen“ untersuchen lassen und die zu eindeutigen Ergebnissen führen. Es handelt sich dabei um Phänomene aus den Bereichen Wortbildung (Kapitel 3.1), Valenz von Verben (Kapitel 3.2) und Satzstrukturen (Kapitel 3.3).

## 3.1 Wortbildung

### 3.1.1 Derivation bei Substantiven und Adjektiven*: Zürcher / Züricher, zürcher / züricher / zürcherisch*

Eine morphologische Variante liegt in der Benennung der Bewohner von Zürich und der Bildung des zur Stadt gehörigen Adjektivs vor. Während in Deutschland und Österreich der Ortsnamen in der Derivation erhalten bleibt (*Züricher*), fällt in der Schweiz das Schwa weg (*Zürcher*) (vgl. Dürscheid et al. 2011, 130 und Ammon et al. 2004, 902). Gemäss dem Duden „Die deutsche Rechtschreibung“ sind beide Varianten erlaubt mit dem Vermerk „*in der Schweiz nur* Zürcher“ (Duden 2009a, 1207). Das Gleiche gilt für das indeklinable Adjektiv (*zürcher/züricher*) und das deklinable Adjektiv (*zürcherisch*). Die Suche nach den Substantiven und Adjektiven lässt sich im Korpus leicht durchführen.

[(lemma="(Z|z)ürcher((in)|(isch)?)")] liefert 3459 Treffer (davon 2486 aus der Schweiz)

[(lemma="(Z|z)üricher((in)|(isch)?)")] liefert 423 Treffer (davon 5 aus der Schweiz)

Es fällt auf, dass die Adjektive *züricher* und *züricherisch* in keinem einzigen Fall verwendet werden. Der Chi-Quadrat-Wert von 41149 bestätigt die Varianten mit Schwa-Elision als signifikante Frequenzhelvetismen. Die Varianten mit Erhalt des Ortsnamens in der Derivation werden – nicht überraschend – in der Schweiz signifikant weniger häufig verwendet als in den anderen deutschsprachigen Ländern (Chi Quadrat = 13.1).

### 3.1.2 s-Suffix bei Adverbien: *durchwegs / durchweg*

Innerhalb des deutschsprachigen Gebiets besteht die Variationsmöglichkeit, ob Adverbien wie *öfter(s), durchweg(s), durchgehend(s)* und *weiter(s)* mit oder ohne s-Suffix realisiert werden (vgl. Dürscheid et al. 2011, 130 und die zu diesem Phänomen in der Datenbank zum Forschungsprojekt „Variantengrammatik des Standarddeutschen“ aufgelis-

tete Literatur: https://www-gewi.uni-graz.at/variantengrammatik/variante/533 <9.9.2013>). Da es sich beim s-Suffix um ein distinktives morphologisches Merkmal handelt, gestaltet sich eine Untersuchung der Varianten *durchwegs* und *durchweg* am Korpus problemlos.
Sowohl [(word="(D|d)urchweg")]” als auch [(lemma="durchweg")]” finden die gleichen 4963 Treffer. Dasselbe gilt für die Variante *durchwegs*:
Für [(word="(D|d)urchwegs")] und [(lemma="durchwegs")] resultieren je 1340 Treffer. Hier zeigt sich alles „durchwegs“ positiv. Die Suche nach den Lemmata findet in beiden Fällen alle Tokens, die Variante *durchwegs* wird als eigenes Lemma erkannt und ist frequent in Liechtenstein, Österreich und der Schweiz (Chi Quadrat = 2105.6).

## 3.2 Verben

### 3.2.1 Absolute Verwendung eines reflexiven Verbs: *sich rentieren / rentieren*

Einige reflexive Verben wie z.B. *sich rentieren*, *sich ändern* und *sich vorbehalten* können in der Schweizer Standardvarietät auch absolut (*rentieren, ändern, vorbehalten*) verwendet werden (vgl. Dürscheid/Hefti 2006, 137, zu *sich rentieren* vgl. Ammon 1995, 271 und die in der Datenbank zum Forschungsprojekt „Variantengrammatik des Standarddeutschen“ aufgeführte Literatur:
https://www-gewi.uni-graz.at/variantengrammatik/variante/1271 <11.9.2013>).
Der Wegfall des Reflexivpronomens *sich* gegenüber dem kodifizierten bundesdeutschen Standard, mit dem die morpho-syntaktischen Tagger trainiert wurden, lässt sich über die Ausschlussfunktion korpusbasiert untersuchen. Anders sieht es aus in der inversen Situation, wenn ein absolutes Verb reflexiv verwendet wird. Die gegenüber dem bundesdeutschen Standard zusätzlichen Reflexivpronomen bereiten den Taggern grosse Annotationsprobleme (vgl. Kapitel 4.3.1 zu *sich nerven*).
Zum Lemma *rentieren* [lemma="rentieren"] finden sich insg. 1089 Treffer. Von diesen werden, falls vor und hinter dem Lemma je eine Spannweite von 10 Wörtern angenommen wird, innerhalb derer das Reflexivpronomen *sich* vorkommen muss, 906 reflexiv verwendet:
([lemma="rentieren"] []{0,10} [rfpos="PRO.Refl.*"])|([rfpos="PRO.Refl.*"] []{0,10} [lemma="rentieren"]) within c (906 Treffer). Die verrechneten absoluten Treffer ergeben einen Chi-Quadrat-Wert von 13.5, der anzeigt, dass *rentieren* in der Schweiz signifikant weniger häufig reflexiv verwendet wird als in Deutschland und Österreich.

Diese Abfrage könnte *false positives* enthalten, bei denen sich das Reflexivpronomen nicht auf *rentieren*, sondern auf ein anderes reflexives Verb bezieht. Jedoch sind für diesen Fall, wenn überhaupt, äusserst tiefe Zahlen anzunehmen.
Eine Disambiguierung der Tagger zeigt erfreulicherweise, dass die Reflexivpronomen sowohl vom RFTagger als auch von TreeTagger in allen Fällen korrekt erkannt werden, was nicht weiter erstaunt, zumal ja die reflexive Variante den kodifizierten Standard darstellt. Vgl. zur Disambiguierung die durchgeführten Testabfragen mit dem Reflexivpronomen *sich* hinter resp. vor dem Lemma *rentieren*:

| | |
|---|---|
| [lemma="rentieren"] []{0,3} [rfpos="PRO.Refl.*"] within c | 320 Treffer |
| [lemma="rentieren"] []{0,3} [pos="PRF"] within c | 320 Treffer |
| [rfpos="PRO.Refl.*"] []{0,3} [lemma="rentieren"] within c | 374 Treffer |
| [pos="PRF"] []{0,3} [lemma="rentieren"] within c | 374 Treffer |

Die folgende Abfrage wurde mit dem RFTagger durchgeführt.

Die absolute Variante *rentieren* muss über einen komplizierten Ausschluss des Reflexivpronomens *sich* an einer beliebig grossen Anzahl Stellen vor und hinter dem Lemma *rentieren* recherchiert werden.[1]

(<c> [rfpos!="PRO.Refl.*" & c & !c1]* [c1]* [lemma="rentieren" & c & !c1] [c1]* [rfpos!="PRO.Refl.*" & c & !c1]* </c>) | (<c1> [rfpos!="PRO.Refl.*" & c1 & !c2]* [c2]* [lemma="rentieren" & c1 & !c2] [c2]* [rfpos!="PRO.Refl.*" & c1 & !c2]* </c1> )
(141 Treffer)

Da in Deutschland und Österreich *rentieren* im Zusammenhang mit *Staatsanleihen, Papieren, Investitionen* u.ä. stets absolut verwendet wird, dies aber nicht der gesuchten absoluten Konstruktion entspricht, wurden diese von Hand aus den 141 Treffern aussortiert. Zusätzlich begegneten in diesem Recall Tippfehler (*sie* statt *sich*) und – wie auch bei anderweitigen Recherchen sehr häufig – Clause Fehler, durch die sehr viele *false positives* verursacht werden. Die manuelle Auszählung ergibt 90 relevante absolute Verwendungen von *rentieren* in der Schweiz und 1 aus dem deutschen Südkurier. In allen übrigen Regionen wird dieser Satzbau nicht verwendet, wodurch die eingangs zitierte Zuordnung Ammons (1995, 271), dass *rentieren* nur in der Schweiz absolut verwendet werde, bestätigt wird (abgesehen vom einen Treffer aus Süddeutschland). Der Chi-

[1] Vielen Dank an dieser Stelle an Klaus Rothenhäusler für das Zur-Verfügung-Stellen der Formel.

Quadrat-Wert von 2080.3 widerspiegelt diesen signifikanten grammatischen Frequenzhelvetismus.

### 3.2.2 Verbrektion: *telefonieren* / *telefonieren mit*

Eine ähnliche Strategie wie soeben beschrieben muss angewendet werden, wenn korpusbasiert untersucht werden möchte, wie oft und in welchen Regionen das Verb *telefonieren* die Präposition *mit* regiert. Gemäss Ammon et al. (2004, 787) wird *telefonieren* „in der Bedeutung ‘jmdn. anrufen’ in CH auch ohne Präposition und mit Dativobjekt verwendet, gemeindt. mit der Präposition *mit* und Dativobjekt“. Das Vorhandensein der Präposition *mit* lässt sich mit entsprechenden Abfragen am Korpus untersuchen. Ein anderer Fall liegt bei *anrufen* vor, wo die problematischen Annotationen die Suche nach Akkusativ- resp. Dativobjekten praktisch verunmöglichen (vgl. Kapitel 4.3.2).

Insgesamt gibt es 3982 Treffer zum Lemma *telefonieren/telephonieren* ([(lemma="tele(f|ph)onieren")]). Nur in einem Fall wird *telephonieren* in der alten Schreibweise mit *ph* verwendet.

Davon werden, bei einer angenommen Spannweite von je 0 bis 5 Wörtern zwischen *telefonieren* und *mit*, 698 Fälle mit der Präposition *mit* realisiert. Die kombinierte Abfrage nach der Präposition *mit* vor oder hinter dem Verb *telefonieren* lautet folgendermassen:

```
([lemma="tele(f|ph)onieren"] []{0,5} [word="mit"]
[rfpos="(PRO.Pers.*Dat.*|ART.*Dat.*|N.*Dat.*)"] [word!="Handy|Natel|Mobiltelefon|Gerät"])
| ([word="mit"] [rfpos="(PRO.Pers.*Dat.*|ART.*Dat.*|N.*Dat.*)"]
[word!="Handy|Natel|Mobiltelefon|Gerät"][]{0,5} [lemma="tele(f|ph)onieren"]) within c
```

In dieser Abfrage wurden die falsch positiven Dative, die die Funktion eines Instrumentalis haben (*mit dem Handy/Natel/Mobiltelefon* etc.), ausgeschlossen. Wahrscheinlich gäbe es noch ein paar weitere technische Geräte, mit denen telefoniert werden kann.

Der Chi-Quadrat-Wert von 26.4 offenbart, dass *telefonieren* in Österreich signifikant häufiger mit der Präposition *mit* verwendet wird und diese Konstruktion in den Schweizer Zeitungen untervertreten ist.

Wenn *telefonieren* nicht die Präposition *mit*, sondern lediglich ein Dativobjekt regiert, müssen Präpositionen in der CQP Syntax über mehrere Positionen hinweg ausgeschlossen werden. Für den Fall, dass das Dativobjekt hinter dem Verb steht, ergeben sich 44 Treffer:

[(lemma="tele(f|ph)onieren")] [rfpos!="APPR.*"]{0,3}
[rfpos="(PRO.Pers.*Dat.*|ART.*Dat.*|N.*Dat.*)"] within c

Auch hier führen viele Clause Fehler zu nicht relevanten Belegen. Zusätzlich befinden sich darunter *false positives* wie *jdm. hinterher telefonieren* oder *jdm. nach telefonieren*. Eine manuelle Durchsicht des Recalls ergibt 3 relevante Treffer aus der Schweiz und 1 aus Liechtenstein.

Wenn das Dativobjekt vor dem Verb steht, gestaltet sich die Abfrage komplizierter:

<c> [rfpos!="APPR.*" & c & !c1]* [c1]*
[rfpos="(PRO.Pers.*Dat.*|ART.*Dat.*|N.*Dat.*)"] [rfpos!="APPR.*" & c & !c1]*
[c1]* [(lemma="tele(f|ph)onieren") & c & !c1] within c

Die 37 Treffer müssen wiederum von Hand kontrolliert werden. Es verbleiben 3 relevante Treffer aus der Schweiz, alle aus dem St. Galler Tagblatt.

Die geringen absoluten Zahlen (insg. 7, davon 6 aus der Schweiz) lassen den Schluss zu, dass die dialektale Verwendung des Verbs *telefonieren* ohne Präposition *mit* (noch) nicht in den Gebrauchsstandard der Schweizer Presse eingegangen ist. Allerdings ist der Referenzwert (18 Realisierungen von *telefonieren* mit der Präposition *mit* in der Schweiz) ebenfalls sehr klein.

## 3.3 Satzstrukturen

### 3.3.1 Verwendung des *am*-Progressivs

Gemäss Van Pottelberge 2004 (zitiert bei Dürscheid/Hefti 2006, 139) ist „der *am*-Progressiv in der Schweiz viel häufiger [...] als in Deutschland“ (vgl. zum *am*-Progressiv auch Elspaß 2010, 131-134). Eine Überprüfung dieser Beobachtung am Zeitungskorpus, das die Basis der „Variantengrammatik des Standarddeutschen“ bildet, gestaltet sich relativ einfach: [lemma="am"] [rfpos="VINF.*"]. Es resultieren 215 Treffer (darunter einige Tippfehler mit kleingeschriebenem Infinitiv und zwei falsch positive Ortsangaben *am Gurten*) mit einer höchst signifikant häufigeren Verwendung in der Schweiz und Liechtenstein (Chi Quadrat = 28.9). Das Korpus bestätigt somit Van Pottelberges Hypothese.

### 3.3.2 *n-jährig* in prädikativer Funktion

Die bisher in der Literatur nicht dokumentierte, aber innerhalb des Projekts „Variantengrammatik“ angestellte Beobachtung, dass nur in der Schweiz das Adjektiv *jährig* in Kombination mit einem Kopulaverb (*sein*) auftritt, lässt sich in zwei Schritten untersuchen. Zuerst wurde nach Konstruktionen der Form *ist n-jährig* (mit Bindestrich) gesucht:

<c> []* [rfpos="VFIN.Sein.*"] []{0,4} [word=".*-j.{1,3}hrig"][rfpos!="VPP.*"] </c>

Da mit rfpos, anders als mit pos, der unter VAFIN beide Auxiliarverben *sein* und *haben* zusammenfasst, das Auxiliarverb *sein* klar identifiziert werden kann, wurde die Suchanfrage über den RFTagger durchgeführt. Um das in Kapitel 4.1.3 diskutierte Umlautproblem zu umgehen, wurde der Umlaut *ä* durch drei optionale Zeichen ersetzt. Es resultieren 33 Treffer (davon zwei identische Belege mit falschem Umlaut aus der Neuen Luzerner Zeitung) – stets Zahlen mit Bindestrich –, die alle aus der Schweiz stammen.

Um zusätzlich die in Worten ausgedrückten Zahlen zu finden, die vor dem Adjektiv *jährig* stehen, wurde, wiederum mit dem RFTagger, eine zweite Abfrage getätigt, wobei die nicht relevanten Kompositionen *volljährig, völljährig* (Tippfehler), *minderjährig, ganzjährig, halbjärig, mehrjährig* und *langjährig* manuell ausgeschlossen wurden:

<c> []* [rfpos="VFIN.Sein.*"] []{0,4} [word=".*j.{1,3}hrig" & word!=
"(v(olö)ll|minder|ganz|halb|mehr|lang)j.{1,3}hrig"][rfpos!="VPP.*"] </c>

Die 55 resultierenden Treffer erweisen sich ebenfalls als klaren Frequenzelvetismus (Chi Quadrat = 964), wodurch sich die Hypothese, dass in Bezug auf das Alter von Menschen die prädikative Verwendung von *n-jährig* in Kombination mit einem Kopulaverb nur in der Schweiz möglich ist, bestätigt (vgl. Powerpointpräsentation „Variantengrammatik des Standarddeutschen“).

### 3.3.3 Temporaladverb *bereits* im Vorfeld

Dürscheid/Hefti (2006, 143f.) weisen darauf hin, dass nur in der Schweiz das Temporaladverb *bereits* im Vorfeld stehen könne (Bsp. *Bereits hat…*).

Da die Satzanfangsposition von *Bereits* vorgegeben ist, bereitet eine korpusbasierte Untersuchung keine grossen Schwierigkeiten. Ein Vergleich der Resultate, die der TreeTagger resp. der RFTagger zu Tage fördern, zeigt, dass, zumindest in diesem spezifischen Fall, wahrscheinlich aber allgemein bei der Annotation von Verben, der RFTagger in Bezug auf die Verben eine bessere Precision bietet.

TreeTagger:

<field_category="VF"> [word="Bereits"] </field_category> [pos="VVFIN|VAFIN|VMFIN"]

(285 Treffer, davon viele *false positives*)

RFTagger:

<field_category="VF"> [word="Bereits"] </field_category> [rfpos = "VFIN.*"]

(234 Treffer, davon alle korrekt).

Eine Disambiguierung der Tagger (<field_category="VF"> [word="Bereits"] </field_category> [pos="VVFIN|VAFIN|VMFIN" & rfpos!="VFIN.*"]) führt zu 60 Treffern, bei denen der TreeTagger in den meisten Fällen fälschlicherweise ein Verb annotiert, obwohl es sich nicht um ein Verb handelt.

Die Distribution bestätigt auf jeden Fall einen eindeutigen grammatischen Frequenzhelvetismus mit einer relativen Frequenz von 9.26 für die Schweiz und von 4.03 für Liechtenstein (Chi Quadrat = 4167).

### 3.3.4 Ellipse des Platzhalter *-es* im Vorfeld in der Konstruktion *Kommt hinzu/dazu*

Ebenfalls das Vorfeld betrifft eine weitere Konstruktion, der Dürscheid/Hefti (2006, 144) den Status einer grammatischen Variante zuschreiben, die nicht nur im mündlichen, sondern auch – die unten dokumentierte Korpusabfrage bestätigt dies – im schriftlichen Sprachgebrauch Verwendung findet. Es handelt sich um die Ellipse des Platzhalter *–es* in den Konstruktionen *Kommt hinzu/dazu, dass* oder *Kommt hinzu/dazu:*.

Diese lassen sich mit folgender Abfragesyntax formalisieren:

<s> [(word="Kommt")][word="(hin|da)zu"][rfpos="SYM.Pun.*"]

Wichtig ist, dass der Satzanfang über den Operator <s> definiert wird, da dieser einen Doppelpunkt vor *Kommt* ausschliesst.

Von den 91 Belegen sind 78 aus der Schweiz, 2 aus Österreich, 1 aus Liechtenstein und 10 aus Deutschland (davon 8 aus den Regionen Südwest und Südost). Die Konstruktion erweist sich somit als eine im Süden des deutschen Sprachraums gebräuchliche Variante mit einer höchst signifikant häufigeren Verwendung in der Schweiz (Chi Quadrat = 1561).

### 3.3.5 Nebensatz mit Verberststellung bei emotional-bewertendem Prädikat in Satzanfangsposition: *Gut/Schön/Toll/Schade*, + finites Verb

Das Phänomen, dass ein übergeordneter Satz, der auf einen Kopulasatz zurückzuführen ist, auf ein Prädikativum reduziert wird und von einem Nebensatz ohne Konjunktion (*dass*) gefolgt wird, wurde in einem Aufsatz von Andreas Lötscher aus dem Jahr 1997 (Lötscher 1997) erstmals aufgegriffen und ist bei Dürscheid/Hefti (2006, 140-143, 147f., 154f.) ausführlich beschrieben und mit ersten Untersuchungen zur Akzeptabilität in Deutschland und der Schweiz belegt (vgl. dazu auch den Abschnitt in Dürscheid et al. 2011, 130f.). Unter der Bedingung, dass die Resultate manuell ausgezählt werden, lässt sich diese Konstruktion am Korpus untersuchen. Vorauszuschicken ist, dass es wahrscheinlich weitere emotional-bewertende Adjektive gäbe, die die Satzanfangsposition einnehmen können. Die unternommene Suche beschränkt sich auf die drei positiv bewertenden Adjektive *gut, schön* und *toll* und das Bedauern ausdrückende Bewertungslexem *schade*.

Gesucht wurde eine Sequenz, in der auf die vier zur Auswahl stehenden Adjektive ein Komma und darauf ein finites Verb folgen. Wiederum ist es wichtig, den Satzanfang in die Abfrage einzubeziehen, da sonst *false positives* resultieren (z.B. direkte Reden, *Toll* als Eigenname).

<s> [word="(Gut|Schön|Toll|Schade)"] [rfpos="SYM.Pun.Comma"] [rfpos="VFIN.*"]

Eine Suche mit fakultativem Komma ist unmöglich, da viel zu viele Belege gefunden werden, die nicht mit der gesuchten Konstruktion übereinstimmen. Auch unter den 101 Treffern, die mit obligatorischem Komma gefunden werden, sind, wie eine manuelle Durchsicht ergibt, 90% *false positives*. Hauptsächlich lassen sich drei verschiedene Ursachen dafür ausmachen.

Zum Ersten enthalten die Zeitungsbelege bei Kommunikationsverben und Verben wie *denken*, *meinen*, *finden* etc. öfter direkte Reden, die nicht mit Anführungs- und Schlusszeichen markiert sind (Bsp. *Toll, freuten sich die Empfänger, fragten sich aber:* [marchcrawlfour1218_87]).

Zum Zweiten werden damit auch Sätze gefunden, die im Nebensatz die Partikel *doch* enthalten und auch im bundesdeutschen Standarddeutsch möglich sind

(Bsp. *Schade, hatten die Mädchen doch erst vor kurzem einen Namen gefunden:* [marchcrawlfour2284_239]).

Zum Dritten haben die satzeinleitenden Bewertungslexeme in mehreren Fällen, denen im Sinn von Koch/Oesterreicher (1994) ein eher konzeptionell mündlicher Status zuzu-

weisen ist, eine adhortative Bedeutung (Bsp. *Gut, nehmen wir die Bahn jetzt mal beim Wort*. [marchcrawlone1388_28]).

Da diese Fälle gemäss meiner Einschätzung nicht mittels einer entsprechenden Abfrage ausgeschlossen werden können, habe ich die 101 Treffer manuell durchgesehen. Unter ihnen finden sich 9 gesuchte Sequenzen mit faktitivem Nebensatz, allesamt aus der Schweiz.

Die wenigen positiven Treffer aus der Schweiz lassen den Schluss zu, dass diese konzeptionell mündliche Satzkonstruktion (noch) nicht in den Gebrauchsstandard der Schweizer online Zeitungen eingegangen ist.

In Vorgriff auf Kapitel 4.3 sei an dieser Stelle darauf hingewiesen, dass der RFTagger, weil er diese Schweizer Konstruktion nicht kennt, in 7 dieser 10 von Hand aussortierten Fälle das einleitende *Gut* fälschlicherweise als N.Reg.Nom.Sg.Neut. annotiert und nicht als ADJD. Eine Suche mit der Bedingung, dass ein Adjektiv am Satzanfang steht, würde diese 7 Belege somit verheerenderweise nicht finden. Diese Problematik, dass nicht bundesdeutsche Varianten aus annotiationstechnischen Gründen „durch’s Netz fallen“, bestätigt genau jene Gefahr, auf die Belica et al. (2010, 461) aufmerksam machen: „there is a danger that without realizing we may end up not analyzing observations of language use but also the tagging tool, the theory and language model behind it, or possibly its imperfect implementation“. Unter dieser Bedingung ist es nötig, die Adjektive über eine Word Abfrage zu formulieren.

Anstelle des Prädikativums kann der übergeordnete Hauptsatz auch ein transitives Verb enthalten (Bsp. *Ich finde [es] gut, kommst du morgen*) (vgl. Dürscheid/Hefti 2006, 140). In der Suchsyntax müssen Partikel und Adverbien (*doch, aber*) im Nebensatz ausgeschlossen werden, da diese Konstruktionen auch im bundesdeutschen Standard möglich sind. Zudem dürfen die finiten Verben des Haupt- und Nebensatzes nicht im Konjunktiv stehen und das Verb des Nebensatzes darf kein Kommunikationsverb sein. Von letzteren konnten nicht alle von Hand ausgeschlossen werden. Unter den 34 Treffern der folgenden Abfrage befinden sich auch zahlreiche andere Kommunikationsverben.

[lemma="sein|finden" & rfpos!=".*Subj.*"] [word="es"]? [word="(gut|schön|toll|schade)"] [rfpos = "SYM.Pun.Comma"] [rfpos="VFIN.*" & lemma!="sagen|meinen||finden|urteilen|fragen|behaupten|betonen|versichern"& rfpos!=".*Subj.*"] [pos!="(ADV|KON)" & c & !c1]* </c> (34 Treffer)

Lediglich in einem Beleg aus der Schweiz (marchcrawlfour502_27) findet sich die gesuchte Konstruktion.

### 3.3.6 Verwendung des Relativpronomens *welcher/welche/welches*

Im Verlauf des Seminars „Korpuslinguistik“ wurde wiederholt gemutmasst, dass in der Schweiz die Relativpronomen *welcher/welche/welches* häufiger verwendet würden als in den anderen deutschsprachigen Regionen. Eine entsprechende Korpusabfrage unterstützt diese Hypothese:

[rfpos="SYM.Pun.Comma"] [lemma="(welche(r|s)?)" & rfpos="PRO.Rel.*"]

Insgesamt wird ein Relativsatz in 37618 Fällen mit den Pronomen *welcher/welche/welches* realisiert mit signifikant hohen Frequenzen in der Schweiz, Liechtenstein und Südtirol (Chi Quadrat = 23975). Ein Vergleich mit der Verwendung der Relativpronomen *der/die/das* in diesen Regionen ist leider nicht möglich, da eine Verarbeitung der mehr als 100000 Belege die rechnerischen Kapazitäten des Korpus übersteigt.

# 4 Evaluation der aufgetretenen Schwierigkeiten und Aufzeigen der Konsequenzen für Korpusabfragen

Im Verlauf meiner Abfragen im Korpus „Variantengrammatik des Standarddeutschen“ bin ich an verschiedenen Stellen Schwierigkeiten, Limitierungen und zum Teil auch Fehlern des aktuellen Stands des Korpus (MarchCrawl) begegnet, welche die korpuslinguistischen Untersuchungen von Variationsphänomenen erschweren und die Ergebnisse unter Umständen verfälschen können. In diesem Kapitel wird deshalb der Versuch unternommen, die aufgetretenen Problemstellungen unter verschiedenen Gesichtspunkten zu bündeln und anhand von Beispielabfragen zu illustrieren, wobei bei bestimmten Phänomenen eine Kombination von mehreren Problembereichen auftreten kann (z.B. falsche Annotation und falsche Lemmatisierung). Das Ziel dieser Evaluation liegt darin, das Bewusstsein für die Restriktionen des Korpus zu schärfen und Konsequenzen für künftige Korpusabfragen aufzuzeigen. Keinesfalls erhebt diese Zusammenstellung Anspruch auf Vollständigkeit. Sie ist das Ergebnis einer subjektiven Recherchetätigkeit, versucht aber immer, so weit als möglich, durch das Aufstellen von Hypothesen über mögliche Analogien den Blick vom Einzelphänomen auf einen grösseren Themenbereich auszuweiten.

## 4.1 Fehlerhaftes Datenmaterial und Probleme beim Crawling

Die Datengrundlage des Korpus „Variantengrammatik des Standarddeutschen“ bilden Artikel, die aus den online Ausgaben der 84 in das Projekt eingeschlossenen Zeitungen über einen sog. *Crawlingprozess* automatisch eingelesen wurden. Fehler auf dieser „untersten“ Schicht sind insofern von Bedeutung, als sie sich negativ auf die Arbeit der Tagger und als Folge davon direkt auf die linguistische Untersuchung auswirken können.

### 4.1.1 Tipp- und Grammatikfehler in den Zeitungstexten

Eine erste Ausprägung von fehlerhaftem Datenmaterial sind Tipp- oder Grammatikfehler in den Zeitungstexten. Da diese Fehler kaum systematisch und deshalb äusserst niedrigfrequent sind, sind sie statistisch vernachlässigbar. Während „simple“ Tippfehler (Bsp. *völljährig* statt *volljährig* [marchcrawlfour681_70]) meist unproblematisch sind, können fehlerhafte Wortformen in den Zeitungstexten die morpho-syntaktischen Tagger

in ihrem Part-of-speech-Tagging-Prozess „verirren“, wodurch Wörter im Umfeld des betreffenden Wortes falsch annotiert werden können.

Als Beispiel für diesen zweiten Fall von Fehlern soll die folgende Belegstelle dienen.

*Jetzt schneit es sogar, da fühle ich mich noch mehr Hause*. (marchcrawlone26_63)

Da in diesem Satz die Präposition *zu* fehlt, fasst der RFTagger das Token *Hause* als finites Verb auf.

Umgekehrt kann der Tagger im folgenden Satz einen Fehler im Quellentext fehlerfrei umgehen.

*An einer Kreuzung stieß der 45-Jährige mit dem Pkw eines 77-jährigen Pensionist aus Deutschland zusammen*. (marchcrawltwo814_99)

Hier erkennt der RFTagger aus dem Kontext (unbestimmter Artikel *eines*) korrekt, dass ein Genitiv erforderlich ist, obwohl das Genitiv Suffix *–en* der schwachen Deklination beim Token *Pensionist* fehlt.

### 4.1.2 Tokenisierung

Neben den Fehlern aus journalistischer Hand können auch im Prozess der *Tokenisierung*, d.h. der Segmentierung des Textes in einzelne Tokens, Fehler auftreten. Gleich drei dieser Worttrennungsfehler finden sich beispielsweise im folgenden Satz.

*W�hrend desKurses ist jeweils eine einst�ndige w�chentliche Austausch- undFrage-runde im GemeindehausSelbitz geplant*. (marchcrawlthree2181_8)

Wie die beiden Tagger mit den zusammengeschriebenen Tokens umgehen, zeigt die untenstehende Tabelle. Die fehlerhaften Annotationen sind rot markiert. Zusätzlich bereitet der falsch kodierte Umlaut in *Während* Probleme (vgl. dazu die Ausführungen unten in Kapitel 4.1.3). In diesem konkreten Beispiel wirken sich die falschen Annotationen nicht auf die Annotaion benachbarter Tokens aus.

| | W�hrend | desKurses | [...] | undFragerunde | [...] | GemeindehausSelbitz |
|---|---|---|---|---|---|---|
| Crawling Problem | Umlaut | Tokenisierung | | Tokenisierung | | Tokenisierung |
| TreeTagger | NE | NN | | NN | | NN |
| RFTagger | N.Reg. Nom.Sg.Masc | ADJA.Pos. Nom.Sg.Neut | | ADJA.Pos. Nom.Sg.Fem | | N.Reg. Dat.Sg.Masc |

Die Ursache der Tokenisierungsfehler dürfte in diesem speziellen Fall (und womöglich auch in weiteren Fällen) im Layout des Artikels in der Frankenpost liegen (http://www.frankenpost.de/lokal/naila/naila/3-Fragen-an;art2443,1897373). Der betref-

fende Satz steht in einem Informationskasten unterhalb des Hauptartikels. Die Zeilenumbrüche innerhalb dieses Kastens werden vom Tagger nicht als Wortgrenzen erkannt.

### 4.1.3 Kodierung von Umlauten und Sonderzeichen

Von grösserer Bedeutung als die bisher genannten scheinen mir die Probleme in der Zeichenkodierung von Diakritika (insb. Umlaute) und anderen Sonderzeichen (z.B. dem deutschen ß) zu sein. Die deutschen Umlaute werden in gewissen Fällen und in ausgewählten Zeitungen besonders gehäuft nicht als Umlaute erkannt. Dies führt einerseits dazu, dass bei einer Abfrage, die einen Umlaut enthält, aus den hauptsächlich betroffenen Zeitungen möglicherweise viele korrekte Belege nicht gefunden werden (*false negatives*). Andererseits kann die fehlerhafte Kodierung eines Umlauts zu falschen Annotation von Wörtern im Umfeld dieses Wortes führen und dadurch unkontrollierbare, im Einzelfall nicht mehr nachvollziehbare Auswirkungen haben. Durch die Rekonstruktion von zwei falsch als Genitiv annotierten Fällen von *Pensionist*, denen je ein Wort mit einem Umlaut vorangeht (*während* resp. *öffnete*), bin ich überhaupt erst auf die Problematik aufmerksam geworden (vgl. marchcrawltwo2052_79 und marchcrawltwo2053_35). Welche Auswirkungen die falsch kodierten Umlaute nach sich ziehen können, soll am Beispiel der Präposition resp. Subjunktion *während* illustriert werden.

Die Suche nach [lemma="während"] ergibt 217677 Treffer, davon entfallen 216365 auf Tokens *während* [word="während"]. Umgeht man mittels einer RegEx Abfrage den Umlaut *ä* und erlaubt zwischen dem *w* und dem *h* 1-3 beliebige Zeichen [word="(W|w).{1,3}hrend"], findet man, neben wenigen Tippfehlern von *während* und 7 falsch positiven *Wehrend,* 2275 zusätzliche Tokens *W*  *hrend* und *w**hrend*, von denen, wie die Abfrage [(word="(W|w).{1,3}hrend") & lemma!="während"] zeigt, keines dem Lemma *während* zugeordnet wird. Diese falsch kodierten *während* besitzen ein eigenes Lemma *W**hrend*. Beim Taggen entstehen dadurch falsche Annotationen (z.B. als N.Reg.*, VFIN oder ADJD). Von diesen 2275 Fällen werden fälschlicherweise 514 als Nomen identifiziert, wie die Abfrage [(word="(W|w).{1,3}hrend") & (rfpos="N.Reg.*")] zeigt. Dies kann weitreichende Auswirkungen auf die Annotation von Satzteilen in der Nähe des betreffenden Wortes haben.

Wenn auch vom Total der Realisierungen von *während* der Umlaut *ä* in lediglich ca. 1% der Fälle falsch kodiert ist, ist das Verhältnis von korrekten und fehlerhaften *wäh-*

*rend* in drei Zeitungen – je eine aus Deutschland, Österreich und der Schweiz – besorgniserregend, wie die untenstehende Tabelle verdeutlicht.

Als korrekte *während* verstehe ich jene, die dem Lemma *während* zugeordnet sind [lemma="während"], als fehlerhaft diejenigen *während* mit falsch kodiertem Umlaut, die nicht dem Lemma *während* zugewiesen sind [(word="(W|w).{1,3}hrend") & lemma!="während"].

| Zeitung | Korrekte *während* | Fehlerhafte *während* |
|---|---|---|
| A-Krone | 2842 (77%) | 839 (23%) |
| CH-Neue Luzerner Zeitung | 650 (68%) | 305 (32%) |
| D-Frankenpost | 1784 (72%) | 703 (28%) |
| Gesamtes Korpus | 217677 (99%) | 2311 (1%) |

Erschwerend kommt hinzu, dass es neben dem soeben beschriebenen Kodierungsfehler ï¿½ mindestens einen weiteren Kodierungsfehler (*ï¿½.*) gibt, der sowohl die Umlaute, als auch das ß (und möglicherweise weitere Sonderzeichen) betrifft. Eine Suche nach allen Wörtern, die diese fehlerhafte Kodierung enthalten [word=".*ï¿½.*"], zeigt, dass zwei österreichische Zeitungen von diesem Fehler betroffen sind: A-Kleine Zeitung (179706 Tokens) und A-Niederösterreichische Nachrichten (83822 Tokens).

Die Auswirkungen dieser beiden Arten von Kodierungsfehlern werden im Folgenden an einigen Beispielen demonstriert. Wiederum sind die falschen Annotationen in der Tabelle rot markiert.

*Wï¿½hrend der Pensionist in der Speisekammer war, stahl der Unbekannte die Brieftasche*. (marchcrawltwo2053_35, Kleine Zeitung)

| | Wï¿½hrend | der | Pensionist |
|---|---|---|---|
| Crawling Problem | Umlaut | | |
| TreeTagger | ADJD statt KOUS | ART | NN |
| RFTagger | N.Reg. Nom.Sg.Neut | ART.Def. Gen.Sg.Fem | N.Reg. Gen.Sg.Fem |

Weil der RFTagger das Token *während* wegen des falsch kodierten Umlauts als Nomen auffasst, wird *der Pensionist* fälschlicherweise als Genitiv Singular feminin annotiert. Der ursprüngliche Kodierungsfehler wirkt sich hier auf benachbarte Tokens aus.

Ein identischer Fall liegt im folgenden Beispiel aus der Kronenzeitung vor.

*W*hrend der Gr*ne Sicherheitssprecher Peter Pilz die sofortige Freilassung der Aktivisten fordert, da "unschuldige und unverd*chtige Menschen" im Gef*ngnis s**en, h*lt die Justiz ihre weitere Anhaltung f*r gerechtfertigt.*
(marchcrawlfour16_82)
*Während* wird hier vom RFTagger wegen dem Kodierungsproblem nicht als Subjunktion, sondern als Nomen annotiert. Als Folge davon wird die Nominalphrase *der Grüne Sicherheitssprecher Peter Pilz* als Genitiv Singular feminin betrachtet.

Ein analoger Fall, jedoch mit einem anderen Kasus, findet sich in einem Artikel aus der Frankenpost.
*W*hrend die alte Tradition in Japan zu verschwinden droht, gehen die Europ*er diesem Hobby mit unheimlich viel Engagement nach.* (marchcrawlthree1451_106)
Hier wird als Folge der falschen Annotation von *während* als Nomen die darauf folgende Nominalphrase *die alte Tradition* in den Akkusativ (statt in den Nominativ) gesetzt.

Im folgenden Fall aus der Kleinen Zeitung tritt das Problem der Umlautkodierung im Verb auf. Wiederum wirkt sich dies auf die Annotation des Satzumfeldes aus, wie die Tabelle deutlich macht.
*An der Bergstation Steinermandl ï¿½ffnete der Pensionist wie vorgesehen den Sicherungsbï¿½gel, um anschlieï¿½end auszusteigen.* (marchcrawltwo2052_79)

| | ï¿½ffnete | der | Pensionist |
|---|---|---|---|
| Crawling Problem | Umlaut | | |
| TreeTagger | NN | ART | NN |
| RFTagger | N.Reg.Nom.Sg.Fem | ART.Def.Gen.Sg.Fem | N.Reg.Gen.Sg.Fem |

| | um | anschlieï¿½end | auszusteigen |
|---|---|---|---|
| Crawling Problem | | Umlaut | |
| TreeTagger | KOUI | ADJD | VVIZU |
| RFTagger | CONJ.SubInf. | N.Reg.Acc.Sg.Neut | VINF.Full.zu |

Insgesamt sind somit (mindestens) drei österreichische Zeitungen, eine deutsche und eine Schweizer Zeitung im Hinblick auf Abfragen von Wörtern, die einen Umlaut oder ein anderes Sonderzeichen enthalten, als kritisch zu betrachten. Falls dieses Kodierungsproblem nicht behoben werden kann, muss man sich bei zeitungs- oder regionenspezifischen Untersuchungen bewusst sein, dass bei der Abfrage eines bestimmten Phänomens im Extremfall die Fehlerrate grösser als 5% sein kann. Möchte

man z.B. analysieren, wie häufig die Präposition *während* mit Genitiv oder Dativ verwendet wird, können Abfragen, an denen die fünf genannten Zeitungen beteiligt sind, verfälschte Resultate liefern, da bei einer Suche nach dem Lemma *während* sehr viele Tokens mit dem beschriebenen Kodierungsfehler nicht gefunden werden (*false negatives*). Die vorgeschlagene Umgehungsstrategie über die Optionalität von Zeichen ist sehr aufwändig und kann unter Umständen zu zu vielen falsch positiven Ergebnissen führen. Darüber hinaus besteht die Gefahr, dass als Folge dieser Kodierungsfehler Varianten falsch annotiert sein können, wie am Beispiel von *Pensionist* demonstriert wurde.

### 4.1.4 Nicht redaktionelle Inhalte

Ebenfalls in den Problembereich des Crawlings fällt die Unterscheidung zwischen redaktionellem Inhalt der Artikel und Funktionen auf der Webpage. Zum Beispiel werden Links auf der Webpage wie *Artikel per E-Mail versenden* oder *Twitter(n)* ebenfalls tokenisiert und in die Datenbank aufgenommen. Die Tokens *E-Mail* und *Mail* treten ausserdem sehr häufig im Adressblock von Kleinanzeigen oder bei der Vorstellung lokaler Betriebe mit Kontaktangaben auf. In besonderem Ausmass betroffen sind die in der Tabelle aufgeführten Zeitungen.

| Zeitung | *E-Mail* (Rel. Freq.) | *Mail* (Rel. Freq.) | *Twitter(n)* (Rel. Freq.) |
|---|---|---|---|
| A-Krone | | 132 | |
| A-Kurier (Wien) | | | 195 |
| A-Salzburger Fenster | | | 3276 |
| A-Vorarlberger Nachrichten | 1610 | | |
| BELG-Grenz-Echo | 146 | | |
| CH-Bernerzeitung.ch | 2658 | | |
| D-Märkische Allgemeine | 172 | | |
| D-Neue Westfälische | 445 | | |
| D-Ostfriesen-Zeitung | 1006 | | |
| D-Schweriner Kurier | 153 | | |
| LUX-Tageblatt | | 2228 | |

Da das Korpus sowohl für *E-Mail* als auch für *Mail* sehr viele nicht relevante Quellentexte enthält, sind statistische Untersuchungen zur Verwendung des Genus (*das/die E-Mail/Mail*) oder zu einer allfälligen Präferenz für die Variante *E-Mail* oder *Mail* in den verschiedenen Regionen unter den aktuellen Voraussetzungen sinnlos, insbesondere da der RFTagger ohne Zusatzinformation aus der Umgebung *E-Mail/Mail* durchwegs als feminin annotiert, was dem bundesdeutschen Standard entspricht.

## 4.2 Teilweise fehlende oder unvollständige Lemmatisierung, insb. von nicht bundesdeutschen Varianten

### 4.2.1 Orthographische Varianten

Der Umstand, dass in der Schweiz und in Liechtenstein die Schreibung eines Doppel-s mit *ß* nicht verwendet wird, kann zu Problemen bei der Lemmatisierung von Tokens führen, da die Tagger mit dem bundesdeutschen Standard trainiert wurden. Der Schweizer Linguist oder die Schweizer Linguistin muss sich bewusst sein, dass Abfragen, bei denen das Lemma in Schweizer Schreibweise mit *ss* eingegeben wird, in der Regel zu keinen Treffern führen. Bei einer Abfrage mit *ß* kann aber der Fall auftreten, dass nicht alle Flexionsformen der Schweizer Schreibweise mit *ss* gefunden werden. Ein Beispiel für diese Problematik liegt in der unvollständigen Lemmatisierung von *Fuß* vor. Während die Schweizer Flexionsformen *Fuss, Fusses, Fusse, Füssen* korrekt dem Lemma *Fuß* zugeordnet werden, wird der Nominativ Plural *Füsse* nicht erkannt. Im Fall von *Spaß* werden ebenso die Singularformen *Spass* und *Spasses* dem Lemma *Spaß* zugeordnet, sowohl der Nominativ Plural *Spässe* als auch der Akkusativ Plural *Spässen* jedoch nicht, obwohl eine Suche nach den entsprechenden Words Treffer findet.

Als Konsequenz dieser uneinheitlichen und unvollständigen Lemmatisierung von orthographischen Varianten ist es notwendig, bei Wörtern mit *ß* alle Abfragen eines Lemmas einzeln mit einer Suche über Word gegen zu prüfen.

### 4.2.2 Lexikalische Varianten

Diatopische Variation auf der lexikalischen Ebene bildet nicht das Erkenntnisinteresse des Projekts „Variantengrammatik des Standarddeutschen“, da sie bereits vom „Variantenwörterbuch des Deutschen“ (Ammon et al. 2004) abgedeckt wird. Es lohnt sich aber, sich der Limitierungen des aktuellen Korpusstands bewusst zu sein, da an variantengrammatischen Untersuchungen lexikalische Varianten beteiligt sein können.

Die Lemmatisierung von Helvetismen (mit Austriazismen dürfte es sich ähnlich verhalten) ist zur Zeit uneinheitlich. So verfügen beispielsweise *Velo* und *Abwart* über ein eigenes Lemma, das jedoch nicht alle Flexionsformen korrekt umfasst.

Im Fall von *Abwart* findet eine Suche nach [lemma="Abwart"] lediglich die 40 Singularformen *Abwart*, nicht aber die 7 Tokens, die auf den Genitiv Singular *Abwarts* oder die Pluralform *Abwarte* entfallen. Lediglich eine RegEx Abfrage [word="Abw.rt.?"], die sowohl Singular- als auch Pluralformen zulässt, fördert alle 47 Tokens zu Tage. Die

Distribution bestätigt *Abwart* als eindeutigen Helvetismus. Von den 47 Treffern finden sich zwar 4 in der Rheinischen Post, jedoch beziehen sich alle auf eine Figur *Veri, der Abwart (CH-Malters)*, also eine stereotype Schweizer Figur, die beim niederrheinischen Kabarettwettbewerb „Das schwarze Schaf“ mitmacht (vgl. http://www.rp-online.de/niederrhein-nord/moers/nachrichten/16-kandidaten-fuer-das-schwarze-schaf-1.1069125, http://www.rp-online.de/niederrhein-nord/kleve/nachrichten/talente-jagen-schwarzes-schaf-1.1069381, http://www.rp-online.de/niederrhein-sued/krefeld/nachrichten/dieter-nuhr-hat-jetzt-einen-vogel-1.2832232 <11.8.2013>).

Anders verhält es sich beispielsweise mit den Helvetismen *Pendenz*, dem dazugehörenden Adjektiv *pendent* und *parkieren*, die alle kein eigenes Lemma haben.
Im Fall von *parkieren* bestätigt das Korpus den Eintrag des „Variantenwörterbuchs des Deutschen“ (Ammon et al. 2004, 556). Zum Begriff *parkieren* in all seinen Flexionsformen finden sich 488 Einträge aus der Schweiz (relative Frequenz 20.56) und Liechtenstein (relative Frequenz 17.45).
Unter dem Frequency Breakdown von Lemma *parken* ist die schweizerische Variante *parkieren* nicht aufgeführt, sie hat aber auch kein eigenes Lemma. Leider kann aus dem Korpus nicht herausgelesen werden, welches Lemma für *parkieren* angenommen wird, es ist auf jeden Fall nicht <*unknown*>. Wiederum ergibt sich die Konsequenz, dass man die Variante *parkieren* kennen und über [word="parkier.*"] suchen muss.

Unter den gegebenen Umständen empfiehlt es sich in jedem Fall, stets unter dem Frequency Breakdown eines Lemmas zu überprüfen, ob alle möglichen Flexionsformen aufgeführt sind. Falls eine oder mehrere Formen fehlen oder falls zu einem Begriff gar kein Lemma zu finden ist, bleibt nur die Möglichkeit, lexikalische Varianten mittels einer RegEx Abfrage über Word, die alle möglichen Flexionsformen umfasst, zu untersuchen. Der Komfort einer Lemma Abfrage kann in diesen Fällen leider nicht ausgenutzt werden.

### 4.2.3 Morphologische Varianten

Die höchste Komplexitätsstufe wird bei der Untersuchung von morphologischen Varianten erreicht. Dies soll am Beispiel der unterschiedlichen Pluralbildungen von *Park* aufgezeigt werden.

Das „Variantenwörterbuch des Deutschen“ (Ammon et al. 2004, 556) verzeichnet für den Plural von *Park* die folgende diatopische Variation: „In der Bedeutung ‘größere Grünfläche in einer Stadt’ lautet der Plural in CH auch *Pärke*, gemeindt. *Parks*. In der Bedeutung ‘großräumiges naturbelassenes geschütztes Gebiet, z.B. Naturpark’ lautet der Plural in STIR auch *Parke*, gemeindt. *Parks*“.

Die Überprüfung dieser Konstellation im Korpus „Variantengrammatik des Standarddeutschen“ bestätigt, dass die Pluralvariante *Pärke* nur in der Schweiz verwendet wird: 56 mal in der Grundform ([word="Pärken?"]) und in 30 Fällen als Kompositum ([word=".+pärken?"]). Die schweizerische Pluralbildung *Pärke* wird jedoch nie auf das Lemma *Park* zurückgeführt ([word="Pärken?" & lemma="Park"] findet 0 Treffer). Es zeigt sich, dass *Pärke* zwar ein eigenes Lemma hat, dass aber die 13 Dativformen *Pärken* nicht auf das Lemma *Pärke* zurückgeführt werden, sondern unter dem eigenen Lemma *Pärken* auftreten ([word="Pärken?" & lemma!="Pärke"] und [lemma = "Pärken"] erzielen die gleichen 13 Treffer).

Analog verhält es sich mit dem Pluralparadigma der Komposita von *Pärke*. Von den 86 Kompositaformen ([word="(Pl.+p)ärken?"]) entfallen 67 auf das Lemma *-pärke* ([lemma="(Pl.+p)ärke"]) und 19 auf das Lemma *-pärken* ([lemma="(Pl.+p)ärken"]).

Zusätzlich zur problematischen Lemmatisierung der Schweizer Variante werden – damit wird nun auf das Kapitel 4.3 vorgegriffen – die Tokens *Pärke* meistens falsch annotiert. So fasst der RFTagger das maskuline Nomen im Plural in exakt der Hälfte der Fälle als femininen Singular auf ([word="Pärken?" & rfpos=".*Sg.Fem.*"] ergibt 28 von 56 Treffern). Von den 30 Komposita *-pärke* werden lediglich 8 fälschlicherweise als Singular annotiert, ebenfalls alle feminin ([word=".+pärken?" & rfpos=".*Sg.Fem.*"]).

Die Pluralform *Parke* tritt am häufigsten im Südtirol auf (relative Frequenz 8.16), es finden sich aber auch, in Differenz zum „Variantenwörterbuch des Deutschen“ (Ammon et al. 2004, 556), Belege in Nordostdeutschland (relative Frequenz 1.02). Die einfache Wortform *Parke* ist dabei mit 18 relevanten Treffern sehr viel seltener als die Verbindung in einem Kompositum (der Löwenanteil entfällt dabei auf *Naturparke* mit 159 und *Nationalparke* mit 35 Treffern).

Um die Komposita von *Parke* zu finden, muss die Abfrage präzisiert werden, indem die Eigennamen und zusätzlich die falsch positiven *Sparke* und *Falschparke* ausgeschlossen werden ([word="(Pl.+p)arke" & word!="Sparke" & rfpos!="N.Name.*" & word!="Falschparke"]). Es resultieren 231 relevante Komposita.

Anders als die schweizerische Pluralform *Pärke* (inkl. Komposita) hat die Südtiroler Variante *Parke* kein eigenes Lemma und wird auch nicht dem Lemma *Park* zugeordnet ([word="(Pl.+p)arke" & lemma= "(Pl.+p)arke"] liefert keine relevanten Treffer).

Die vorangehenden Analysen haben gezeigt, dass es nicht möglich ist, über eine Suche [lemma="(Pl.+p)ark" & rfpos="N.Reg.*.Pl.*"] alle Pluralvarianten von *Parks*, *Pärke* und *Parke* inkl. deren Komposita zu finden (*corpus-driven*). Diese Abfrage findet nur die Pluralformen *Parks* und Komposita (182 Types). Genau diejenigen Informationen, die eine variantengrammatische Untersuchung interessiert, nämlich die *Pärke* und *Parke* und Komposita, treten dabei nicht ans Tageslicht (*false negatives*). Die nichtbundesdeutschen Pluralformen müssen bereits bekannt sein und einzeln nach Word abgefragt werden, da, wie das Beispiel *Parke* zeigt, nicht alle Varianten einzeln lemmatisiert sind (*corpus-based*). Diese Ausgangslage birgt einmal mehr genau jene Gefahr, auf die Belica et al. in der bereits zitierten Stelle (2010, 461) hinweisen.
Das am Beispiel von *Park* demonstrierte Vorgehen ist generell bei der Untersuchung von morphologischen Varianten zu empfehlen.

## 4.3 Tagging: Falsche Annotationen

Falsche Annotationen stellen die grösste Herausforderung bei der korpusbasierten Recherche von Variationsphänomenen dar. Sie betreffen stets ein einzelnes Wort, können aber – und dies ist für variantengrammatische Untersuchungen besonders kritisch – die Abfrage einer ganzen Satzkonstruktion negativ beeinflussen. Besonders gravierend wirken sich die falschen Annotationen bei Untersuchungen zur Rektion, Valenz und der Reflexivität von Verben aus. Bei grossen Recalls, die nicht von Hand ausgezählt werden können, ist es in Extremfällen unter den gegebenen Umständen sogar unmöglich, fundierte korpusbasierte Aussagen zu machen.

Da Simone Ueberwasser bereits ausführlich auf die irreführenden Kasusannotationen bei Präpositionen hingewiesen hat (Ueberwasser 2013), konzentriere ich mich auf dieselbe Problematik in Bezug auf Pronomen.

Auf jeden Fall gilt es, sich stets bewusst zu sein, dass „[i]n linguistics, it is in general of crucial importance to construe annotation data as mere assessments (or ‘opinions’) of (human or automatic) annotators rather than as straightforward observations of language use.“ (Belica et al. 2010, 466).

### 4.3.1 Annotation von Reflexivpronomen: Reflexive Verwendung eines absoluten Verbs: *nerven / sich nerven*

Die Inversion zum Fall, in dem ein reflexives Verb absolut verwendet wird (vgl. *rentieren* in Kapitel 3.2.1), liegt vor, wenn absolute Verben reflexiv verwendet werden (Bsp. *sich nerven, sich ändern, sich erwarten, sich gewohnt/gewöhnt sein*) (vgl. Dürscheid/Hefti 2006, 137, zur Reflexivierungstendenz in Österreich vgl. Ziegler 2010 und Dürscheid et al. 2011, 132f.). Die Schwierigkeiten, die diese Varianten gegenüber der korpusbasierten Untersuchung der absoluten Verwendung von reflexiven Verben bereiten, werden im Folgenden am Beispiel von *nerven* resp. *sich nerven* durchexerziert. Zugrunde liegt dabei die Definition der Duden-Grammatik (2009b, 401) für die reflexiven Varianten von Verben der Gemütsbewegung.

[lemma="nerven"] ergibt 3762 Treffer, die Frequenzen sind in allen Ländern etwa ähnlich hoch. Anzumerken ist, dass dieser Ausdruck für den Sprachgebrauch in Zeitungen ziemlich dezidiert sein dürfte. Zum Vergleich: Die „mildere“ Form *sich ärgern* [lemma="ärgern"] wird im Korpus insg. 17187 mal verwendet. Dieser Vergleich ist auch im Hinblick auf das Annotationsproblem interessant. Da *sich ärgern* als reflexives Verb

kodifiziert ist, annotieren die Tagger die Reflexivpronomen korrekt. Anders sieht es aus bei der nicht bundesdeutschen reflexiven Variante *sich nerven*. Die grossen Divergenzen der beiden Tagger bei der Annotation der Pronomen in der näheren Satzumgebung des zur Diskussion stehenden Verbs mögen die folgenden beiden Beispielabfragen illustrieren, in denen sowohl mit dem TreeTagger als auch mit dem RFTagger nach Reflexivpronomen, die im näheren Umfeld hinter dem Verb *nerven* auftreten, gesucht wurde:

[lemma="nerven"] []{0,3} [pos="PRF"] within c 150 Treffer

[lemma="nerven"] []{0,3} [rfpos="PRO.Refl.*"] within c 83 Treffer

Bei beiden Taggern handelt es sich im grössten Teil der Fälle – zumindest gemäss der Kategorisierung der reflexiven Verben der Duden-Grammatik, auf die oben verwiesen wurde – nicht um Reflexivpronomen, sondern um Akkusativobjekte (Bsp. *Das nervt mich* anstatt *Ich nerve mich*).

Ich habe die Recalls beider Tagger manuell durchgesehen und bei beiden die identischen 23 Belege identifiziert, bei denen es sich um korrekte Reflexivpronomen im gesuchten Sinn handelt. Beim Rest handelt es sich um Personalpronomen im Akkusativ. Der RFTagger bietet in diesem Fall somit im Vergleich zum TreeTagger noch eine relativ höhere Precision. Von diesen 23 Treffern stammen mit einer Ausnahme aus dem Liechtensteiner Vaterland alle, wie erwartet, aus Schweizer Zeitungen.

Die untenstehende Tabelle verdeutlicht die Fehlerrate bei der Annotation der Reflexivpronomen (rot markiert) am Beispiel von *sich nerven* auf der Basis der beiden oben angegebenen Abfragen.

| | Total Treffer | Korrekte Reflexivpronomen | Personalpronomen im Akkusativ |
|---|---|---|---|
| RFTagger | 83 | 23 (28%) | 60 (72%) |
| TreeTagger | 150 | 23 (15%) | 127 (85%) |

Dass umgekehrt korrekte Reflexivpronomen fälschlicherweise als Personalpronomen annotiert wären und dadurch bei einer Suche nach Reflexivpronomen nicht gefunden würden, kann nahezu ausgeschlossen werden. Um dies zu überprüfen, habe ich über die Word Suche nach Tokens *sich* gesucht ([lemma="nerven"] []{0,3} [word="sich"] within c, 58 Treffer) und den Recall überprüft. Er enthält wiederum mehrere Fälle, in denen sich *sich* nicht auf *nerven* bezieht. Alle relevanten *sich* sind vom RFTagger korrekt als PRO.Refl. annotiert. *False negatives* sind aufgrund dieser stichprobenartigen Kontrolle nicht zu erwarten.

Die Suche nach einem Reflexivpronomen vor dem Verb *nerven* ergibt beim TreeTagger 134 Treffer ([pos="PRF"] []{0,3} [lemma="nerven"] within c), beim RFTagger 87 Treffer ([rfpos="PRO.Refl.*"] []{0,3} [lemma="nerven"] within c). Der RFTagger liefert auch hier einen präzisieren Recall. Wiederum wurde die optionale Range mit 0 bis 3 Wörtern zwischen dem Reflexivpronomen und dem Verb relativ klein gehalten, damit die Zahl derjenigen Fälle, in denen sich das Reflexivpronomen nicht auf *nerven*, sondern ein anderes reflexives Verb bezieht, möglichst gering ist. Trotzdem finden sich unter den Ergebnissen viele falsch positive Partizipien und Adjektive *genervt* (Bsp. *ist genervt, fühlt sich genervt, fragt sich genervt* etc.). Eine manuelle Durchsicht findet 4 Fälle, in denen das Reflexivpronomen vor *nerven* steht. Alle stammen, wie nicht anders erwartet, aus der Schweiz.

Beim Zustandspassiv *genervt sein* zeigt sich ein leicht anderes Bild. Das Auxiliarverb *sein* kann entweder vor oder hinter *genervt* stehen.
Vorne: [rfpos="VFIN.Sein.*"] []{0,3} [lemma="nerven"] within c 518 Treffer
Für diese Konstruktion ist die Frequenz in Deutschland signifikant höher als in der Schweiz und Österreich (Chi Quadrat = 29).
Hinten: [lemma="nerven"] [rfpos="VFIN.Sein.*"] within c 108 Treffer
Wegen einer geringen Precision aufgrund vieler Clause Fehler musste die Einschränkung festgesetzt werden, dass das Auxiliarverb *sein* direkt auf *genervt* folgen muss. Es zeigt sich keine auffällige Distribution, allerdings sind die absoluten Zahlen sehr klein (kein Beleg in der Schweiz).

Als Schlussfolgerung aus diesen Analysen lässt sich festhalten, dass bei Phänomenen, in denen gegenüber dem kodifizierten Standard ein Pronomen (oder auch eine andere Wortart) hinzukommt, dessen Annotationen mit allerhöchster Vorsicht zu geniessen und die gefundenen Treffer auf jeden Fall manuell zu kontrollieren sind. Zudem wäre es, falls dies technisch machbar ist, für syntaktische Untersuchungen eine grosse Erleichterung, wenn die Präzision der Tagger bei der Erkennung von Clause Grenzen (Nebensätzen) erhöht werden könnte.

### 4.3.2 Annotation von Personalpronomen: *anrufen* mit Dativ

Ob und falls ja, in welchen Regionen und mit welchen Frequenzen das Verb *anrufen* anstatt eines Akkusativs einen Dativ regiert (Ammon et al. 2004, 44f.: „wird in CH und im Grenzfall des Standards in D-südwest mit einem Dativobjekt verbunden, gemeindt. mit einem Akkusativobjekt“), lässt sich im Korpus nur mühsam untersuchen. Das Hauptproblem liegt darin, dass der RFTagger (der TreeTagger unterstützt ja gar keine Kasusdifferenzierung) die synkretistischen Fälle, in denen die Kasusformen des Dativs und Akkusativs identisch sind (z.B. *Er rief uns an.)*, entweder als Akkusativ oder Dativ annotiert. Dasselbe Problem in Bezug auf die Kasusrektionen von Präpositionen hat Simone Ueberwasser im bereits zitierten Aufsatz (Ueberwasser 2013) dokumentiert. Wünschbar wäre für Synkretismen meines Erachtens statt einer absoluten, mehr oder weniger willkürlichen Zuordnung zu einem von mehreren möglichen Kasus eine Kategorie „unentscheidbar“. Wie das Verhältnis der Annotationen als Akkusativ bzw. Dativ bei Synkretismen genau aussieht, habe ich am Beispiel des Personalpronomens *uns*, das in einem Abstand von 0 bis 3 Wörtern hinter dem Verb *anrufen* auftritt, überprüft. Die tabellarische Übersicht offenbart zudem, dass die beiden Tagger das Personalpronomen *uns* in einer gewissen Anzahl Fälle als Reflexivpronomen auffassen (rot markiert). Die 19 Dativobjekte, die der RFTagger zu identifizieren glaubt, sind sämtlich Synkretismen.

| | Total Treffer | Personalpronomen im Akkusativ | Personalpronomen im Dativ | Reflexivpronomen (im Dat. oder Akk.) |
|---|---|---|---|---|
| RFTagger | 115 | 59 (51%) | 19 (17%) | 37 (32%) |
| TreeTagger | 115 | 97 (84%) (keine Kasusdifferenzierung) | | 18 (16%) |

Bei den Personalpronomen im konkreten Beispiel zeigt, anders als bei den Reflexivpronomen oben (vgl. *sich nerven*), der TreeTagger eine höhere Precision als der RFTagger. Jedoch ist mit dem TreeTagger keine nach Dativ und Akkusativ differenzierte Suche möglich.

Um die Zahlen in der Tabelle nachvollziehen zu können, seien hier die einzelnen Abfragen dokumentiert.

RFTagger:

[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [word="uns"] []{0,3} [word="an" & pos="PTKVZ"] within c returned 115 matches

Davon Akkusativ:

[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [word="uns" & rfpos="PRO.Pers.*Acc.*"] []{0,3} [word="an" & pos="PTKVZ"] within c returned 59 matches

Davon Dativ:
[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [word="uns" & rfpos="PRO.Pers.*Dat.*"] []{0,3} [word="an" & pos="PTKVZ"] within c returned 19 matches → alles Synkretismen
Weder Akkusativ noch Dativ → Reflexivpronomen im Dativ oder Akkusativ:
[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [word="uns" & rfpos!="PRO.Pers.*(Dat.|Akk.)*"] []{0,3} [word="an" & pos="PTKVZ"] within c returned 37 matches

TreeTagger:
Personalpronomen:
[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [word="uns" & pos="PPER"] []{0,3} [word="an" & pos="PTKVZ"] within c returned 97 matches
Nicht Personalpronomen → Reflexivpronomen:
[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [word="uns" & pos!="PPER"] []{0,3} [word="an" & pos="PTKVZ"] within c returned 18 matches

Da es unter den aktuellen Bedingungen unmöglich ist, über eine Suche nach einem Dativobjekt nur die eindeutigen Dativ Kasusformen zu finden, müssen diese manuell identifiziert und von den Synkretismen (die nicht nur Artikel, Nomen und Pronomen, sondern auch Eigennamen betreffen) getrennt werden. Die folgenden Abfragen zeigen die Ergebnisse. In den Suchabfragen wurden einer grösseren Precision wegen je 0 bis 3 Wörter zwischen dem Verb *anrufen* und dem Dativobjekt erlaubt und Präpositionen vor dem Dativobjekt ausgeschlossen (Bsp. *bei ihm angerufen*).
In lediglich einem Artikel aus dem Schwarzwälder Boten (marchcrawlfour792_161) wird ein expliziter Dativ hinter dem Lemma *rufen* und vor dem Verbzusatz *an* verwendet, der Rest der 49 Treffer sind Synkretismen:
[(lemma="rufen")] [rfpos!="APPR.*"]{0,3} [(rfpos="PRO.Pers.*Dat.*") | (rfpos="ART.*Dat.*") | (rfpos="N.*Dat.*")] []{0,3} [word="an" & pos="PTKVZ"] within c (49 Treffer)
Auch unter den 54 Treffern für den Fall, dass das Dativobjekt vor dem Verb *anrufen* steht, findet sich lediglich ein relevanter Beleg aus der Bernerzeitung (marchcrawlthree1529_41):
<c> [rfpos!="APPR.*" & c & !c1]* [c1]* [rfpos="(PRO.Pers.*Dat.* | ART.*Dat.* | N.*Dat.*)"] [rfpos!="APPR.*" & c & !c1]* [c1]* [(lemma="anrufen") & c & !c1] within c (54 Treffer)

Zum Schluss sei lediglich noch am Rande erwähnt, dass das Lemma *anläuten*, welches im „Variantenwörterbuch des Deutschen“ (Ammon et al. 2004, 42) zum „Grenzfall des Standards“ gezählt wird, 24 Tokens zusammenfasst. Von diesen entfallen 19 auf Österreich (10 Mitte, 3 Ost, 5 Süd, 1 West) und 5 auf Deutschland (3 Südwest, 1 Südost, 1 Mittelost). Kein Beleg stammt aus einer Schweizer Zeitung. In keinem dieser Belege hat *anläuten* die Bedeutung *anrufen*. Entweder ist es verwendet als Synonym zu *an der Tür klingeln* oder es geht um Glocken, die gegen etwas *anläuten*. Das Korpus offenbart somit andere Verwendungen, als im „Variantenwörterbuch des Deutschen“ (Ammon et al. 2004, 42) und im Duden „Die deutsche Rechtschreibung“ (Duden 2009a, 204) vermerkt sind. Bei beiden ist nur die Bedeutung *jmdn. telefonisch anrufen* angegeben. Unter der Definition auf Duden online hingegen findet sich auch die Bedeutung *an der Tür klingeln* (http://www.duden.de/rechtschreibung/anlaeuten <14.9.2013>).

Als Schlussfolgerung dieser Ausführungen lässt sich festhalten, dass die Synkretismen, die in grossem Mass als Dativ annotiert werden, jegliche Abfragen zu Valenz und Rektion massgeblich erschweren. Alle vom RFTagger als Dativ annotierten Varianten sind unbedingt von Hand zu überprüfen, da Aussagen zur Frequenz von Dativen sonst stark verfälscht sein können. Noch einmal könnte an dieser Stelle auf die bereits mehrfach zitierte „Warnung“ von Belica et al. (2010, 461) verwiesen werden.

## 4.4 Vergleich von TreeTagger und RFTagger

Im Verlauf der präsentierten Analysen hat sich je nach zu untersuchendem Phänomen in gewissen Fällen der RFTagger und in anderen der TreeTagger als geeigneter erwiesen. Aufgrund meiner Recherchetätigkeit ist es nicht möglich, allgemeine Empfehlungen abzugeben, bei welcher Wortart welcher der beiden Tagger eine höhere Precision zeigt (vgl. das Beispiel der Pronomen, wo bei den Reflexivpronomen der RFTagger und bei den Personalpronomen der TreeTagger präzisere Resultate lieferte). Der TreeTagger verfügt allerdings über wesentlich weniger Spezifizierungsmerkmale hinsichtlich des Kasus, Numerus und Genus bei Substantiven, Pronomen, Adjektiven etc. und des Tempus, Numerus und Modus bei Verben (eine Übersicht zu den Merkmalen der beiden Tagger findet sich in der Dokumentation von Klaus Rothenhäusler zum VG-Korpus [2013]). Dieser Nachteil des TreeTaggers führt dazu, dass in vielen Fällen der RFTagger eingesetzt werden muss. Auf jeden Fall ist es ratsam, bei komplexen Phäno-

menen die Resultate beider Tagger zu vergleichen (disambiguiren) und den als präziser erachteten Tagger für die Detailuntersuchungen auszuwählen.

# 5 Schlusswort

Im Verlauf meiner Recherchen hat sich herauskristallisiert, dass mit den aktuellen Möglichkeiten und Limitierungen des Korpus „Variantengrammatik des Standarddeutschen“ Phänomene, bei denen sich die Varianten an eindeutigen sprachlichen Merkmalen manifestieren, und Satzstrukturen, bei denen die Position innerhalb des Satzes vorgegeben ist, am einfachsten untersuchen lassen. Sobald man jedoch auf die morpho-syntaktischen Annotationen der Tagger angewiesen ist, wie dies insbesondere bei linguistischen Untersuchungen zur Flexion, Rektion und Reflexivität von Verben der Fall ist, müssen die Resultate mit grösster Vorsicht genossen und wenn immer möglich manuell kontrolliert werden. Es gilt sich stets bewusst zu sein, dass die Annotationen lediglich Interpretationen der Tagger sind, die nicht der sprachlichen Wirklichkeit entsprechen müssen (vgl. Belica et al. 2010, 466). Als besonders kritisch hat sich die nicht nachvollziehbare Annotation von Synkretismen erwiesen.

Generell droht die Gefahr, dass Varianten, die nicht dem bundesdeutschen Standard entsprechen, mit dem die Tagger trainiert wurden, falsch oder unpräzise erkannt werden. Dies stellt im Rahmen einer variantengrammatischen Untersuchung, die sich gerade für Phänomene an der Peripherie der deutschen Standardsprache(n) interessiert, eine ernstzunehmende Schwierigkeit dar.

Meine Untersuchungen haben den Status von stichprobenartigen Beobachtungen von Einzelphänomenen. Wenn ich auch stets Hypothesen über mögliche Analogien aufgestellt habe, so müsste man, um die Ursachen der verschiedenen aufgedeckten Problematiken systematisch ergründen zu können, mit computerlinguistischen Methoden in die Tiefenstrukturen des Korpus eindringen können und allfällige Muster zu identifizieren versuchen.

Trotz der gegebenen Einschränkungen konnten im Zeitungskorpus verschiedene Variationsphänomene nachgewiesen werden, wobei gewisse analysierte Varianten, die eher konzeptionell mündlich sind, in sehr kleinen Frequenzen auftraten. Dies erlaubt die Annahme, dass Konstruktionen aus dem Dialekt oder der Umgangssprache (noch) nicht in den Gebrauchsstandard der Presse eingegangen sind.

Auf jeden Fall zeigt das Korpus, dass nationale und regionale Variation Realität im Sprachgebrauch der Presse ist. Man darf gespannt sein auf weitere korpusbasierte Erkenntnisse zur grammatischen Variabilität im Standarddeutschen.

# 6 Bibliographie

## 6.1 Sekundärliteratur

Ammon, Ulrich (1995): Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten. Berlin, New York: de Gruyter.

Ammon, Ulrich / Bickel, Hans / Ebner, Jakob et al. (2004): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin, New York: de Gruyter.

Belica, Cyril et al. (2011): The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.-24.09.2009. Ed. by Marek Konopka, Jacqueline Kubczak, and Christian Mair. (= Korpuslinguistik und interdisziplinäre Persepektiven auf Sprache 1). Tübingen: Narr, 451-469.

Duden (2009a): Die deutsche Rechtschreibung. 25., völlig neu bearb. und erw. Aufl. Hrsg. v. der Dudenredaktion. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Duden (2009b): Duden – Die Grammatik. Unentbehrlich für richtiges Deutsch. (= Duden 4). 8., überarb. Aufl. Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

Dürscheid, Christa (2009): Variatio delectat? Die Plurizentrizität des Deutschen als Unterrichtsgegenstand. In: Clalüna, Monika / Etterich, Barbara (Hrsg.): Deutsch unterrichten zwischen DaF, DaZ und DaM. Sondernummer Rundbrief AkDAF. Stallikon: Käser, 59-69.

Dürscheid, Christa / Elspaß, Stephan / Ziegler, Arne (2011): Grammatische Variabilität im Gebrauchsstandard – das Projekt «Variantengrammatik des Standarddeutschen». In: Konopka, Marek / Kubczak, Jacqueline / Mair, Christian / Štícha, František / Waßner, Ulrich H. (Hrsg.): Grammatik und Korpora 2009 / Grammar & Corpora 2009. Tübingen: Narr (= Corpus Linguistics and Interdisciplinary Perspectives on Language 1), 123-140.

Dürscheid, Christa / Giger, Nadio (2010): Variation in the case system of German – linguistic analysis and optimality theory. In: Lenz, Alexandra N./Plewnia, Albrecht (Hrsg.): Grammar between Norm and Variation. Frankfurt: Lang (= VarioLingua 40), 167-192.

Dürscheid, Christa / Hefti, Inga (2006): Syntaktische Merkmale des Schweizer Standarddeutsch. Theoretische und empirische Aspekte. In: Dürscheid, Christa / Businger, Martin (Hrsg.) (2006): Schweizer Standarddeutsch. Beiträge zur Varietätenlinguistik. Tübingen: Narr, 131-161.

Elspaß, Stephan (2010): Regional standard variation in and out of grammarians' focus. In: Lenz, Alexandra N. / Plewnia, Albrecht (Hrsg.): Grammar between norm and variation. (= VarioLingua 40). Frankfurt a.M. u.a.: Peter Lang, 127-144.

Evert, Stefan (2010): The IMS Open Corpus Workbench (CWB). CQP Query Language Tutorial. CWB Version 3.0. Version vom 17. Februar 2010.
Online unter: http://cwb.sourceforge.net/ <15.9.2013>

Haas, Walter (2000): Die deutschsprachige Schweiz. In: Bickel, Hans/Schläpfer, Robert (Hrsg.): Die viersprachige Schweiz. Aarau et al.: Sauerländer (= Reihe Sprachlandschaft 25), 57-138.

Koch, Peter; Oesterreicher, Wulf (1994): Schriftlichkeit und Sprache. In: Günther, Hartmut; Ludwig, Otto (Hrsg.): Schrift und Schriftlichkeit. Writing and Its Use. Ein interdisziplinäres Handbuch internationaler Forschung. An Interdisciplinary Handbook of International Research. 2 Bde. Berlin, New York: de Gruyter, 1. Halbband, 587-604.

König, Werner (1989): Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland. 2 Bde. Ismaning: M. Hueber Verlag.

Lötscher, Andreas (1997): "Guet, sind Si doo". Verbstellungsprobleme bei Ergänzungssätzen im Schweizerdeutschen. In: Ruoff, Arno/Löffelad, Peter (Hrsg.): Syntax und Stilistik der Alltagssprache. Beiträge der 12. Arbeitstagung zur alemannischen Dialektologie. Tübingen: Niemeyer, 85-95.

Ueberwasser, Simone (2013): Corpus Annotations & the German Case System: Can Morphosyntactic Case Annotations meet the Needs of German Variational Grammar Research? A Case Study. In: Linguistik Online (eingereicht zur Publikation).

Ziegler, Arne (2010): „Er erwartet sich nur das Beste ...“. Reflexivierungstendenz und Ausbau des Verbalparadigmas in der österreichischen Standardsprache. Zu einer Variantengrammatik des Deutschen. In: Bittner, Dagmar / Gaeta, Livio (Hrsg.): Kodierungstechniken im Wandel: Das Zusammenspiel von Analytik und Synthese im Gegenwartsdeutschen. (= Linguistik – Impulse & Tendenzen 34). Berlin, New York: de Gruyter, 67-81.

## 6.2 Projektinterne Dokumente

Dürscheid, Christa: Das Deutsche und seine Grammatiken (Powerpoint Präsentation).

Dürscheid, Christa: D-A-CH-Projekt: Variantengrammatik des Standarddeutschen (Powerpoint Präsentation).

Dürscheid, Christa / Elspaß, Stephan / Ziegler, Arne: Variantengrammatik des Standarddeutschen (Powerpoint Präsentation).

Rothenhäusler, Klaus (2013): VG-Korpus. Version vom 27.3.2013.

Semtracks (2010): Anleitung CQPWeb. Version vom 4. April 2010.

Ueberwasser, Simone (2012): Pluri-Regional German Grammar: A Corpus based approach. Grammar and Corpora: 4th International Conference, Prague, Version vom 29.11.2012.

Ueberwasser, Simone (2013a): Fehlerquellen beim Arbeiten mit Korpora. Überlegungen zum VG-Projekt. Version vom 18.4.2013.

Ueberwasser, Simone (2013b): Mustervorgehen im Projekt Variantengrammatik des Standarddeutschen. Version vom 28.6.2013.

## 6.3 Internetlinks

Datenbank zum Projekt „Variantengrammatik des Standarddeutschen“
https://www-gewi.uni-graz.at/variantengrammatik/ <15.9.2013>

Homepage des Projekts „Variantengrammatik des Standarddeutschen“
http://www.variantengrammatik.net/ <15.9.2013>

Korpus „Deutsch heute“ im Projekt „Variation des gesprochenen Deutsch“ des IDS-Mannheim
http://www1.ids-mannheim.de/prag/AusVar/Deutsch_heute/ <15.9.2013>

Zeitungskorpus des Projekts „Variantengrammatik des Standarddeutschen“
http://corpora.semtracks.org/marchcrawl/ <15.9.2013>