Simon Clematide

Wozu Kasusrektion auszeichnen bei Präpositionen?

1 Einleitung

Die Identifizierung von Kasus bei kasustragenden, deklinierbaren Wörtern (Pronomen, Artikel, Nomen, Adjektive) ist eine entscheidende Anforderung an die Sprachverarbeitung für flektierende Sprachen wie Deutsch. Neben grundlegenden syntaktischen Funktionen (Subjekte im Nominativ, Objekte im Akkusativ, Dativ oder Genitiv), welche vom Verb regiert werden und nominalen Modifikatoren im Genitiv sind Präpositionen mit ihren Rektionseigenschaften kasusbestimmend bzw. kasusregierend. Im folgenden Beispiel sind alle Präpositionen und alle kasustragenden Wörter mit entsprechenden Kasus-Tags markiert[1]:

> Mit/Dat dieser/Dat neuen/Dat Praxis/Dat reagiert/- das/Nom Gericht/Nom auf/Acc die/Acc öffentliche/Acc Schelte/Acc jüngster/Gen Urteile/Gen ./-

Um das Bestimmen von Kasusmerkmalen für deutsche Sätze zu lernen, kann die in deutschen Baumbanken annotierte morphologische Information verwendet werden. Die älteste und mit gut 20.000 Sätzen kleinste Baumbank NEGRA[2] (Skut et al., 1997) enthält nur sehr wenig und partielle, teilweise ambige morphologische Annotation, z.B. unaufgelöste Nominativ-Akkusativ-Alternativen. Die mit rund 50.000 Sätzen mehr als doppelt so große Baumbank TIGER[3] (Brants et al., 2004) verwendet mit kleineren Ausnahmen das originale große STTS-Tagset[4], das sowohl Wortarten als auch detaillierte morphologische Merkmale spezifiziert (Schiller et al., 1999; Teufel und Stöckert, 1996). Allerdings weicht TIGER in einem wichtigen Punkt von der STTS-Spezifikation ab und enthält keine Information zu Kasusrektion bei Präpositionen.

Die zentrale Frage, welche in dieser Studie untersucht wird, lautet, ob diese Abweichung vom STTS-Standard einen nennenswerten Nachteil darstellt für sprachtechnologische Systeme, welche aus einer solchen Baumbank die Zuweisung von Kasus lernen möchten. Eine Annotation, welche auf die Kasusauszeichnung von Präpositionen verzichtet, ergibt für obigen Beispielsatz die folgende Annotation:

> Mit/- dieser/Dat neuen/Dat Praxis/Dat reagiert/- das/Nom Gericht/Nom auf/- die/Acc öffentliche/Acc Schelte/Acc jüngster/Gen Urteile/Gen ./-

[1] Die folgenden Kasuskürzel (d.h. Kasus-Tags) werden in diesem Artikel verwendet: Nom (Nominativ), Acc (Akkusativ), Dat (Dativ), Gen (Genitiv), - (kein Kasus).

[2] Siehe `http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html`

[3] Siehe `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html`

[4] Für Informationen zum Stuttgart-Tübingen-Tagset (STTS) siehe `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/GermanTagsets.html`

Mit der Tübinger Baumbank TüBa-D/Z 7.0[5] (Telljohann et al., 2004) (im Folgenden jeweils mit dem Kürzel TUEBA bezeichnet), liegt mit rund 65.000 Sätzen eine noch größere Ressource vor, welche wie TIGER durchgängig mit morphologischen Kategorien annotiert ist. Auch wenn die morphologischen Merkmale der TUEBA im Gegensatz zu den Wortarten mit anderen als vom STTS geforderten Kürzeln notiert sind, lassen sich diese Merkmale leicht auf das STTS-Format abbilden. Eine für diese Arbeit zentrale Eigenschaft der Annotation der TUEBA ist, dass die Kasusrektion bei Präpositionen annotiert ist. Das erlaubt auf eine einfache Art Experimente durchzuführen, welche den Nutzen der Angabe von Kasusrektion bei Präpositionen evaluieren. Zu diesem Zweck wird die Kasusinformation bei Präpositionen (STTS-Tag APPR) einmal beibehalten und einmal entfernt.

1.1 Fragestellung

Weshalb sollte man in einer Baumbank die Kasusrektion von Präpositionen explizit annotieren?

Ein guter Grund wäre, dass grundsätzlich eine möglichst hohe linguistische Explizitheit angestrebt wird, indem nicht bloß die kasustragenden, sondern auch die kasusfordernden Elemente ausgezeichnet werden. Diese Information wird auch in den meisten Lexika und auch in elektronischen Ressourcen wie dem morphologischen Analysewerkzeug GERTWOL (Haapalainen und Majorin, 1994) zur Verfügung gestellt.

Ein weiterer Grund wäre, dass man bessere sprachtechnologische Systeme erzeugen kann, welche supervisierte Lernverfahren auf dem Material von Baumbanken anwenden. Grundsätzlich ist es zwar möglich, auch aus Baumbanken wie TIGER, welche keine Kasusrektion enthalten, in den meisten Fällen analoge Information aus den kasustragenden Elementen in der von der Präposition abhängigen Phrase abzuleiten. Allerdings reichen einfache Heuristiken nicht aus, welche beispielsweise den Kasus des am nächsten rechts stehenden Tokens übernehmen. Nicht selten enthalten abhängige Nominalphrasen komplexe pränominale Modifikatoren mit abweichendem Kasus (1) oder attributive Relativpronomen (2), welche mit Genitiv ausgezeichnet sind.

1. Zu Velazquez' Lebzeiten stand dort die Kirche San Juan, in/Dat deren/Gen Krypta/Dat der Maler 1660 beigesetzt wurde.

2. Rückschlag für/Acc St./Gen Paulis/Gen Amateure/Acc

Eine exakte Rekonstruktion der Kasusrektion erfordert deshalb eine nicht-triviale tiefergehende Analyse der abhängigen Phrase.

In dieser Arbeit soll nun experimentell untersucht werden, welche sprachtechnologischen Vorteile durch die explizite Kasusrektion bei Präpositionen (APPR) entstehen. Diese Frage wird operationalisiert durch systematische Experimente zur Qualität der Kasusklassifikation im TUEBA-Korpus (d.h. in deutschen Zeitungstexten) mit verschiedenen frei verfügbaren Systemen, welche eine hohe und dem Stand der Technik entspre-

[5] Siehe `http://www.sfs.uni-tuebingen.de/de/ascl/ressourcen/corpora/tueba-dz.html`

chende Performanz aufweisen. Kasusklassifikation wird dabei analog zum Tagging von Wortarten als ein Problem der Klassifikation von Sequenzen von Tokens betrachtet.

1.2 Verwandte Arbeiten

Kasusklassifikation als isolierte Anwendung für deutsche Texte wird nach unserer Kenntnis nur vom Kasus-Tagger gemacht, welcher Teil der Durm-Lemmatisierungsapplikation (vgl. Perera und Witte, 2006) ist. Normalerweise wird Kasus in Kombination mit anderen morphologischen Merkmalen analysiert, meist auch in Kombination mit der Wortart. Ein älteres Werkzeug ist MORPHY (Lezius et al., 1998), das ebenfalls wie das Durm-System auf Hidden-Markov-Modellen basiert. Ein aktuelles State-of-the-Art-System stellt das Modell des RFTaggers[6] für Deutsch dar, dessen statistische Komponente in den untenstehend beschriebenen Experimenten in dieser Arbeit benutzt wird. Die Software-Distribution dieses Taggers enthält auch Modelle für slawische Sprachen, und insbesondere für solche morphologisch reicheren Sprachen gibt es auch noch weitere Literatur (Hajič et al., 2001).

2 Ressourcen und Methoden

2.1 TUEBA

Für die Experimente verwenden wir die syntaktisch und morphologisch annotierte Tübinger Baumbank Tüba-D/Z 7.0, welche aus Zeitungstexten besteht und 65.524 Segmente (Sätze, Titel usw.) mit rund 1,2 Millionen Tokens enthält.

Die Abbildung 1 zeigt die Verteilung der verschiedenen Kasusklassen. Darin eingeschlossen ist der Fall, dass kein Kasus markiert ist, was auf 44% aller Tokens der TUEBA zutrifft. Die TUEBA verwendet im Original eigene Kürzel zur Kasusmarkierung, welche aber für diese Studie auf STTS-Kürzel abgebildet werden.

Wie ambig sind Präpositionen bezüglich Kasus im TUEBA-Korpus? Die Tabelle 1 zeigt die Distribution auf der Ebene der Tokens und der verschiedenen Types. Die Types der Ambiguitätsklassen 2 und 3 machen zwar nur 42% aller Types aus, enthalten aber diejenigen Präpositionen, welche gerade sehr häufig vorkommen und 87% aller APPR-Tokens ausmachen. Nur für die Präpositionen „statt“ und „außer“ sind alle vier Möglichkeiten im Korpus belegt.

Die Zahlen in Tabelle 1 sagen nicht besonders viel über die Schwierigkeit der Disambiguierung der Kasusrektion von Präpositionen aus, solange wir nichts über die Distribution der Kasus-Tags von ambigen Wortformen wissen. Falls nur einzelne Ausreißer eine Präposition ambig machen, können einfache Maximum-Likelihood-Entscheidungen das Problem sehr gut lösen.

Um die Schwierigkeit der Kasusdisambiguierung einer Präposition besser einschätzen zu können, kann die Entropie ihrer Kasusverteilung berechnet werden. Dieses Maß drückt die Unsicherheit aus, welche es beim Disambiguieren einer Präposition bezüglich

[6] Siehe `http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger`

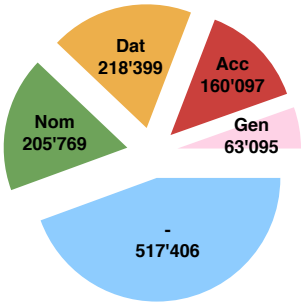

Abbildung 1: Verteilung der Kasus über allen 1.164.766 Tokens der TUEBA. Kasuslose Tokens („-“) dominieren stark. Der häufigste Kasus ist Dativ (Dat), gefolgt von Nominativ (Nom), Akkusativ (Acc) und Genitiv (Gen).

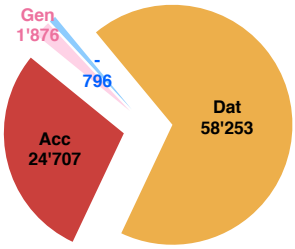

Abbildung 2: Verteilung der Kasus über allen 85.632 mit APPR klassifizierten Tokens der TUEBA (7,4% aller Tokens haben das STTS-Tag APPR). Der häufigste Kasus Dativ deckt 68% aller Fälle ab, Akkusativ 29% und Genitiv gut 2%. Nur 1% der Token, welche mit APPR getaggt sind, haben keine Kasusrektionsinformation.

Ambiguität	# Tokens	in %	# Types	in %
1	11.176	13,1	75	56,4
2	37.380	43,7	47	35,3
3	36.695	42,9	9	6,8
4	381	0,4	2	1,5
Total	85.632	100,0	133	100,0

Tabelle 1: Ambiguitätsrate der mit APPR getaggten Tokens in der TUEBA. Eine Ambiguitätsrate von 1 heißt, dass nur eine der folgenden vier möglichen Kasus-Tags vorkommt: „Acc“, „Dat“,„Gen“,„-“.

Präposition	$H(w)$	**Freq/Kasus**	**Freq**
statt	1,29	176/Gen 70/Dat 22/- 4/Acc	272
außer	1,04	87/Dat 8/- 8/Acc 6/Gen	109
bis	1,02	604/- 551/Acc 2/Dat	1157
einschließlich	1,00	8/Dat 7/Gen	15
anstatt	1,00	2/- 2/Gen	4
zuzüglich	1,00	1/Dat 1/Gen	2
getreu	1,00	1/Dat 1/Gen	2
auf	0,99	3940/Acc 2801/Dat 4/-	6745
binnen	0,99	14/Dat 11/Gen	25
dank	0,95	34/Gen 20/Dat	54
mittels	0,95	15/Gen 9/Dat	24
südlich	0,95	10/- 6/Gen	16
an	0,94	2291/Dat 1288/Acc	3579
plus	0,92	8/Dat 4/Acc	12
nordwestlich	0,92	2/- 1/Gen	3
anhand	0,90	13/Gen 6/-	19
entlang	0,90	11/Gen 5/Dat	16

Tabelle 2: Präpositionen mit höchster Entropie

ihrer Kasusrektion aufzulösen gilt. Formal ergibt sich die Entropie H einer Präposition w aus der negativen Summe aller Wahrscheinlichkeiten P, dass die Präposition einen bestimmten Kasus regiert, multipliziert mit der logarithmierten Wahrscheinlichkeit.

$$H(w) = -\sum_{t \in \text{Tagset}} P_w(t) \times \log_2 P_w(t)$$

Aus der obigen Formel ergibt sich einerseits, dass die Entropie größer wird, je mehr Kasus eine Präposition regieren kann. Zudem vergrößern auch gleichwahrscheinliche Kasus die Entropie.

Die Tabelle 2 zeigt die Präpositionen mit der höchsten Entropie aus der TUEBA. Sie illustriert schön, wie die beiden oben erwähnten Faktoren zusammenspielen: Hohe Entropie erscheint einerseits bei hochambigen Präpositionen und andererseits bei Präpositionen mit zwei gleich häufigen Kasus. Dabei muss es sich wie etwa bei „zuzüglich“ nicht um häufige Präpositionen handeln. Bei „statt“ ist ein stark schwankender Gebrauch von Kasus zu beobachten. Dies hängt teilweise mit fehlender expliziter Kasusmarkierung in der abhängigen Nominalphrase zusammen, welche die Kasusrektion nur unzureichend spezifiziert.

Am anderen Ende der Skala mit einer minimalen Entropie von 0 finden sich die bezüglich Kasus eindeutigen Präpositionen. Die häufigsten Vertreter aus der TUEBA sind in der Tabelle 3 aufgelistet. Die hochfrequente Präposition „mit“ ist leicht mehrdeutig

Präposition	$H(w)$	**Freq/Kasus**	**Freq**
nach	0,00	3761/Dat 2/- 1/Acc	3764
zu	0,00	2798/Dat 1/- 1/Acc	2800
seit	0,00	1012/Dat 1/-	1013
mit	0,00	8279/Dat 3/-	8282
aus	0,00	3653/Dat	3653
bei	0,00	3091/Dat	3091
gegen	0,00	1786/Acc	1786
durch	0,00	1536/Acc	1536
ohne	0,00	650/Acc	650
-	0,00	78/Acc	78
samt	0,00	36/Dat	36
namens	0,00	36/Gen	36
anläßlich	0,00	31/Gen	31
entgegen	0,00	24/Dat	24
seitens	0,00	20/Gen	20

Tabelle 3: Häufige Präpositionen mit niedriger Entropie

aufgrund von Fehlannotationen. Die Entropie von (gerundet) 0,00 bedeutet nicht, dass die Wortform „mit“ im Korpus immer eindeutig mit Dativ zu kennzeichnen wäre. Die Wortform „mit“ kann zusätzlich sowohl als abgetrenntes Verbpräfix (200 PTKVZ) wie als Adverb (72 ADV) erscheinen. Eine andere auffällige Präposition ist der Bindestrich „-“, welcher die Präposition „bis“ ersetzen kann.

Die Gesamtschwierigkeit der Bestimmung der Kasusrektion kann als kumulative Entropie aller Vorkommen von Präpositionen formalisiert werden, wobei die Entropie der einzelnen Präposition für jedes ihrer Vorkommen aufsummiert wird. Die Tabelle 4 zeigt diejenigen Präpositionen, welche in der TUEBA die höchste kumulative Entropie aufweisen. Dies sind erwartungsgemäß insbesondere die hochfrequenten kasusambigen Präpositionen. Es ist Teil der Aufgabe der Sprachmodelle, welche in den supervisierten Lernverfahren berechnet werden, diese Masse der Entscheidungs-Unsicherheit auf die korrekte Lösung hin zu reduzieren.

2.2 Externe und interne Tagsets

Ein externes Tagset bezeichnet das minimale Tagset, das für eine Evaluation oder für die eigentliche Verwendung in einer nachfolgenden Applikation benutzt wird. Unter einem internen Tagset versteht man ein reicheres und feineres Tagset, welches zum Trainieren und Taggen gebraucht wird. Das interne Tagset wird danach für die Evaluation oder Anwendung auf das externe Tagset abgebildet (*tag mapping*). Schon frühe Experimente von Brants (1997) mit statistischem Tagging für Deutsch haben gezeigt, dass die Verwendung von internen Tagsets über 1% Leistungsverbesserung erbringen können

Präposition	$H(w)$	**Freq/Kasus**	**Freq**	$\sum H(w)$
in	0,65	14563/Dat 2942/Acc 1/-	17506	11450,9
auf	0,99	3940/Acc 2801/Dat 4/-	6745	6650,2
an	0,94	2291/Dat 1288/Acc	3579	3373,5
bis	1,02	604/- 551/Acc 2/Dat	1157	1174,5
über	0,36	2379/Acc 178/Dat	2557	932,0
vor	0,23	2474/Dat 93/Acc 2/-	2569	600,4
unter	0,44	1222/Dat 124/Acc	1346	597,0
wegen	0,77	423/Gen 122/Dat	545	418,1
statt	1,29	176/Gen 70/Dat 22/- 4/Acc	272	351,8
für	0,04	7068/Acc 31/-	7099	287,6
von	0,02	9583/Dat 17/-	9600	179,9
trotz	0,68	198/Gen 44/Dat	242	165,5
um	0,08	1857/Acc 14/Gen 2/Dat	1873	141,6
innerhalb	0,77	140/Gen 35/- 1/Dat	176	135,2
hinter	0,50	225/Dat 28/Acc	253	127,0
außer	1,04	87/Dat 8/- 8/Acc 6/Gen	109	113,7
zwischen	0,12	829/Dat 14/Acc	843	102,8
ab	0,31	218/Dat 13/Acc	231	72,2
während	0,23	230/Gen 9/Dat	239	55,3
neben	0,18	293/Dat 8/Acc	301	53,3
dank	0,95	34/Gen 20/Dat	54	51,4
aufgrund	0,51	86/Gen 11/-	97	49,5
mit	0,00	8279/Dat 3/-	8282	38,6
nach	0,01	3761/Dat 2/- 1/Acc	3764	38,0
per	0,31	100/Acc 4/Dat 1/-	105	32,6
laut	0,17	154/Dat 4/Gen	158	26,9
zu	0,01	2798/Dat 1/- 1/Acc	2800	25,8
angesichts	0,21	113/Gen 4/-	117	25,2
binnen	0,99	14/Dat 11/Gen	25	24,7
mittels	0,95	15/Gen 9/Dat	24	22,9
jenseits	0,50	41/Gen 5/-	46	22,8
anhand	0,90	13/Gen 6/-	19	17,1
inklusive	0,77	17/Dat 5/Gen	22	17,0
außerhalb	0,31	50/Gen 3/-	53	16,6
pro	0,10	154/Acc 2/Dat	156	15,4
südlich	0,95	10/- 6/Gen	16	15,3
einschließlich	1,00	8/Dat 7/Gen	15	15,0

Tabelle 4: Präpositionen mit der höchsten kumulativen Entropie in der TUEBA

Tagset	Größe	Beispiel
Kasus	5	Frage/Dat
Kasus Numerus	15	Frage/Dat.Sg
Kasus Genus	20	Frage/Dat.Fem
Kasus Genus Numerus	50	Frage/Fem.Dat.Sg
Kasus Wortart	113	Frage/NN.Dat
Kasus Wortart Numerus	**197**	Frage/NN.Dat.Sg
Kasus Wortart Genus	277	Frage/NN.Dat.Fem
Kasus Wortart Numerus Person	296	ihn/PPER.Acc.Sg.3
Kasus Wortart Genus Numerus	460	ihn/PPER.Masc.Acc.Sg
Kasus Wortart Genus Numerus Person	492	ihn/PPER.Masc.Acc.Sg.3

Tabelle 5: Empirische, d.h. belegte Größen einiger interessanter interner Tagsets in der TUEBA. Fett ausgezeichnet ist das optimale interne Tagset, über dem vergleichend evaluiert werden kann.

gemessen an der Genauigkeit auf dem externen Tagset. Kasus-Tagging-Experimente im Rahmen von Clematide (2013) mit der systematischen Kombination von morphologischen Kategorien wie Wortart, Geschlecht, Numerus und Person haben gezeigt, dass für die verwendeten Werkzeuge ein internes Tagset mit den zusätzlichen Kategorien Wortart und Numerus optimal ist. Die Tabelle 5 zeigt, wie sich die Größe der internen Tagsets verändert, wenn zusätzliche morphologische Kategorien verwendet werden. Das optimale Tagset enthält 197 verschiedene Tags.

Eine alternative Verfeinerung des Tagsets wäre denkbar, bei der die Kasus-Tags der häufigsten oder aller Präpositionen lexikalisiert werden, d.h. mit dem Lemma der Präposition angereichert werden.

> Mit/mit- dieser/Dat neuen/Dat Praxis/Dat reagiert/- das/Nom Gericht/Nom auf/auf- die/Acc öffentliche/Acc Schelte/Acc jüngster/Gen Urteile/Gen ./-

Diese Technik hat für das große morphologische Tagset in den Experimenten von Schmid und Laws (2008) eine Leistungsteigerung erbracht, da damit die Präpositionen auf der Ebene der Tags mehr Information einbringen können.

2.3 Werkzeuge zur supervisierten Wortartenklassifikation

Für die Beantwortung unserer Fragestellung beziehen wir in dieser Arbeit nur frei verfügbare und rein statistische Werkzeuge mit ein, welche alle Informationen aus demjenigen Teil der TUEBA beziehen, mit dem das Sprachmodell trainiert wird. Die meisten praktischen Werkzeuge, mit denen Tokenfolgen (d.h. Sätze oder vergleichbare Textsegmente) mit linguistischer Information wie Wortart oder morphologischer Information wie in unserem Fall klassifiziert werden, erlauben das Hinzufügen von größeren

externen Lexika, mit denen bessere Resultate erreicht werden können. Um die Untersuchungsresultate vergleichbar und leichter reproduzierbar halten zu können, verzichten wir auf externe Ressourcen und nehmen die entsprechende Leistungsreduktion in Kauf. Im Zentrum dieser Arbeit steht ein vergleichendes methodisches Erkenntnisinteresse und weniger die Idee, Kasus mit Hilfe von möglichst allen verwendbaren lexikalischen Ressourcen optimal zu klassifizieren.

Im Rahmen dieser Arbeit können keine vollständigen technischen Beschreibungen der verwendeten Methoden gegeben werden. Wir verweisen auf die jeweilige Literatur und erwähnen nur die für diese Arbeit wichtigen Eigenschaften der Werkzeuge.

2.3.1 `hunpos`: Ein Tagger mit klassischem Hidden-Markov-Modell (HMM) (Halácsy et al., 2007)

Bei `hunpos` handelt es sich um eine quellfreie Reimplementation des bekannten älteren statistischen Trigramm-Taggers TnT (vgl. Brants, 2000). `hunpos` erlaubt allerdings eine flexiblere Parametrisierung bezüglich der Kontextgröße, welche für die Bestimmung der Übergangswahrscheinlichkeiten der Tags verwendet wird.

Die Abbildung 3 illustriert das Kontextmodell eines N-gramm-Taggers, das folgendermaßen funktioniert. Die Klasse, d.h. das Tag t_n eines Tokens w_n ergibt sich aus:

- der Verteilung der möglichen Tags vom Token w_n aus dem Tagger-Lexikon,
- den bereits berechneten Tags der $N-1$ vorangehenden Tokens und den Wahrscheinlichkeiten, ein bestimmtes Tag für w_n an der n-ten Stelle zu haben,
- der Verteilung der Tag-Wahrscheinlichkeiten für unbekannte, d.h. im Trainingsmaterial nicht vorgekommenen Wörter, welche anhand eines Endungsbaums (*suffix tries*) berechnet wird.

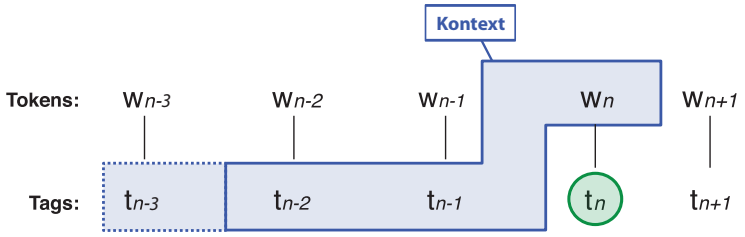

Abbildung 3: Kontextmodell eines Trigramm- bzw. Quadrigramm-Taggers

Im Rahmen einer größer angelegten vergleichenden Studie zur Kasusklassifikation im Deutschen (vgl. Clematide, 2013) zeigte sich eine Kontextgröße von 3 vorangehenden

Tags (Quadrigramm-Tagger) als optimal. Dies im Vergleich zur normalen, kleineren Kontextgröße 2 eines Trigramm-Taggers oder zu noch größeren Kontextfenstern, wo sich die nur spärlich vorhandenen Daten beim statistischen Tagging negativ auswirken.

2.3.2 `RFTagger`: Ein HMM- und entscheidungsbaumbasierter Tagger für große morphologische Tagsets (Schmid und Laws, 2008)

Beim `RFTagger` handelt es sich um einen statistischen State-of-the-Art-Tagger, welcher für den Umgang mit großen morphologischen Tagsets optimiert ist, welche im Deutschen oder auch in slawischen Sprachen (vgl. Erjavec et al., 2003) auftreten. Mit einem guten externen Lexikon erreicht dieser Tagger auf dem vollen morphologischen Tagset auf Zeitungstexten, wie sie dem TIGER-Korpus zugrunde liegen, eine Genauigkeit von 91,1% (vgl. Schmid und Laws, 2008). In unseren Experimenten verwenden wir wie erwähnt keine externen Lexika, allerdings benötigt der `RFTagger` einen einfachen Wortarten-Guesser. Wir verwenden denjenigen, welcher in der Software-Distribution des Taggers für Deutsch mitgeliefert wird.

Der Grund für die Leistungsfähigkeit dieses Taggers im Zusammenhang mit großen Tagsets liegt darin, dass die Tags nicht als unstrukturierte atomare Symbole betrachtet werden (z.B. „NN.Fem.Sg“), sondern als Vektoren von morphologischen Merkmalen, d.h. „Wortart=NN“,„Genus=Fem“,„Numerus=Sg“. Die Verwendung von Entscheidungsbaumtechniken mit Pruning erlaubt dann, dass potentiell große Kontexte gezielt nach relevanter Information abgeprüft werden können und dabei die morphologischen Merkmale separat verrechnet werden. Für Kasusklassifikation wurde in den Experimenten in Clematide (2013) eine optimale Kontextgröße von 4 für die TUEBA festgestellt. Als einziges betrachtetes Werkzeug ist `RFTagger` fähig mit einem internen Tagset optimal umzugehen, welches auch noch die morphologische Kategorie „Person“ enthält. Der Leistungsunterschied ist aber so gering, dass wir uns aus Gründen der Vergleichbarkeit auf ein gemeinsames internes Tagset für alle Werkzeuge beschränkt haben.

2.3.3 `wapiti`: Ein generisches Conditional-Random-Field-Werkzeug mit einem selbsterstellten Modell (Lavergne et al., 2010)

Bei `wapiti` handelt es sich um ein generisches Werkzeug zum Erstellen von sequentiellen Conditional-Random-Fields (CRFs) (vgl. Sutton und McCallum, 2012). CRFs sind bekannt für ihre State-of-the-Art-Leistung bei der Klassifikation von Sequenzen. Im Gegensatz zu den beiden oben erwähnten HMM-basierten Taggern muss vom Benutzer von Hand bestimmt werden, welche Kontextmerkmale für die Vorhersage der Klassifikationstags in Betracht gezogen werden. Der Benutzer muss sich auch selbst um geeignete Merkmale für den Umgang mit unbekannten, d.h. im Trainingsmaterial nicht vorgekommenen Tokens kümmern, was bei spezialisierten Wortarten-Taggern standardmässig mit Hilfe von Endungsbäumen gelöst ist.

Im Gegensatz zu HMM-basierten Taggern ist dafür das Kontextmodell von CRFs viel flexibler und global: Beliebige Information kann aus jeder Position der Tokenebene

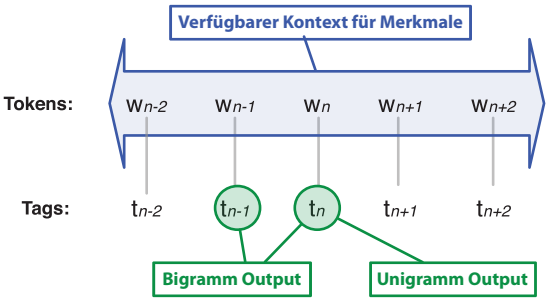

Abbildung 4: Kontextmodell eines sequentiellen CRF-Taggers

als Feature extrahiert und kombiniert werden. Diese Information kann als Unigramm-Merkmal auf das aktuelle Klassifikations-Tag oder als Bigramm-Merkmal auch auf das aktuelle und vorangehende Klassifikations-Tag bezogen werden. Bigramm-Merkmale können jedoch bei großen Tagsets schnell zu einer Merkmalsexplosion führen, weshalb sie nur sehr gezielt eingesetzt werden sollten. Die Abbildung 4 illustriert das unrestringierte Kontextmodell von sequentiellen CRFs und den Unterschied von Unigramm- und Bigramm-Merkmalen.

Das in Clematide (2013) entwickelte eigene Modell benutzt folgende Merkmalsgruppen:

- Das aktuelle Token sowie seine Präfixe und Suffixe der Längen 1 bis 3.
- Die Kombination vom ersten und letzten Buchstaben des aktuellen Tokens.
- Die Information, ob das aktuelle Token groß oder klein geschrieben ist.
- Nachbartokens bis zu 2 Positionen nach rechts und links.
- Wortbigramme aus aktuellem Token und linkem/rechtem Nachbarn.
- Die Kombination der 2 letzten Buchstaben vom linken, aktuellen und rechten Token.

Für die Modellbildung wurde der Algorithmus `rprop-` von `wapiti` verwendet (mit dem Standardwert für die L1-Regularisierung). Dieser Algorithmus ist schneller und speicherschonender als das Limited-Memory-BFGS-Verfahren und liefert nur geringfügig schlechtere Resultate.

	Mit Kasus			Ohne Kasus		
	Internes Tagset			Internes Tagset		
System	K,W,N	Kasus	Differenz	K,W,N	Kasus	Differenz
hunpos	91,2	84,9	-6,3	90,8	84,4	-6,4
RFTagger	91,6	84,9	-6,6	91,3	84,5	-6,8
wapiti	93,8	92,5	-1,3	93,9	93,1	-0,8

Tabelle 6: Einfluss der Verwendung von einem reichen internen Tagset (Kasus, Wortart (STTS-Tag), Numerus) für Training und Tagging im Vergleich zur direkten Verwendung des externen Tagsets (Kasus) für das Training und Tagging. Die Spalte „Mit Kasus“ zeigt die Resultate für die TUEBA, bei denen die Kasusrektionsangabe von Präpositionen entfernt wurde. Die Spalte „Ohne Kasus“ zeigt die Resultate, bei denen diese Angaben beibehalten wurden.

3 Resultate und Diskussion

In dieser Arbeit soll die Frage operationalisiert und beantwortet werden, ob explizit annotierter Rektionskasus bei Präpositionen (APPR) das statistische Tagging von Kasus über allen Tokens insgesamt verbessert oder nicht. Um diese Frage zu beantworten, sind auf TUEBA-Daten Experimente mit 10-facher Kreuzvalidierung durchgeführt worden, wobei einerseits bei allen APPR-getaggten Tokens das Kasusmerkmal beibehalten („mit Kasusrektion“) oder gelöscht („ohne Kasusrektion“) worden ist.

Für `hunpos` und `RFTagger` wird jeweils 90% der TUEBA als Trainingsmaterial und 10% als Testmaterial verwendet. Für `wapiti` ist ein Entwicklungsset notwendig, welches aus 1/5 des Trainingsmaterials besteht, d.h. nur 72% der Gesamtdaten stehen als Trainingsmaterial zur Verfügung. Alle Aufteilungen sind sequentiell zusammenhängende Korpusstücke. Bei einem zufälligen Ziehen von einzelnen Sätzen (Segmenten) würde eine zu optimistische, d.h. für echte Anwendungen nicht realistische lexikalische Abdeckung entstehen.

Evaluiert wird auf den arithmetischen Mittelwerten der Genauigkeit (*accuracy*) der 10 Testkorpora aus der Kreuzvalidierung. Mit Hilfe eines einseitigen Wilcoxon-Mann-Whitney-Test (R Core Team, 2013) wird geprüft, ob bei Systemen mit höherem Mittelwert die Genauigkeit statistisch signifikant besser ist. Der nicht-parametrische Wilcoxon-Test wird anstelle des T-Tests verwendet, weil die Differenzen der Mittelwerte nicht in allen Fällen eine Normalverteilung aufweisen. In den Resultatstabellen wie Tabelle 7 auf Seite 57 wird zudem die Verbesserung der unteren Grenze des Konfidenzintervalls (95%) angegeben.

3.1 Internes und externes Tagset

In der Tabelle 7 sind die Resultate dargestellt, welche durch das Training mit dem optimalen, vergleichbaren internen Tagset auf dem externen Tagset resultieren. Um den

System	Mean	SD	$_{\Delta}$abs	P-value	$_{\Delta}\mathrm{CI}_{low}$
hunpos, ohne Kasus	90,77	0,87			
hunpos, mit Kasus	91,17	0,29	+0,41	0,0020	+0,08
RFTagger, ohne Kasus	91,30	0,85			
RFTagger, mit Kasus	91,58	0,23	+0,29	*0,1377*	-0,03
wapiti, **mit** Kasus	93,79	0,17			
wapiti, **ohne** Kasus	**93,85**	0,64	+0,06	0,0420	+0,21

Tabelle 7: Evaluationsresultate der Werkzeuge auf den 5 Kasus-Tags. Die Reihenfolge der Daten ist pro System geordnet nach aufsteigender mittlerer Genauigkeit. Man beachte die abweichende Reihenfolge bei `wapiti`, welche sich durch dieses Kriterium ergibt. Das Training erfolgte mit dem internen Tagset mit STTS-Wortart, Numerus und Kasus. Die Spalte SD enthält die Standardabweichung. Die Spalte $_{\Delta}$abs enthält die absolute Differenz zur vorangehenden Zeile. Die Spalte P-value enthält den entsprechenden Wert des Wilcoxon-Tests. Die Spalte $_{\Delta}\mathrm{CI}_{low}$ die relative Verbesserung (bzw. Verschlechterung im Fall von negativen Werten) bezüglich der unteren Schranke des Konfidenzintervalls (95%), d.h. die Leistungverbesserung gegenüber dem Vergleichssystem, welche mindestens in 95 von 100 Fällen erreicht werden sollte.

Nutzen dieses internen Tagsets zu quantifizieren, wurde auch direkt auf dem externen Tagset trainiert und evaluiert. Die Tabelle 6 zeigt die Leistungsdifferenz. Reine HMM-basierte Ansätze profitieren mit über 6% von der Verwendung eines reicheren internen Tagsets. Für den CRF-Ansatz ergeben sich deutlich geringere Leistungsgewinne im Bereich von 1%.

3.2 Einfluss der Kasusrektion für die Vorhersage der Kasus-Tags auf allen Tokens der TUEBA

Zuerst soll die Frage beantwortet werden, wie gut sich die Kasusklassen (Nom, Akk, Dat, Gen, –) von allen Tokens vorhersagen lassen, wenn man auf 2 Versionen der TUEBA trainiert. Einer Version, bei der die Kasusrektion von Präpositionen (APPR) im Trainingsmaterial behalten wird und einer Version, bei der die Kasusrektion von Präpositionen entfernt ist (d.h. auf den Wert „–“ gesetzt).

Die Tabelle 7 zeigt, dass die Angabe von Kasusrektion für `hunpos` zu einer absoluten Verbesserung von 0,41% führt, welche statistisch signifikant ist. Beim `RFTagger` ist der Mittelwert zwar leicht besser mit der Angabe von Kasusrektion, allerdings liegt der Wilcoxon-Test mit 0,14 deutlich über der Standard-Signifikanzschwelle von 0,05. Dies bringt die *Senkung* (!) der unteren Grenze des Konfidenzintervalls (95%) ebenfalls zum Ausdruck. Das bedeutet, dass sich für den `RFTagger` keine statistisch signifikante Leistungssteigerung durch explizite Kasusrektion ergibt. Wie ist das möglich, obwohl die mittlere Leistungssteigerung vom `RFTagger` $(+0,29)$ beinahe fünf Mal größer ist als der Leistungsunterschied bei `wapiti` $(+0,06)$, welche statistisch signifikant ist? Mit dem Wilcoxon-Text werden die paarweisen Differenzen der beiden Systemvarianten

(mit/ohne Kasus) über den einzelnen Testdatensets der Kreuzvalidierung verglichen. Falls diese Differenzen ein uneinheitliches Bild ergeben, steigt die Wahrscheinlichkeit, dass die resultierende mittlere Leistungsverbesserung rein zufällig zu beobachten war. Diese Wahrscheinlichkeit wird durch den P-Value ausgedrückt.

Bei `wapiti`, dem System mit der deutlich besten Leistung ingesamt, liegt der Fall sogar umgekehrt und die Gesamtleistung sinkt statistisch signifikant, wenn die Kasusrektion hinzugefügt wird beim Lernen.

Was sind die möglichen Gründe für diese Zahlen? Ein wichtiger Punkt ist sicher, dass die Vorhersage der 74.456 potentiell ambigen APPR-Tokens (von insgesamt 85.632 APPR-Tokens) nicht ohne Fehler geschieht. Ein anderer Punkt ist, dass das Kontextmodell der klassischen HMM-Verfahren keine direkte Evidenz von Tokens rechts vom aktuellen Token einbeziehen kann.[7] Die Rektionsmarkierung von bezüglich Kasus eindeutigeren Präpositionen könnte nach rechts für die abhängige Phrase eine diskriminierende Information darstellen. Allerdings zeigt der `RFTagger`, der ein vergleichbares Kontextmodell besitzt, ein leicht anderes Verhalten als `hunpos`. Dem CRF-Modell, welches eine unrestringiertere Sicht auf die Token-Ebene hat, nützt die nach rechts vorstrukturierende Rektionsinformation insgesamt nichts. Die erhobenen Resultate erlauben keine Erklärung, wieso diese Effekte auftreten. Man kann einfach festhalten, dass der sprachtechnologische Nutzen vom verwendeten Modell abhängig ist.

In Tabelle 7 lässt sich auch gut ablesen, dass die Standardabweichung für die mittleren Genauigkeitswerte der 3 Tagger deutlich kleiner ist, wenn die Kasusrektion bei den Präpositionen vorhanden ist. Die Kasusinformation scheint also eine stabilisierende Wirkung zu haben auf die Leistung der Klassifikatoren.

3.3 Qualität der Vorhersage der Kasus-Tags aller APPR-Tokens

Wie gut kann der Kasus von APPR-Tokens überhaupt bestimmt werden von den 3 Werkzeugen? Um diese Frage zu beantworten, haben wir nur diejenigen 85.632 Token evaluiert, welche im Goldstandard das Tag APPR aufweisen. Auf den Beispielsatz (1) bezogen wurden also nur die zwei Tokens in (2) evaluiert.

1. Mit/Dat dieser/Dat neuen/Dat Praxis/Dat reagiert/- das/Nom Gericht/Nom auf/Acc die/Acc öffentliche/Acc Schelte/Acc jüngster/Gen Urteile/Gen ./-

2. Mit/Dat auf/Acc

Die Tabelle 8 zeigt die Resultate. Erstaunlicherweise ist `hunpos` in dieser Teilaufgabe leicht besser als der `RFTagger`. `wapiti` liegt deutlich vorne und klassifiziert auch stabiler,

[7] Damit soll nicht behauptet werden, dass ein HMM-Tagger insgesamt die Tagging-Entscheidung nur auf Grund des aktuellen Tokens und des linken Kontexts fällt. Sonst wäre es nicht möglich, dass das Wort „Das“ im Satz „Das ist richtig“ korrekt als Demonstrativpronomen (PDS) getaggt würde, obwohl der linke Kontext und die lexikalische Wahrscheinlichkeit von „Das“ die massiv häufigere Wortart als Begleiter (ART) bevorzugen würden. HMM-Tagger bestimmen die global beste Sequenz von Wortarten und fällen ihre Entscheidungen nicht Wort für Wort. Das heißt, für die Entscheidung PDS vs. ART spielt es eine wichtige Rolle, dass ein finites Hilfsverb (VAFIN) nach ART viel unwahrscheinlicher ist als PDS.

System	Mean	SD
RFTagger	92,8	0,48
hunpos	93,4	0,52
wapiti	96,8	0,39

Tabelle 8: Genauigkeit der Kasusklassifikation auf den 85.632 APPR-Tokens

System	Mean	SD	$_{\Delta}$abs	P value	$_{\Delta}CI_{low}$
hunpos, ohne Kasus	81,09	1,83			
hunpos, mit Kasus	84,19	0,58	+3,10	0,0010	+2,41
RFTagger, ohne Kasus	82,44	1,65			
RFTagger, mit Kasus	85,08	0,51	+2,64	0,0010	+2,03
wapiti, ohne Kasus	88,20	1,26			
wapiti, mit Kasus	**89,40**	0,33	+1,20	0,0010	+0,71

Tabelle 9: Evaluationsresultate der Werkzeuge auf den 5 Kasus-Tags. Die Reihenfolge der Daten ist pro System geordnet nach aufsteigender mittlerer Genauigkeit. Das Training erfolgte mit dem internen Tagset mit STTS-Wortart, Numerus und Kasus. Die Spalte SD enthält die Standardabweichung. Die Spalte $_{\Delta}$abs enthält die absolute Differenz zur vorangehenden Zeile. Die Spalte P-value enthält den entsprechenden Wert des Wilcoxon-Tests. Die Spalte $_{\Delta}CI_{low}$ die relative Verbesserung bezüglich der unteren Schranke des Konfidenzintervalls (95%), d.h. die Leistungverbesserung gegenüber dem Vergleichssystem, welche mindestens in 95 von 100 Fällen erreicht werden sollte.

was durch die niedrigere Standardabweichung belegt wird. Die Fehlerrate beim Kasus-Tagging von Präpositionen ist durchaus in einem Bereich, der die Gesamtgenauigkeit spürbar drücken kann, da die 85.632 Präpositionen ja 7,4% aller Tokens ausmachen.

3.4 Einfluss der Kasusrektion auf die Vorhersage der Kasus-Tags aller kasustragenden Tokens der TUEBA

Statt der Kasusklassifikation der Präpositionen kann auch die Kasusklassifikation aller *kasustragenden* Tokens evaluiert werden. Es geht um die Frage, wie gut die drei Werkzeuge diejenigen Wortarten klassifizieren können, welche echte Kasusmerkmale tragen können. Um diese Frage zu beantworten, haben wir nur diejenigen Tokens in die Evaluation einbezogen, welche im Goldstandard mit einem der folgenden STTS-Tags getaggt sind: ADJA, APPRART, ART, NE, NN, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWAT, PWS. Auf den Beispielsatz (1) bezogen wurden also nur die zehn Tokens in (2) evaluiert.

1. Mit/Dat dieser/Dat neuen/Dat Praxis/Dat reagiert/- das/Nom Gericht/Nom auf/Acc die/Acc öffentliche/Acc Schelte/Acc jüngster/Gen Urteile/Gen ./-

2. dieser/Dat neuen/Dat Praxis/Dat das/Nom Gericht/Nom die/Acc öffentliche/Acc Schelte/Acc jüngster/Gen Urteile/Gen

Die Tabelle 9 zeigt die entsprechenden Resultate. `hunpos` profitiert am meisten von der Rektionsinformation der Präpositionen, der `RFTagger` profitiert ebenfalls sehr deutlich. `wapiti` profitiert am wenigsten, aber die Kasusrektion ergibt für die (im Goldstandard) mit einer kasustragenden Wortart klassifizierten Tokens immerhin eine statistisch signifikante Verbesserung auf der mittleren Genauigkeit von ca. 1,2 Punkten. D.h. für die Untermenge der im Goldstandard mit kasustragenden Wortarten getaggten Tokens ergibt sich auch für `wapiti` eine Verbesserung, wenn man die Kasusrektion beim Training benutzt.

Die Standardabweichung der Klassifikationsgenauigkeit ist auch in diesem Fall wieder deutlich geringer, wenn die Präpositionsrektion verwendet wird. D.h. die Klassifikatoren sind nicht nur genauer, sondern auch stabiler.

4 Schluss

Die mit der TUEBA durchgeführten Experimente haben gezeigt, dass die Frage, ob es aus sprachtechnologischer Sicht Sinn macht, Kasusrektion bei Präpositionen zu annotieren (oder zu rekonstruieren, wie es für TIGER notwendig wäre), nicht eindeutig zu beantworten ist. Relativ einfache und auch Ressourcen schonende Systeme wie die HMM-basierten Tagger `hunpos` oder `RFTagger` profitieren. CRF-basierte Verfahren, welche die besten Resultate erbringen, aber auch massiv aufwändiger in der Berechnung sind, profitieren insgesamt nicht, d.h. unter Berücksichtigung aller Tokens. Nur wenn man die Evaluation auf die im Goldstandard als kasustragend klassifizierten Tokens einschränkt, wird auch für das CRF-basierte Modell eine Leistungssteigerung messbar. Die Rektionsinformation bei APPR scheint somit bei einfachen HMM-Modellen zu helfen, welche eine limitierte Sicht auf die direkte Evidenz rechts vom zu taggenden Token aufweisen. Um die genauen Faktoren und Gründe für das unterschiedliche Verhalten der betrachteten Ansätze zu verstehen, sind weitergehende Fehleranalysen und Untersuchungen notwendig. Unabhängig davon bleibt die linguistische Motivation nach möglichst expliziter Annotation von Kasusinformation bestehen.

Unser eigenes CRF-Modell löst das Problem der Kasusbestimmung auf der TUEBA mit einer Genauigkeit von 93,8% und schlägt damit das State-of-the-Art-Werkzeug `RFTagger` deutlich, das 91,6% erreicht. Wenn nur die Kasusdisambiguierung von Präpositionen betrachtet wird, schlägt unser Modell mit 96,8% sowohl den klassischen HMM-Tagger `hunpos` mit einem Quadrigramm-Modell (93,4%) als auch den `RFTagger` mit seinen 92,8% Genauigkeit.

Es wurde weiter gezeigt, dass bei allen Taggern die Ergebnisse der Kasusklassifikation verbessert werden, wenn das von den Taggern intern verwendete Tagset zusätzlich zum Kasus auch Numerus und Genus enthält. Dabei profitieren HMM-Tagger mit 6% wesentlich stärker als der CRF-Tagger mit 1%.

Eine interessante Frage für weitere Untersuchungen wäre, ob und wie stark die Benutzung von morphologischer Information, wie sie etwa von Morphologieanalysesystemen

wie GERTWOL (Haapalainen und Majorin, 1994) geliefert werden, die Kasusbestimmung weiter optimieren kann.

Danksagung

Herzlichen Dank an die Gutachter, welche wertvolle und anregende Rückmeldungen gegeben haben.

Literatur

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G. und Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Brants, T. (1997). Internal and external tagsets in part-of-speech tagging. In: *Proceedings of Eurospeech*, Seiten 2787–2790, Rhodos, Griechenland.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seiten 224–231, Seattle, WA, USA.

Clematide, S. (2013). A case study in tagging case in German: an assessment of statistical approaches. In: Mahlow, C. und Piotrowski, M. (Hgg.), *Systems and Frameworks for Computational Morphology: Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings*, Seiten 22–34. Springer.

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M. und Vitas, D. (2003). The MULTEXT-east morphosyntactic specifications for Slavic languages. In: *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, MorphSlav '03, Seiten 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Haapalainen, M. und Majorin, A. (1994). *GERTWOL: Ein System zur automatischen Wortformenerkennung deutscher Wörter*. Lingsoft Oy, Helsinki.

Hajič, J., Krbec, P., Květoň, P., Oliva, K. und Petkevič, V. (2001). Serial combination of rules and statistics: a case study in Czech tagging. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, Seiten 268–275, Stroudsburg, PA, USA. Association for Computational Linguistics.

Halácsy, P., Kornai, A. und Oravecz, C. (2007). Hunpos: an open source trigram tagger. In: *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Seiten 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lavergne, T., Cappé, O. und Yvon, F. (2010). Practical very large scale CRFs. In: *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seiten 504–513. Association for Computational Linguistics.

Lezius, W., Rapp, R. und Wettler, M. (1998). A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. In: *Proceedings of COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Band 2, Seiten 743–748, Montreal, Kanada.

Perera, P. und Witte, R. (2006). *The Durm German Lemmatizer*. Universität Karlsruhe.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Wien, Österreich.

Schiller, A., Teufel, S. und Stöckert, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technischer Bericht, Universität Stuttgart und Universität Tübingen.

Schmid, H. und Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Seiten 777–784, Manchester, UK.

Skut, W., Krenn, B., Brants, T. und Uszkoreit, H. (1997). An annotation scheme for free word order languages. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Seiten 88–95, Washington, D.C., USA.

Sutton, C. A. und McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Telljohann, H., Hinrichs, E., Kübler, S. et al. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Seiten 2229–2235, Lissabon, Portugal.

Teufel, S. und Stöckert, C. (1996). ELM-DE: EAGLES Specifications for German morphosyntax: Lexicon Specification and Classification Guidelines. (`http://www.ilc.cnr.it/EAGLES96/pub/eagles/lexicons/elm_de.ps.gz`).