Volume 28 — Number 1 — 2013 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Das Stuttgart-Tübingen Wortarten-Tagset – Stand und Perspektiven

The Stuttgart-Tübingen Part-of-Speech Tagset – Current Status and Plans

Herausgegeben von /*Edited by*
Heike Zinsmeister,
Ulrich Heid und Kathrin Beck

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 28 – 2013 – Heft 1
Herausgeber	Heike Zinsmeister, Ulrich Heid, Kathrin Beck
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Heike Telljohann, Yannick Versley, Kathrin Beck, Erhard Hinrichs, Thomas Zastrow

STTS als Part-of-Speech-Tagset in Tübinger Baumbanken

1 Einleitung

Das Stuttgart-Tübingen Tagset (STTS, Schiller et al., 1999) ist der De-facto-Standard für das Tagging von Wortarten in deutschen Texten, und die überwiegende Mehrzahl der POS-annotierten Ressourcen fürs Deutsche – darunter die Baumbanken NeGra (Skut et al., 1997), TIGER (Brants et al., 2002), TüBa-D/S (Hinrichs et al., 2000) und TüBa-D/Z (Hinrichs et al., 2004), und viele andere Korpora – verwenden dieses Tagset.

In dieser Rolle stellt das STTS in dreierlei Hinsicht einen wichtigen Referenzpunkt dar: Zum einen als ausgewiesenes Tagset für die moderne Standardsprache, das die Interoperabilität mit einem komplexen Gefüge an Werkzeugen sowohl zur Wortartenauszeichnung als auch zur darauf aufbauenden Auszeichnung syntaktischer und anderer Strukturen. Zum anderen ist das STTS Ausgangspunkt für Arbeiten jenseits der geschriebenen Standardsprache, die standardsprachliche Konstrukte im Sinne der ursprünglichen Richtlinien annotieren und nur dort abweichen, wo Phänomene in der Standardsprache der Gegenwart untypisch sind oder als ungrammatisch gelten (Buchstabierungen in der gesprochensprachlichen TüBa-D/S; auseinandergeschriebene Komposita in den frühneuhochdeutschen Texten der Mercurius-Baumbank, siehe Pauly et al., 2012; zu weiteren Beispielen siehe weitere Artikel dieser Ausgabe).

Weiterhin dient das STTS solchen Annotationsvorhaben als Referenzpunkt, die aufgrund ihrer unterschiedlichen Fragestellung eine andere Granularität der Tags anstreben. Beispiele hierfür sind das Historische Tagset (HiTS; Dipper et al., diese Ausgabe), das feingranulare Wortartentags für die Analyse früherer Sprachstufen des Deutschen bereitstellt, oder das sprachübergreifende Tagset von Petrov et al. (2012), das zur Vereinheitlichung zwischen Sprachen eine wesentlich gröbere Granularität als das STTS verwendet.

In diesem Artikel soll es darum gehen, eine Bestandsaufnahme des STTS vor allem in der Rolle als Tagset für Standardsprache, insbesondere anhand der in Tübingen erstellten Korpora, vorzunehmen. Eine solche Bestandsaufnahme soll verdeutlichen, welche Aspekte neben der deskriptiven Adäquatheit und der grundsätzlichen Anwendbarkeit wichtig sind, aber nur langfristig durch kontinuierliche Inspektion und Revision sichtbar werden.

Die Frage, was jenseits der ursprünglichen Tagsetdefinition zu einer konsistenten Anwendung des STTS gehört, reicht dabei hinein in die ebenfalls wichtige Frage der Interoperabilität mit bestehenden Werkzeugen und Ressourcen, die sich realiter auf eine bestimmte Ausdeutung des Standards bezieht und über die ursprünglichen Richtlinien hinausgeht.

Auch wenn die technischen Voraussetzungen gänzlich andere sind als bei Entstehung des STTS — Verfahren zur automatischen Verfeinerung von Tags durch unüberwachtes Lernen (Huang et al., 2009) beziehungsweise zum Tagging mit Verfeinerungen des STTS (Schmid und Laws, 2008; Müller et al., 2013) gehören mittlerweile zum "state of the art" der Standardsprache — profitieren auch (oder gerade) neuere Methoden sowohl von der Menge als auch der Konsistenz bestehender annotierter Daten (vgl. Manning, 2011).

Im Folgenden geben wir einen kurzen Überblick über das Stuttgart-Tübingen Tagset (Abschnitt 2), um dann einen Überblick über die Verwendung des STTS in Tübinger Korpora verschiedener Genres zu geben (Abschnitt 3). In Abschnitt 4 wird der Frage nachgegangen, welche Verwechslungen von POS-Tags in der Annotation auftreten, die durch einen langjährigen Revisionsprozess wie den der TüBa-D/Z zutage treten. Diese Art Datengrundlage bildet einen Kontrast zu Studien, in denen nur der erste Schritt einer Nachkorrektur automatisch zugewiesener POS-Tags ausgeführt wird und die kaum auf Fragen weniger offensichtlicher Ambiguitäten eingehen können. Abschnitt 5 enthält abschließende Betrachtungen.

2 Das Stuttgart-Tübingen Tagset (STTS)

Das Stuttgart-Tübingen Tagset entwickelte sich als allgemein akzeptierter Vorschlag zur Auszeichnung von Wortarten in den Projekten Elwis (Tübingen) und TC (Stuttgart) (Thielen und Schiller, 1994), nachdem das 1980 im SFB 100 "Elektronische Sprachforschung" entstandene SADAW/SATAN/SALEM-System aus linguistischen wie Performanzgesichtspunkten verworfen wurde (Hinrichs et al., 1995). Innerhalb des Elwis-Projekts wurde das STTS-Tagset unter anderem zur Auszeichnung von Text aus deutschen Newsgruppen verwendet (Feldweg et al., 1995).

Im Gegensatz zu komplexen morphosyntaktischen Tagsets beschränkt sich das Stuttgart-Tübingen Tagset auf eine Bestimmung der Wortart, während weitergehende morphosyntaktische oder semantische Information in Baumbanken wie TIGER oder TüBa-D/Z in einer separaten Annotationsebene (in TüBa-D/Z: Morphologie, Lemmata, Eigennamen-Ebene) kodiert ist. Auch das STTS macht im sogenannten "großen Tagset" (Schiller et al., 1999) einen Vorschlag, wie Morphologie und Derivation zu repräsentieren sind, der auf den Tags des üblicherweise verwendeten "kleinen Tagsets" aufbaut.

In Verarbeitungstools, die weitergehende morphosyntaktische Information nutzen, wie etwa dem RFTagger (Schmid und Laws, 2008) oder dem unlexikalisierten PCFG-Parser von Versley (2005), die feinere Unterscheidungen produzieren, wird üblicherweise ein hierarchisches Tagset verwendet, dessen Information sich mit wenig Mehraufwand in ein STTS-konformes POS-Tag und weitere morphologische Information splitten lässt.

3 STTS in der Praxis: Tübinger Baumbanken und annotierte Korpora

Die auf Basis des STTS getaggten Tübinger Ressourcen unterteilen sich in ausschließlich automatisch annotierte Korpora sowie manuell annotierte Korpora, bei denen Wortarten automatisch vorannotiert und anschließend in mehreren Durchgängen manuell korrigiert

wurden. Das Wortartentagging in den Tübinger Ressourcen wurde mit größtmöglicher Anlehnung an die Definitionen des STTS-Tagsets (Schiller et al., 1999) durchgeführt. Dabei sind nur wenige teilweise korpusspezifische Änderungen vorgenommen worden.

3.1 Überblick über die nach dem originalen STTS-Tagset getaggten Tübinger Ressourcen

In diesem Abschnitt werden die gemäß des STTS getaggten Tübinger Ressourcen vorgestellt.[1]
Folgende Ressourcen sind ausschließlich automatisch annotiert worden:

- Das *Tübinger Partiell Geparstes Korpus des Deutschen / Zeitungskorpus* – TüPP-D/Z (Müller, 2004) wurde mit einem auf dem TnT-Tagger (Brants, 2000) basierten Ensemble-Tagger automatisch annotiert. Die Daten bestehen aus einer Sammlung von Artikeln aus der Zeitung "*die tageszeitung*" (taz), mit einem Umfang von mehr als 200 Millionen Wörtern. Die Textdaten sind der Wissenschaftsausgabe der taz aus dem Jahr 1999 entnommen.

- *web-news* (Versley und Panchenko, 2012) wurde mit dem RFTagger (Schmid und Laws, 2008) und dem MaltParser (Hall et al., 2006) automatisch annotiert. Die Tags des RFTaggers wurden hierbei nachträglich in Wortarten nach STTS und Morphologie-Tags konvertiert. Das Korpus besteht aus 1,7 Milliarden Wörtern, die Nachrichten- und Blogsites im WWW entstammen.

- Die *Tübinger Baumbank des Deutschen / Diachrones Corpus* - TüBa-D/DC (Hinrichs und Zastrow, 2012) wurde mit dem TreeTagger (Schmid, 1995) automatisch annotiert. Es enthält mehr als 250 Mio. Wörter, deren Quelle das Projekt Gutenberg-DE ist.[2]

Folgende Ressourcen sind automatisch annotiert und manuell bearbeitet worden:

- Die *Tübinger Baumbank des Deutschen / Spontansprache* – TüBa-D/S (Hinrichs et al., 2000) ist ein syntaktisch annotiertes Korpus, das aus ca. 38.000 Sätzen besteht. Sie wurde im Projekt Verbmobil (maschinelle Übersetzung von Spontansprache) erstellt und hat spontansprachliche Dialoge als Datenbasis.

- Die *Tübinger Baumbank des Deutschen / Zeitungskorpus* – TüBa-D/Z (Telljohann et al., 2012) ist ein linguistisch annotiertes Korpus, das derzeit ca. 75.400 Sätze umfasst. Die Daten basieren auf Artikeln aus der deutschen Zeitung "*die tageszeitung*" (taz).

[1] Weitere Informationen und Lizensierungsmöglichkeiten der Tübinger Baumbanken sind verfügbar unter `http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora.html`.

[2] Gutenberg-DE: siehe `http://gutenberg.spiegel.de/`. Ein manuell bearbeitetes Sample aus der TüBa-D/DC bestehend aus ca. 3.800 Sätzen aus insgesamt sechs Texten unterschiedlicher Epochen wurde zur internen Evaluation der automatischen Annotation der TüBa-D/DC verwendet.

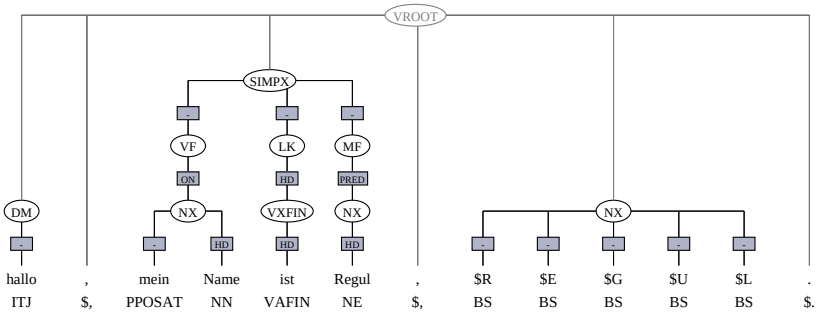


Abbildung 1: Baumbeispiel aus der TüBa-D/S

3.2 Die manuell bearbeiteten Baumbanken TüBa-D/S und TüBa-D/Z

Die Baumbanken TüBa-D/S und TüBa-D/Z sind manuell erstellte, syntaktisch annotierte Korpora. Die TüBa-D/S ist als Teil des im Jahr 2000 abgeschlossenen Projekts *Verbmobil* entwickelt worden (Hinrichs et al., 2000). Das in der TüBa-D/S verwendete Annotationsschema diente als Grundlage für die TüBa-D/Z. Da es sich bei der TüBa-D/S um Dialogdaten gesprochener Sprache handelte, musste das TüBa-D/S-Annotationsschema für die Erfordernisse von Zeitungstexten an die Charakteristika geschriebener Sprache angepasst und erweitert werden.

Die linguistische Annotation beider Baumbanken umfasst neben der hier diskutierten Wortannotation weitere (syntaktische, semantische und diskursbezogene) Annotationsebenen, für die wir aus Platzgründen auf die Annotationsmanuals von Stegmann et al. (2000) sowie Telljohann et al. (2012) verweisen.

In den manuell bearbeiteten Tübinger Korpora konnten alle Tokens eindeutig einem STTS-Tag zugeordnet werden. Es gab nahezu keinen Bedarf an weiteren, bisher nicht enthaltenen Tags (einzige Ausnahme: BS (Buchstabe) in der TüBa-D/S) oder an feineren Unterscheidungen der vorhandenen Tags.

Die primäre Segmentierungseinheit der TüBa-D/S besteht in Äußerungen (Dialog-Turns), da im Gegensatz zu schriftsprachlichen Korpora die Charakteristika gesprochener Sprache (z. B. Sprechfehler, Wiederholungen oder 'false starts") berücksichtigt werden müssen. Abbildung 1 ist ein Beispiel aus der TüBa-D/S. Der Dialog-Turn besteht aus einem Diskursmarker (DM), einem Satz (SIMPX) sowie einer Nominalphrase (NX). Der buchstabierte Name in der Nominalphrase (*R-E-G-U-L*)[3] zeigt die Verwendung des POS-Tags BS (Buchstabe[4]), um das das STTS-Tagset erweitert wurde. Die Satzstruktur des Beispiels ist aufgebaut aus Tokens, Phrasen (NX, VXFIN – *finite Verbphrase*),

[3] Entsprechend den in Verbmobil verwendeten Konventionen sind die Buchstaben mit $ markiert.

[4] In den Terminabsprache-Dialogen von Verbmobil ist Buchstabierung vergleichsweise häufig, weist dabei gleichzeitig eine Struktur auf, die als Kette von Nomina oder Nichtworten weniger adäquat abgebildet würde.

topologischen Feldern (VF – *Vorfeld*, LK – *linke Satzklammer*, MF – *Mittelfeld*) und dem Satz (SIMPX).[5]

Die TüBa-D/Z enthält über die TüBa-D/S hinaus Ebenen mit relevanten Merkmalen der Flexionsmorphologie, mit Lemma-Informationen, sowie auf syntaktischer Ebene eine Named-Entity-Kennzeichnung mit semantischen Klassen (s. Telljohann et al., 2012).

Der Beispielbaum aus der TüBa-D/Z in Abbildung 2 (unten) enthält neben der syntaktischen Struktur und den mit POS-Tags gekennzeichneten Tokens auch die Ebenen der morphologischen Annotation (z. B. nsm – *Nominativ, Singular, Maskulin*, 3sis – *3. Person, Singular, Indikativ, Präsens*) sowie der Lemmata (z. B. 'gelten').

3.3 Anwendungsunterschiede von POS-Tags in der TüBa-D/Z und der TIGER-Baumbank

In diesem Abschnitt werden ausgewählte Beispiele von Anwendungsunterschieden der STTS-POS-Tags in der TüBa-D/Z und der in Stuttgart entwickelten TIGER-Baumbank (Brants et al., 2002) demonstriert, die beide als theorieneutrale Baumbanken auf dem STTS-Tagset basieren. Version 2.1 der TIGER-Baumbank umfasst ca. 50.000 Sätze. Als Datenmaterial liegt ihr die Tageszeitung *Frankfurter Rundschau* zugrunde. Auf der syntaktischen Ebene unterscheiden sich die beiden Baumbanken im wesentlichen in der Behandlung der relativ freien Wortstellung des Deutschen und der Darstellung diskontinuierlicher Konstituenten: Topologische Felder und ein kontextfreies Gerüst ohne kreuzende Kanten in der TüBa-D/Z; dagegen eine weniger eingeschränkte Baumstruktur ohne topologische Felder, die kreuzende Kanten zulässt, in der TIGER-Baumbank.

Bei den POS-Tags wenden beide Baumbanken z. B. die attribuierenden Indefinitpronomina PIDAT und PIAT unterschiedlich an. Das STTS (Schiller et al., 1999, S. 41) definiert für die attribuierenden Indefinitpronomina als Kriterium, ob das Indefinitpronomina mit direkt vorangehendem oder folgendem Determiner auftreten kann (PIDAT) oder nicht (PIAT): Beispiel ohne Determiner: *etliche*/PIAT *Dinge*, *zuviele*/PIAT *Fragen*; mit möglichem Determiner: *all*/PIDAT *die Bücher*, *beide*/PIDAT *Fragen*; *op.cit.*, S. 41).

Satz (1a) zeigt einen Satz aus der TüBa-D/Z mit dem POS-Tag PIDAT: *Eine solche*/PIDAT *Veranstaltung* ... Attribuierende Indefinitpronomen, die nicht neben einem Determiner auftreten können, werden gemäß STTS als PIAT getaggt, wie beispielsweise *keine*/PIAT *Chance* ... in Satz (1b). In der TIGER-Baumbank hingegen wird (möglicherweise als Konzession an den Annotationsprozess) keine Unterscheidung zwischen PIAT und PIDAT gemacht, da die Unterscheidung des STTS zwischen PIAT und PIDAT anhand der Wortform stets rekonstruierbar ist. Stattdessen wird PIAT für attribuierende Indefinitpronomina mit und ohne Determiner verwendet, wie Satz (2) demonstriert: *ein solches*/PIAT *Verhalten*.

[5] Der virtuelle `VROOT`-Knoten in den Abbildungen hat lediglich die formale Funktion, alle Knoten des Satzes unter ein gemeinsames Element zu fassen.

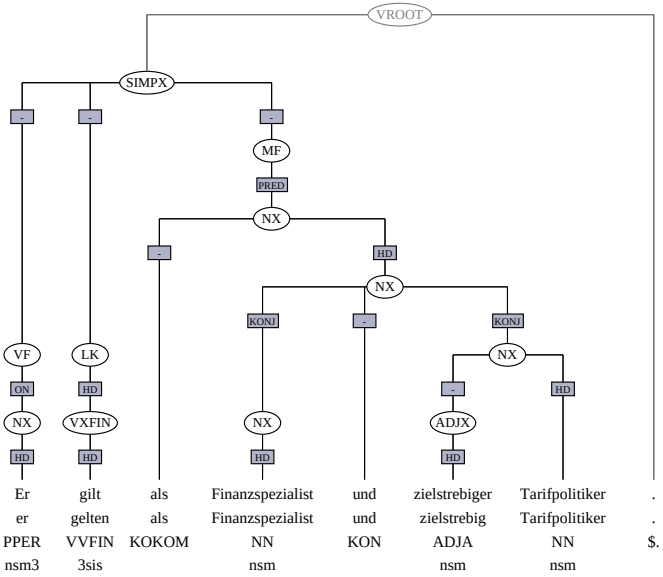


Abbildung 2: POS-Tagging des nichtkomparativen *als* in der TüBa-D/Z

(1) TüBa-D/Z:
a. *Eine solche*/PIDAT *Veranstaltung werden wir leider wiederholen müssen.*
b. *Die Opposition hat keine*/PIAT *Chance, die Mehrheit zu bekommen..*

(2) TIGER-Baumbank:
Was bewirkt Ihrer Ansicht nach ein solches/PIAT *Verhalten?*

Einen weiteren Anwendungsunterschied finden wir beim Tagging von nichtkomparativen *als*-Phrasen. Dagegen wird das Komparativ-*als* in beiden Baumbanken als Vergleichspartikel (KOKOM) getaggt, z. B. "*größer als*/KOKOM *100 Hektar*" (TIGER), "*länger als*/KOKOM *fünf Jahre*" (TüBa-D/Z). Im STTS (Schiller et al., 1999, S. 62) werden ausschließlich *als* und *wie* als KOKOM definiert. Das POS-Tag KOKOM umfasst alle *als* und *wie*, die nicht satzeinleitend verwendet werden. Eine weitere Einteilung von KOKOM in Partikel mit Vergleichssemantik und ohne Vergleichssemantik wird hier nicht getroffen, da diese Unterscheidungen vage sind. Das STTS gibt u.a. folgende Beispiele für KOKOM: *er gilt als*/KOKOM *fleißig*; *entpuppte sich als*/KOKOM *stimmenträchtiges Zugpferd*; *er arbeitet als*/KOKOM *Bauer* (*op. cit.*, S. 62). Gemäß dieser Definition sind nichtkomparative *als*-Phrasen in der TüBa-D/Z mit *als* als KOKOM annotiert. Die syntaktische Kategorie der jeweiligen *als*-Phrase wird von der enthaltenen Phrase bestimmt, z. B. durch eine Nominalphrase (*Finanzspezialist und zielstrebiger Tarifpolitiker*) wie

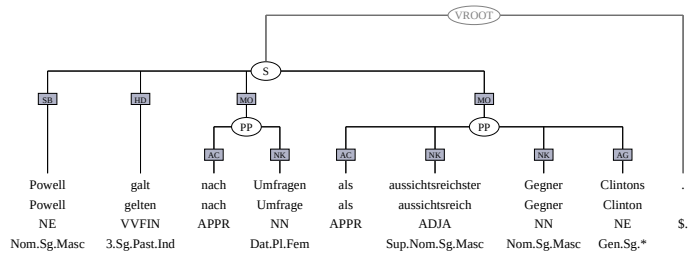


Abbildung 3: POS-Tagging des nichtkomparativen *als* in der TIGER-Baumbank

in Abbildung 2. In der TIGER-Baumbank hingegen wird bezüglich komparativen und nichtkomparativen *als*-Phrasen eine Unterscheidung getroffen. Nichtkomparative *als*-Phrasen sind hier Präpositionalphrasen (PP) und werden, vom STTS abweichend, mit *als* als APPR (Präposition) getaggt, wie in Abbildung 3 gezeigt wird.

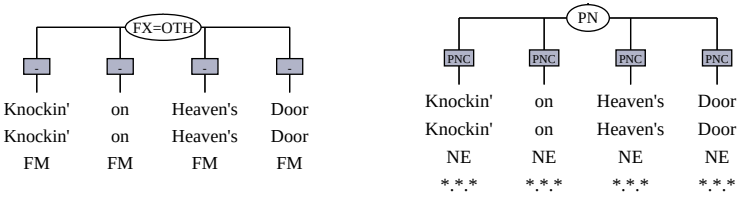


Abbildung 4: Named Entity in der TüBa-D/Z (links) und der TIGER-Baumbank (rechts)

Auch eine unterschiedliche Kategorisierung fremdsprachlicher Named Entities führt dazu, dass derselbe Eigenname in beiden Baumbanken verschiedene POS-Tags erhält. Das STTS (Schiller et al., 1999, S. 75) definiert als fremdsprachliches Material (FM) größere Textstücke, die einer fremden Sprache angehören und nicht als Eigennamen (NE) klassifiziert werden können. Als Beispiele werden neben fremdsprachlichen Ausdrücken wie *lazy*/FM auch fremdsprachliche Filmtitel aufgeführt: *mujer*/FM *de*/FM *Benjamin*/NE und *A*/FM *fish*/FM *called*/FM *Wanda*/NE. Die enthaltenen Eigennamen, die als solche erkannt werden, werden als NE getaggt.

Entsprechend sind in der TüBa-D/Z z. B. fremdsprachliche Buch- und Filmtitel als FM getaggt, im Gegensatz zu Firmen- oder Bandnamen, welche mit dem Tag NE gekennzeichnet sind. Die gesamte Phrase wird mit einem Knotenlabel versehen, das Information über die semantische Klasse des Eigennamens liefert, wie z. B. FX=OTH (other) für den Filmtitel *Knockin' on Heaven's Door* im linken Teil von Abbildung 4. Die TIGER-Baumbank verwendet dagegen für fremdsprachliche Filmtitel das POS-Tag NE. Derselbe Eigenname *Knockin' on Heaven's Door* weist dann eine Annotation mit NE für alle Tokens und dem Knotenlabel PN (proper noun) für die gesamte Phrase auf.

4 Part-of-Speech-Variation in einer Baumbank

Automatische – zum Teil auch manuelle – Annotation von Part-of-Speech-Tags birgt oft Fehler, im Sinne einer Abweichung von einer idealisierten 'wahren' Annotation.[6] Bei der Entwicklung oder Weiterentwicklung eines Tagsets wie des STTS kann es hilfreich sein, derartige Information einzubeziehen, um durch geeignete Vergröberungen oder Verfeinerungen Fehlerquellen auszuschließen, oder um die Beschreibung der Kategorien geeignet zu ergänzen.

4.1 Automatische und manuelle Fehlererkennung

Bestehende Arbeiten zur Fehlerkorrektur in handannotierten Daten, wie etwa van Halteren (2000); Květoň und Oliva (2002); Dickinson und Meurers (2003) stützen sich wesentlich auf das Prinzip der Konsistenz: über verschiedene ähnliche Kontexte hinweg soll das gleiche Tag verwendet werden. Ähnliche Kontexte werden in diesem Fall (wie bei Dickinson und Meurers) als sich wiederholende Wortsequenzen (n-Gramme) definiert, oder als Kontext, der für die Entscheidung eines statistischen POS-Taggers relevant ist (van Halteren). Loftsson (2009) stellt ein neueres Beispiel für die Anwendung dieser Ansätze dar. In Abwesenheit von Evidenz für "mögliche" und "unmögliche" Taggings (wie von Květoň und Oliva als Eingabe für ihr Verfahren gefordert) stellen Konsistenzkriterien in der Regel die einzige Handhabe dar, um Fehler in Korpora — oder zumindest Zweifelsfälle — ohne weitere Hilfsmittel zu finden.

Eine Garantie der Übereinstimmung mit formalen Richtlinien bieten derartige Konsistenzprüfungen allerdings nicht, letztendlich ist die Beurteilung durch menschliche Experten notwendig. Im Fall von Dickinson (2006), der einen Algorithmus zur Korrektur von n-gram-Varianten vorstellt, wird dieser durch die Neuannotation von 300 Tokens in 'verdächtigen' Kontexten validiert, die durch die Methode der Tagvariation in gleichen n-Grammen gefunden wurden.

Rein automatische Verfahren zum Auffinden von Kandidaten für falsche Tags wie das der Tagvariation in n-Gram-Kontexten sind für bestimmte Fehlerarten blind: Wenn zwei unterschiedliche Wortformen zueinander inkonsistent getaggt werden, oder eine Wortform nicht mehrfach in demselben Typ von Kontext vorkommt, kann dieser Fehler nicht durch den Variationsansatz entdeckt werden.

Die laufende manuelle Revision — sei es eine kritische Durchsicht mit Fokus auf bestimmte Phänomenbereiche oder das Finden von Fehlern bei weiterer Annotation — kann solche Fehler durchaus entdecken und ist gegenüber Selektionsverzerrungen weniger anfällig.

Weitere Hinweise auf mögliche oder tatsächliche POS-Fehler ergibt der Abgleich der Part-of-Speech-Annotation mit anderen Annotationsebenen durch Abfragen, die POS

[6] Wir nehmen vereinfachend an, dass es zu einem gegebenen Tagset genau eine intendierte Interpretation gibt, zu der hin — genügend Zeit und Aufwand vorausgesetzt — die Annotation konvergieren würde. Manning (2011) merkt mit Bezug auf transitive Adjektive in der *Penn Treebank* an, dass eine solche Interpretation dort, wo sich Grammatiken nicht einig sind, zwar keine absolute Wahrheit, aber doch zumindest Konsistenz für sich beanspruchen kann.

und die syntaktische Struktur zusammen mit einbeziehen, insbesondere in Fällen, wo Synkretismen nur durch Abhängigkeit vom syntaktischen Kontext aufgelöst werden können.

Ein Beispiel für solche Synkretismen findet sich bei Verbformen, die ambig sind zwischen finiter (V·FIN) und nicht-finiter Verwendung (V·INF): in Trigrammen wie "*Schwierigkeiten gelernt haben*$_{\{\text{VAFIN,VAINF}\}}$" ist zwar Variation feststellbar, die Ambiguität aber ohne Berücksichtigung des strukturellen Kontextes nicht auflösbar.

In der TüBa-D/Z werden diese Ambiguitätsklassen anhand der Felderstruktur des Satzes behandelt: zum einen darf ein Satz in dessen linker und rechter Satzklammer (LK, VK) nur ein einziges finites Verb enthalten. Sätze mit Komplementierer-Feld (C) sind immer finite Verbletztsätze, so dass auch hier Ambiguitäten durch strukturelle Eigenschaften aufgelöst werden können.

Die Unterscheidung zwischen Relativsätzen (R-SIMPX) und anderen Arten von Sätzen (SIMPX) dient dazu, Verwendungen von "*was*" und "*welche*" als Relativpronomina zu überprüfen, die sonst schwer zu identifizieren wären.

Im Fall von Postpositionen und Zirkumpositionen ("*der Reihe nach*/APPO", "*von Anfang an*/APZR") hilft die syntaktische Struktur, Fälle zu erkennen, bei denen Appositionen am Ende falsch getaggt sind. Die Anbindung von Adjektiven liefert Hinweise, ob es sich um ein attributives Adjektiv (mit Anbindung an eine Nominalphrase) oder um ein prädikatives Adjektiv (Anbindung im Satz oder an eine Adverbialphrase) handelt.

In manchen Fällen lassen sich anhand der Wortform oder anhand von Reihenfolgebeschränkungen sinnvolle Konsistenzkriterien für Part-of-Speech-Tags bestimmen (etwa: Zu-Infinitive – VVIZU – müssen ein *zu* enthalten), in vielen Fällen werden POS-Fehler bei der Annotation von morphologischer oder Lemma-Information aufgedeckt.

Zusammenfassend sei festgestellt: es existieren eine ganze Anzahl von Ansätzen, die helfen, Fehler in POS-Annotationen zu finden oder auch Richtlinien für Annotatoren zu präzisieren. Diese Ansätze reichen von solchen, die wie der Ansatz von Dickinson weitgehend ohne Annahmen über Tagset oder sprachliche Struktur auskommen, bis hin zu solchen, die auf expliziten Annahmen (Květoň und Oliva) oder Struktur auf anderen Annotationsebenen (TüBa-D/Z) basieren. Für die Annotation neuer Korpora mit POS-Tags ist es interessant, eine möglichst verzerrungsfreie Abschätzung zu bekommen, wo und welche Fehler/Unsicherheiten zu erwarten seien, wie auch die Frage, welche Mittel helfen können, um von einer (manuellen oder semi-automatischen) Erstannotation zu einem fehlerärmeren Korpus zu kommen.

4.2 Verwendete Ressourcen

In diesem Abschnitt stellen wir eine Studie vor, die Daten verschiedener Versionen der TüBa-D/Z-Baumbank vergleicht und so eine große Stichprobe von Korrekturen liefert, wie sie durch manuelle Revision offenbar werden. Als Material benutzen wir den Text der ersten 766 Artikel der TüBa-D/Z, auf denen das erste Release der Baumbank beruht und die in den folgenden Releases (in korrigierter Form) enthalten sind. Dieser Vergleich erlaubt es uns, alle Änderungen in Folge von manuellen oder

Name	Beschreibung	#Tokens (766 Art.)
tueba1	TüBa-D/Z Release 1	266 441
tueba5	TüBa-D/Z Release 5	266 646
tueba8	TüBa-D/Z Release 8	266 665
treetagger	R8 / TreeTagger	266 665
pcfg	R8 / PCFG+SMOR	266 665

Tabelle 1: Getaggte Varianten der TüBa-D/Z-Texte

semiautomatischen Inspektionen der Baumbank — einschließlich Inkonsistenzen, die bei der Arbeit an anderen Annotationsebenen wie Koreferenz, Lemmatisierung oder Diskurs entdeckt wurden — zu erkennen und in Bezug zu der Menge und Art von Änderungen zu setzen, die insgesamt nötig wären, um von automatisch zugewiesenen POS-Tags zum "Gold"standard des neuesten Release der TüBa-D/Z zu gelangen.

Wie in Tabelle 1 ersichtlich, weicht die Tokenisierung, und mit ihr die Anzahl Tokens, zwischen verschiedenen Varianten der Baumbank geringfügig (um ca. 0.1%) voneinander ab. Hintergrund ist vor allem die Nachtragung von Zwischenüberschriften und Bildunterschriften, die in Release 1 fehlen; von Release 5 zu Release 8 umfassen die Unterschiede vor allem die Trennung von "z.B." in zwei Tokens und die Umtokenisierung von Zahlenbereichen wie "2-3" in einzelne Tokens.

Zum Vergleich mit automatischen Methoden der Zuweisung von Part-of-Speech-Tags wurde die Release-8-Version des Baumbankabschnitts zusätzlich durch zwei automatische Systeme getaggt:

- **treetagger** benutzt den TreeTagger (Schmid, 1995) mit dem Standardmodell und dem in der TreeTagger-Distribution enthaltenen Skript zur Tagkorrektur bei VVFIN/VVINF-Fehlern.

- **pcfg** benutzt ein unlexikalisiertes PCFG-Modell ähnlich dem von Versley (2005), bei dem Part-of-Speech-Tags für unbekannte Wörter durch eine Kombination von SMOR (Schmid et al., 2004) und Gazetteer-Listen vorhergesagt werden. Das PCFG-Modell wurde auf den Sätzen 15266ff. trainiert, die nicht in der zum Testen verwendeten Portion (entsprechend Release 1 der Baumbank) enthalten sind.

4.3 Diskussion

Die Abbildungen 5 bis 7 (folgende Seiten) veranschaulichen Änderungen von jeweils einer getaggten Version der Texte (die ersten 766 Artikel der TüBa-D/Z, in den oben erwähnten Versionen) im Vergleich zum Tagging in Release 8 der Baumbank als gerichtete Graphen. Die Kanten der Graphen zeigen jeweils die Anzahl der von alter zu neuer Version geänderte Tokens (R1/R8) an, beziehungsweise die Fehler, die von einem automatischen Tagger im Vergleich zur Referenzversion (R8) vorliegen.

Zwischen den POS-Kategorien wurde eine Kante eingefügt, sobald eine Mindestanzahl von geänderten Tokens erreicht wurde (Baumbank: 15; TreeTagger/PCFG: 70).

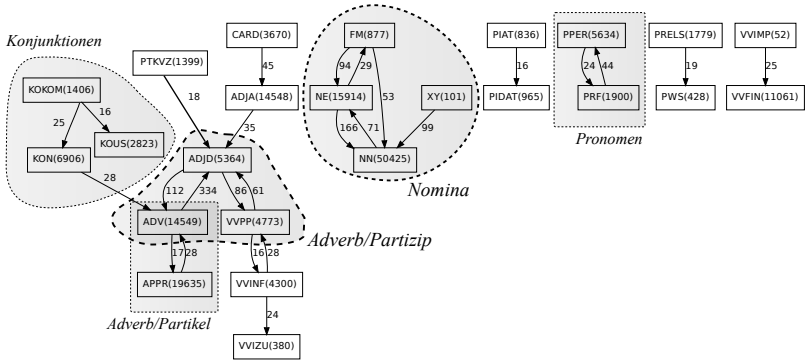


Abbildung 5: Häufigste POS-Änderungen zwischen Release 1 und Release 8

Man erkennt mehrere Cluster, innerhalb derer Wörter im POS-Tag ambig sein können:

- **Nomina:** oft sind neue oder unbekannte Worte ambig zwischen Appelativum (NN), Eigenname (NE) oder fremdsprachlichem Material (FM). Die Kriterien des STTS legen in vielen Fällen fest, wie mit Zweifelsfällen zu verfahren ist. Im Fall von Ambiguität zwischen Kategoriename und Firmenname oder Produktbezeichnung (*Bahn*) sowie im Fall von fremdsprachlichen Namen (*Pastoral Way*, *Kitty Yo*) sind hier jedoch im Einzelfall Festlegungen erforderlich. Einzelne Nomina wie *Ageism*(FM), *Carnigglio*(FM) oder *Fingerfood*(NN) zeigen Zwischenstadien zwischen fremdsprachlichem und nativem Gebrauch, wofür Schiller et al. (1999) Flektierbarkeit und (in der Gebersprache unübliche) Großschreibung als Anhaltspunkte geben.

- **Adverbiale:** Wörter, die adverbial gebraucht werden, sind gegebenenfalls ambig zwischen Adverb (ADV), adverbialem/prädikativem Adjektiv (ADJD) und partizipialer Verbform (VVPP).

- **Verbformen:** Bei Vollverben sind oft finite Verbformen (VVFIN) sowie nichtfinite Verbformen (VVPP, VVINF) gleich in der Oberflächenform. Diese Fälle sind nicht im eigentlichen Sinne ambig, da das korrekte Tag im Kontext eindeutig sein sollte.

- **Pronomen:** In vielen Formen (*mir*, *uns*) sind Akkusativ- und Dativ-Pronomen ambig zwischen reflexivem Pronomen (PRF) und nicht-reflexivem Personalpronomen (PPER). Welche der beiden Möglichkeiten zutrifft, hängt von der Modellierung des Argumentrahmens des regierenden Verbs ab. Dementsprechend gibt es in dieser Kategorie Fälle, in denen linguistische Expertise vonnöten ist.

- **Konjunktionen:** Wörter wie *wie*, *wo* und *als* können satzeinleitend verwendet werden (KOUS), tauchen jedoch auch regulär als Vergleichspartikel (KOKOM:

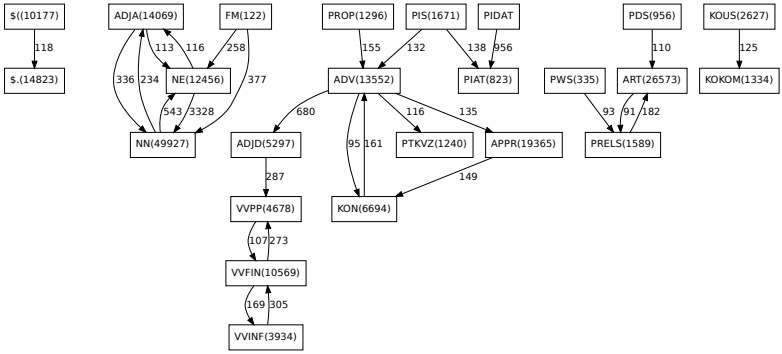


Abbildung 6: Häufigste POS-Fehler des TreeTagger

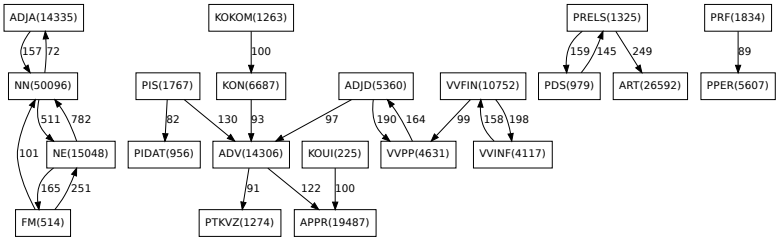


Abbildung 7: Häufigste POS-Fehler des PCFG-Parsers

wie, *als*) oder als Frageadverb (PWAV: *wie*, *wo*) auf. Auch hier ist eine genaue Betrachtung des syntaktischen Kontext erforderlich, um eigentlich nicht-ambige Fälle korrekt zuzuordnen.

- **Verbpartikel:** Zwischen Funktionsverbgefügen mit Adverb ("*klar machen*", "*bekannt werden*") und Partikelverben mit untypischem Verbpartikel ("*klarmachen*", "*bekanntwerden*") besteht ein gewisser Graubereich, innerhalb dessen sowohl eine Lesart als Funktionsverbgefüge als auch die Verbpartikel-Lesart vom Autor verwendet und vom Leser auch in nicht ambigen Konstruktionen als akzeptabel empfunden werden. In V2-Sätzen sind diese Konstruktionen ambig zwischen beiden syntaktischen Lesarten.

alt	neu	Anz.	Wörter
ADV	ADJD	334	künftig (24), täglich (22), völlig (17)
ADJD	ADV	112	genau (13), wirklich (11), scheinbar (7)
ADJD	VVPP	86	geplant (5), geboren (3), gegeben (2)
VVPP	ADJD	61	überzeugt (12), betroffen (6), geeignet (3)
NE	NN	166	Bahn (11), Ex-Jugoslawien (6), Bayern (4)
XY	NN	99	R (53), D (46)[a]
FM	NE	94	Way (9), Pastoral (8), Drum (4)
FM	NN	53	Fingerfood (3), Eyecatcher (3), GIs (3)
NE	FM	29	Underground (3), Sir (2), Yo (2)
VVINF	VVPP	28	bekommen (8), erfahren(2), entfallen (2)
VVIMP	VVFIN	25	Lesen (2), gehen (1), reichen (1)
VVINF	VVIZU	24	einzuhauchen (1), auszurüsten (1), aufzuklären (1)
VVPP	VVINF	16	gefallen (5), erhalten (3), enthalten (2)
APPR	ADV	28	über (16), mit (3), unter (2)
PTKVZ	ADJD	18	bekannt (6), bereit (5), ernst (2)
ADV	APPR	17	namens (8), über (1), Abseits (1)
PRF	PPER	44	uns (20), mich (20), mir (4)
PPER	PRF	24	mir (9), uns (5), dich (5)
PRELS	PWS	19	was (18), wer (1)
PIAT	PIDAT	16	wenige (5), beide (3), ebensoviele (1)
KON	ADV	28	etc. (9), usw. (5), Aber (4)
KOKOM	KON	25	wie (13), als (12)
KOKOM	KOUS	16	wie (12), Wie (2), als (2)

Tabelle 2: Wörter mit POS-Unterschieden zwischen Release 1 und Release 8

[a] Regie, Darsteller

Tabelle 2 fasst die wichtigsten Kategorien von Tag-Änderungen zusammen und listet die Wortformen, die am häufigsten mit dieser Änderung vorliegen.

Tabelle 3 enthält eine quantitative Auswertung der Übereinstimmungen und Unterschiede zwischen Release 1/Release 5 und Release 8 einerseits, sowie der automatisch getaggten Varianten mit der Gold-Annotation in Release 8.

Name	Acc. (%)	ADJD	ADV	FM	NE	VVIMP	VVINF	VVIZU	VVPP
tueba1	99.22	0.94	0.98	0.90	0.99	0.67	0.99	0.96	0.98
tueba5	99.79	0.99	0.99	0.93	0.99	0.97	1.00	1.00	1.00
treetagger	94.99	0.87	0.94	0.22	0.84	0.38	0.93	0.99	0.92
pcfg	97.30	0.92	0.97	0.61	0.94	0.56	0.95	1.00	0.94

Tabelle 3: Quantitativer Vergleich zwischen Release 8 und anderen Varianten (F_1 für einzelne Tagkategorien)

Schaut man sich die quantitative Auswertung dieses Vergleichs an, so ist offensichtlich, dass insbesondere seltene Kategorien wie FM und VVIMP für das automatische Tagging problematisch sind und nicht immer zuverlässig erkannt werden.

Infolgedessen sind ambige Formen ("*Geht*", "*Fragt*") automatisch nicht gut zu desambiguieren, während das korrekte Tag für linguistische Experten zweifelsfrei ist.

5 Abschließende Betrachtung

In diesem Artikel stellten wir eine Betrachtung des Stuttgart-Tübingen Tagset (STTS) in dessen Anwendung auf verschiedene Tübinger Ressourcen vor, unter besonderer Berücksichtigung von Faktoren, die bei der Annotation neuer Korpora von Interesse sein dürften. In einem zweiten Teil eine Auswertung der Typen von Tag-Änderungen, die in der laufenden Revision der geschriebensprachlichen TüBa-D/Z aufgetreten sind.

Die vorgestellten empirischen Daten legen nahe, dass eine formbasierte Ausdeutung von Unterschieden dort, wo distributionelle Kriterien kein klares Bild ergeben, ein notwendiger Seiteneffekt der Annotation ist, dessen einzige Alternative unmotivierte Inkonsistenzen sind, da die distributionelle Evidenz gerade bei gradierten Unterschieden zuweilen ein unklares Bild ergibt. Der in der TüBa-D/Z verfolgte Ansatz, diese formbasierte Ausdeutung zu dokumentieren und in der Annotation von neuen Texten diese Information — sei es aus Vorkommen im bisher annotierten Korpus oder aus eigens zusammengestellten Tabellen — zu berücksichtigen, ist eine effektive Lösung für dieses Problem, bedeutet aber, dass der Standard in der Praxis (d.h. bei Korpusrecherchen oder in automatischer Tools) eine Ergänzung durch veröffentlichte Korpora erfährt.

Erweiterungsvorschläge für das STTS sind somit nicht allein mit Bezug auf die ursprünglichen Tagging-Richtlinien zu sehen, sondern auch in Bezug auf veröffentlichte Korpora, die eine Datenbasis für formbasierte Unterscheidungen darstellen, beziehungsweise Werkzeuge und Ressourcen, die diese Unterscheidungen prinzipbedingt umsetzen. Dies ist auch dann empfehlenswert, wenn Vorschläge zunächst nur als inkrementelle Erweiterung des STTS-Dokuments formuliert sind, da Interoperabilität und Konsistenz mit existierenden Ressourcen spätestens in der tatsächlichen Anwendung einen wichtigen, wenn auch oft unterschätzen, Aspekt darstellen.

Literatur

Brants, S., Dipper, S., Hansen, S., Lezius, W. und Smith, G. (2002). The TIGER treebank. In: *Proceedings of the First Workshop on Treebanks and Linguistic Threories*, Sozopol, Bulgarien.

Brants, T. (2000). TnT — A Statistical Part-of-Speech Tagger. In: *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000)*, Seiten 224–231.

Dickinson, M. (2006). From detecting errors to automatically correcting them. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Seiten 265–272, Trient, Italien.

Dickinson, M. und Meurers, D. (2003). Detecting errors in part-of-speech annotation. In: *Proceedings of EACL-2003*.

Feldweg, H., Kibiger, R. und Thielen, C. (1995). Zum Sprachgebrauch in deutschen Newsgruppen. *Osnabrücker Beiträge zur Sprachtheorie (OBST)*, 50:143–154.

Hall, J., Nivre, J. und Nilsson, J. (2006). Discriminative classifiers for deterministic dependency parsing. In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Seiten 316–323.

Hinrichs, E. W., Bartels, J., Kawata, Y., Kordoni, V. und Telljohann, H. (2000). The Tübingen Treebanks for Spoken German, English, and Japanese. In: Wahlster, W. (Hg.), *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin.

Hinrichs, E. W., Feldweg, H., Boyle-Hinrichs, M. und Hauser, R. (1995). Abschlußbericht ELWIS: Korpusgestützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik. Abschlussbericht für das Ministerium für Wissenschaft und Forschung Baden-Württemberg, Seminar für Sprachwissenschaft, Universität Tübingen.

Hinrichs, E. W., Kübler, S., Naumann, K., Telljohann, H. und Trushkina, J. (2004). Recent Developments of Linguistic Annotations of the TüBa-D/Z Treebank. In: *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen.

Hinrichs, E. W. und Zastrow, T. (2012). Linguistic Annotations for a Diachronic Corpus of German. *Linguistic Issues in Language Technology (LiLT)*, 7(7).

Huang, Z., Eidelman, V. und Harper, M. (2009). Improving a simple bigram HMM part-of-speech tagger by latent annotation and self-training. In: *Proceedings of the 2009 Annual Conference of the NAACL*, Seiten 213–216.

Květǒn, P. und Oliva, K. (2002). Achieving an almost correct pos-tagged corpus. In: Sojka, P., Kopeček, I. und Pala, K. (Hgg.), *Text, Speech and Dialogue: 5th International Conference, TSD 2002*, Band 2448 von *Lecture Notes in Computer Science*.

Loftsson, H. (2009). Correcting a POS-tagged corpus using three complementary methods. In: *12th Conference of the European Chapter of the ACL (EACL 2009)*.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In: *12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Seiten 171–189.

Müller, F. H. (2004). Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z). Technischer Bericht, Seminar für Sprachwissenschaft, Universität Tübingen.

Müller, T., Schmid, H. und Schütze, H. (2013). Efficient higher-order crfs for morphological tagging. In: *Proceedings of EMNLP 2013*, Seiten 323–332.

Pauly, D., Senyuk, U. und Demske, U. (2012). Strukturelle Mehrdeutigkeit in frühneuhochdeutschen Texten. *Journal for Language Technology and Computational Linguistics*, 27(2):65–82.

Petrov, S., Das, D. und McDonald, R. (2012). A universal part-of-speech tagset. In: *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

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

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In: *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Irland.

Schmid, H., Fitschen, A. und Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In: *Proceedings of LREC 2004*.

Schmid, H. und Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In: *Proceedings of COLING 2008*.

Skut, W., Krenn, B., Brants, T. und Uszkoreit, H. (1997). An annotation scheme for free word order languages. In: *5th Applied Natural Language Processing Conference (ANLP 1997)*, Seiten 88–95.

Stegmann, R., Telljohann, H. und Hinrichs, E. W. (2000). Stylebook for the German Treebank in VERBMOBIL. Verbmobil-Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H. und Beck, K. (2012). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technischer Bericht, Seminar für Sprachwissenschaft, Universität Tübingen.

Thielen, C. und Schiller, A. (1994). Ein kleines und erweitertes Tagset fürs Deutsche. In: Feldweg, H. und Hinrichs, E. (Hgg.), *Lexikon & Text*, Seiten 215–226. Niemeyer, Tübingen.

van Halteren, H. (2000). The detection of inconsistency in manually tagged text. In: *Proceedings of the COLING-2000 Workshop on Linguistically Annotated Corpora (LINC-00)*.

Versley, Y. (2005). Parser evaluation across text types. In: *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*.

Versley, Y. und Panchenko, Y. (2012). Not Just Bigger: Towards Better-Quality Web Corpora. In: *Proceedings of the 7th Web as Corpus Workshop at WWW2012 (WAC7)*, Seiten 44–52, Lyon, Frankreich.

Sandra Kübler, Wolfgang Maier

Über den Einfluss von Part-of-Speech-Tags auf Parsing-Ergebnisse

1 Einleitung

Lange Zeit konzentrierte sich die Forschung im datengetriebenen statistischen Konstituenzparsing auf die Entwicklung von Parsingmodellen für das Englische, genauer gesagt, für die Penn Treebank (Marcus et al., 1993). Einer der Gründe dafür, warum sich solche Modelle nicht ohne Weiteres auf andere Sprachen generalisieren lassen, ist die eher schwach ausgeprägte Morphologie des Englischen: Probleme, die sich bei Parsen einer morphologisch reichen Sprache wie z.B. Arabisch oder Deutsch stellen, stellen sich für das Englische nicht. Vor allem in den letzten Jahren erfuhr die Forschung zu Parsingproblemen, die sich auf komplexe Morphologie beziehen, ein gesteigertes Interesse (Kübler und Penn, 2008; Seddah et al., 2010, 2011; Apidianaki et al., 2012).

In einer Baumbank sind Wörter im allgemeinen Information annotiert, die Auskunft über die Wortart (Part-of-Speech, POS) und morphologischen Eigenschaften eines Wortes gibt. Wo, sofern vorhanden, die Trennlinie zwischen Wortart und morphologischer Information gezogen wird und wie detailliert annotiert wird, hängt von der Einzelsprache und dem Annotationsschema ab. In einigen Baumbanken gibt es keine separate morphologische Annotation (wie z.B. in der Penn Treebank), in anderen sind Part-of-Speech- und Morphologie-Tagsets getrennt (z.B. in den deutschen Baumbanken TiGer (Brants et al., 2002) und NeGra (Skut et al., 1997)), und in anderen ist wiederum nur ein Tagset vorhanden, das sowohl POS- als auch Morphologie-Information enthält (z.B. in der Szeged Treebank (Csendes et al., 2005)). Die Anzahl verschiedener Tags für Sprachen mit einer komplexen Morphologie kann in die Tausende gehen, so z.B. für Tschechisch (Hajič et al., 2000), während für die Modellierung der Wortarten von Sprachen mit wenig bis keiner Morphologie nur wenige Tags ausreichen, z.B. 33 Tags für die Penn Chinese Treebank (Xia, 2000). Wir schließen der Einfachheit halber alle Annotationstypen ein, wenn wir ab hier von *Part-of-Speech-Annotation* sprechen.

Die Part-of-Speech-Tags nehmen eine Schlüsselrolle beim Parsen ein als Schnittstelle zwischen lexikalischer Ebene und dem eigentlichen Syntax-Baum: Während des Parsingvorgangs wird der eigentliche Konstituenzbaum nicht direkt über den Wörtern, sondern über der Part-of-Speech-Annotation erstellt. Ein Part-of-Speech-Tag kann als eine Äquivalenzklasse von Wörtern mit ähnlichen distributionellen Charakteristika angesehen werden, die über die individuellen Wörter abstrahiert und damit die Anzahl der Parameter beschränkt, für die Wahrscheinlichkeiten gelernt werden müssen. Die eigentlichen Wörter finden bei lexikalisierten Parsern Eingang in das Wahrscheinlichkeitsmodell. Es ist offensichtlich, dass die Part-of-Speech-Annotation direkten Einfluss auf die Qualität des Parsebaums hat. Nicht nur die Qualität des Taggers spielt hierbei eine Rolle, sondern auch die Granularität des Tagsets an sich. Es muss ein Kompromiss

gefunden werden zwischen zu hoher Abstraktion, die wichtige Unterscheidungen unterdrückt, und zu hoher Detailgenauigkeit, die durch die Trainingsdaten nicht unbedingt abgedeckt ist.

In den letzten Jahren sind daher einige Arbeiten entstanden, die zum Ziel haben, sprachübergreifend die lexikalische Ebene zu einer möglichst sicheren Basis für die Erstellung eines Parsebaums zu machen. Diese reichen von der Einführung des Universal Tagsets (UTS) (Petrov et al., 2012), einem reduzierten, sprachübergreifenden Tagset, über Arbeiten zu lexikalischem Clustering, siehe z.B. Koo et al. (2008), bis zum Entwurf von faktorisierten Parsingmodellen, in denen versucht wird, Parsing und lexikalische Annotation separat voneinander zu modellieren, wie z.B. Chen und Kit (2011).

Wenig untersucht ist bis jetzt der Zusammenhang zwischen der Granularität des POS-Tagsets, verschiedenen Taggern und Parsingergebnissen innerhalb eines „Pipeline"-Ansatzes, in dem ein Parser die Ausgabe eines Taggers als Eingabe erhält. Im vorliegenden Artikel untersuchen wir diese Fragestellung anhand von verschiedenen Variationen des Stuttgart-Tübingen-Tagsets (STTS) (Schiller et al., 1995; Thielen und Schiller, 1994), das für die Annotation der beiden großen deutschen Baumbanken TiGer und TüBa-D/Z (Telljohann et al., 2012) verwendet wurde, und des Universal Tagsets (Petrov et al., 2012). Des Weiteren verwenden wir verschiedene POS-Tagger, namentlich TnT (Brants, 2000), SVMTool (Giménez und Màrquez, 2004), Mallet-CRF (McCallum, 2002) und Stanford-MaxEnt (Toutanova und Manning, 2000), sowie den Berkeley Parser (Petrov et al., 2006). In zwei Experimentgruppen betrachten wir den Einfluss der Granularität von STTS auf die Qualität von Tagger- und Parserausgaben. In zwei weiteren Experimentgruppen untersuchen wir verschiedene Untermengen von morphologischen Merkmalen des STTS, sowie den Einfluss der morphologischen Merkmalssets auf die Parsingqualität.

Der Artikel ist wie folgt strukturiert. Im folgenden Abschnitt stellen wir einen Ausschnitt der vorhandenen Literatur über POS-Tagging und Parsing vor. Abschnitt 3 ist der Vorstellung des Stuttgart-Tübingen-Tagsets und des Universal Tagsets gewidmet. In Abschnitt 4 erklären wir unseren Experimentaufbau, und in Abschnitt 5 analysieren wir die Ergebnisse der Experimente. Abschnitt 6 ordnet unsere Ergebnisse in den Stand der Forschung ein, und Abschnitt 7 schließt den Artikel.

2 Bisherige Arbeiten

In diesem Abschnitt stellen wir einen Ausschnitt der Literatur vor, die den Zusammenhang zwischen POS-Tagging und Parsing unter verschiedenen Aspekten untersucht. Während diese Übersicht keinen Anspruch auf Vollständigkeit erhebt, so sollte sie doch einen Eindruck über vorherige Arbeiten schaffen.

Eine Grundfrage, die sich die meisten Arbeiten stellen, ist, wie am besten (in Bezug auf Parsingergebnisse und Parsinggeschwindigkeit) zwischen mehreren Tags für ein einzelnes Wort disambiguiert werden kann.

Einige Arbeiten untersuchen, ähnlich wie wir, die Rolle des POS-Tagging für Parsing innerhalb eines „Pipeline"-Ansatzes, bei dem die Ausgabe eines POS-Taggers als

Eingabe für einen Parser fungiert. So diskutieren Charniak et al. (1996) die optimale Wahl eines Taggers für PCFG-Parsing. Sie kommen zu dem Hauptergebnis, dass Markov-Modell-basierte Tagger, die jedes Wort mit einem einzelnen Tag versehen (also komplett disambiguieren), am besten geeignet sind. Die Autoren zeigen, dass PCFG-Parser schlechter zwischen POS-Tags disambiguieren und außerdem einen höheren Rechenaufwand verursachen. Anders als Charniak et al. verwendet Maier (2006) keinen separaten Tagger, sondern Gold-POS-Tags in der Parsereingabe (sprich, eine perfekte Tag-Disambiguierung). Er führt PCFG-Parsingexperimente mit den zwei Baumbanken NeGra und TiGer durch und kommt zu einem vergleichbaren Ergebnis: Mit Gold-POS-Tags steigt die Ausgabequalität des Parsers und der Rechenaufwand verringert sich.

Andere Arbeiten untersuchen, ebenfalls innerhalb eines „Pipeline"-Ansatzes, Möglichkeiten zur Reduktion von Ambiguität über die Modifikation von Tagsets bzw. des Lexikons durch Tagset-Reduktion oder Wort-Clustering. Lakeland (2005) beschäftigt sich mit lexikalisiertem Parsing à la Collins (1999). Ähnlich der neueren Arbeiten z.B. von Koo et al. (2008) oder von Candito und Seddah (2010) geht er die Frage nach der für das Parsen optimalen Disambiguierung durch Clustering auf lexikalischer Ebene an. Ein Wort-Cluster wird hierbei als Äquivalenzklasse von Wörtern gesehen und übernimmt gewissermaßen die Funktion eines POS-Tags, kann aber den Trainingsdaten angepasst werden. Le Roux et al. (2012) beschäftigen sich mit Datenknappheit auf lexikalischer Ebene beim PCFG-Parsing der morphologisch reichen Sprache Spanisch. Die Autoren benutzen eine Neuimplementierung des Berkeley Parsers. Sie zeigen, dass Parsingergebnisse sowohl durch eine Vereinfachung des POS-Tagsets als auch durch Lemmatisierung verbessert werden können, da beide Vorgehensweisen die Datenknappheit reduzieren.

Wie schon erwähnt, kann ein POS-Tag als eine Äquivalenzklasse von Wörtern gesehen werden. Da im „Pipeline"-Ansatz der Parsebaum über den POS-Tags erstellt wird, ist es jedoch möglich, dass ein POS-Tagset zwar aus linguistischer Sicht optimal ist, sich jedoch in Bezug auf Parsingergebnisse nicht optimal verhält, da für den Parsebaum relevante lexikalische Information durch das POS-Tagset verdeckt wird. Während lexikalisches Clustering wie bei Koo et al. (2008) dieses Defizit dadurch überwindet, dass (semi-)automatisch „bessere" Cluster gesucht werden, kopieren andere Arbeiten lexikalische Information durch Baumbanktransformationen in den eigentlichen Baum. Dafür wird auch die in einigen Baumbanken bereits vorhandene Annotation von grammatischen Funktionen benutzt. Derartige Transformationen werden z.B. von Versley (2005) und Versley und Rehbein (2009) beschrieben. Seeker und Kuhn (2013) stellen einen Ansatz vor, der das „Pipeline"-Modell (unter Benutzung eines Dependenz-Parsers (Bohnet, 2010)) um eine zusätzliche Komponente ergänzt, die Kasus-Information als Filter für den Parser verwendet. Sie erreichen Verbesserungen für Ungarisch, Deutsch und Tschechisch und stellen dabei fest, dass es verschiedene Arten von morphologischer Komplexität gibt, die beim Parsing unterschiedlich behandelt werden müssen.

Der Zusammenhang von POS-Tagging und Parsing wurde nicht nur im Rahmen des baumbankbasierten Parsing untersucht, sondern auch im Rahmen des Parsing mit einer handgeschriebenen Grammatik und einer Disambiguierungskomponente.

Dalrymple (2006) untersucht die Rolle des POS-Tagging für das Parsing auf Basis des Systems von Riezler et al. (2002), das aus einer englischen *Lexical Functional Grammar*, einem Constraint-basierten Parser und einer Disambiguierungskomponente besteht. Die Autorin benutzt keinen separaten POS-Tagger. Sie bildet Äquivalenzklassen von Parserausgaben basierend auf deren Tag-Sequenzen. Aus der Zahl der gefundenen Äquivalenzklassen für einen einzelnen Satz schließt sie darauf, inwieweit perfektes Tagging für die Disambiguierung des Satzes helfen würde: Eine hohe Anzahl von Klassen lässt darauf schließen, dass die syntaktischen Analysen eines Satzes durch die POS-Tags differenzierbar sind, eine niedrige Anzahl deutet in die gegenteilige Richtung. Sie kommt zu dem Ergebnis, dass mit einem perfekten Tagger eine fünfzigprozentige Reduzierung der Parseambiguität zu erreichen wäre.

Watson (2006) stellt unter Benutzung des RASP-Systems (Briscoe und Carroll, 2002) weitergehende Untersuchungen zum Zusammenhang zwischen verschiedenen Modellen der Tagauswahl (einzelne/mehrere Tags pro Wort, Tagauswahl durch Parser oder Tagger) und Parsing an. Sie zeigt einerseits, ebenso wie andere Arbeiten, dass POS-Tagger eine bessere Tagauswahl treffen als Parser, und andererseits, dass es einen Trade-Off zwischen der Qualität von Parsingergebnissen und dem Zulassen multipler Tags pro Wort gibt. Prins und van Noord (2003) betrachten eine verwandte Frage für einen HPSG-Parser. Sie verwenden einen auf Parserausgaben trainierten Markov-Modell-POS-Tagger, um Parsereingaben vorzutaggen. Dies wirkt sich günstig sowohl auf Parsingergebnisse als auch auf die Parsingzeit aus. Curran et al. (2006) beschäftigen sich mit der Rolle von POS-Tagging für CCG- und TAG-Parsing. Sie berichten, dass eine zu frühe Disambiguierung von POS-Tags sich dann schlecht auf Parsingergebnisse auswirkt, wenn einzelne Tags sehr informativ sind (vgl. Supertagging gegenüber „normalem“ POS-Tagging (Bangalore und Joshi, 1999)). Die Arbeit von Yoshida et al. (2007) geht in dieselbe Richtung. Die Autoren verwenden einen HPSG-Parser mit vorgeschaltetem POS-Tagger und zeigen, dass das Zulassen von mehreren Tags pro Wort, d.h. das teilweise Übertragen der Tag-Disambiguierung an den Parser, sich unter bestimmten Bedingungen günstig auf Parsingergebnisse auswirkt.

Eine Reihe von Arbeiten entwirft Modelle für gleichzeitiges POS-Tagging, bzw. morphologische Segmentierung, und Parsing. Besonders interessant ist hier die Arbeit von Chen und Kit (2011). Sie gehen in gewisser Weise ebenfalls die Frage der Disambiguierung auf lexikalischer Ebene an. Basierend auf Arbeiten von Ratnaparkhi (1996) und Toutanova und Manning (2000) gehen die Autoren davon aus, dass lokale Merkmale für die Qualität von POS-Tagging entscheidend sind. Nichtsdestotrotz wird dies nicht berücksichtigt, wie sie bemerken, wenn auf ein „Pipeline“-Modell verzichtet wird, d.h. wenn dem Parser auch die Aufgabe des Tagging zufällt. Auf dieser Basis stellen sie ein erfolgreiches faktorisiertes Modell für das PCFG-Parsing vor, das das Parsing in ein diskriminatives lexikalisches Modell (mit lokalen Merkmalen) und das eigentliche Parsingmodell trennt. Die Modelle werden mittels eines *product-of-experts* (Hinton, 1999) kombiniert.

Kombinierte Modelle für gleichzeitiges POS-Tagging und Parsing lassen sich besonders in der Dependenzparsing-Literatur finden; hier zeigt sich vor allem eine Konzentration

auf Sprachen, die noch eine zusätzliche Segmentierung auf der Wortebene erfordern, so wie Chinesisch (Hatori et al., 2011) oder Hebräisch (Goldberg und Tsarfaty, 2008). Ein neuerer Ansatz von Bohnet und Nivre (2012) wurde auch auf dem Deutschen evaluiert. Ergebnisse zum POS-Tagging und Parsing des Deutschen mittels einer Constraint-Grammatik finden sich in Daum et al. (2003) sowie Foth et al. (2005). Da diese Arbeiten den Gegenstand unserer Arbeit jedoch nur am Rand berühren, verzichten wir auf einen weiteren Überblick.

3 Die Tagset-Varianten

In diesem Abschnitt werden die verschieden Tagset-Varianten beschrieben, die wir in unseren Experimenten verwenden. Wir beginnen mit dem Stuttgart-Tübingen-Tagset (STTS) (Schiller et al., 1995; Thielen und Schiller, 1994), das sich zum Standard POS-Tagset für das Deutsche entwickelt hat. Des Weiteren beschreiben wir die morphologische Erweiterung des STTS. Da die beiden hier verwendeten Baumbanken, TiGer und TüBa-D/Z (siehe Abschnitt 4.1), unterschiedliche Morphologie-Annotationen aufweisen, stellen wir beide Versionen der Morphologie vor. Die kleinste POS-Tagset-Variante ist das Universal Tagset (UTS) (Petrov et al., 2012).

Das UTS besteht aus 12 grundlegenden Tags, die in Tabelle 1 aufgeführt sind. Es wurde entwickelt, um als gemeinsames Tagset für eine große Anzahl von Sprachen verwendet zu werden, z.B. um sprachübergreifende POS-Tagger zu entwickeln, oder um Sprachen in Bezug auf das POS-Tagging zu vergleichen. Bei diesem Tagset fällt auf, dass nur sehr grundlegende Wortarten vertreten sind, es wird z.B. keine Unterscheidung zwischen verschieden Arten von Pronomen gemacht, und koordinierende und subordinierende Konjunktionen werden gemeinsam unter CONJ gruppiert. Dieses Tagset sollte eine hohe Qualität beim POS-Tagging garantieren, weil nur wenige, grobe Unterscheidungen getroffen werden. Jedoch stellt sich die Frage, inwieweit diese grobe Granularität genügend Information für eine syntaktische Analyse liefert.

Das STTS ist ein Tagset, das hauptsächlich basierend auf distributionellen Regularitäten des Deutschen entwickelt wurde. Es umfasst 54 Tags und modelliert damit wesentlich feinere Unterschiede als das UTS. Die STTS-Tags sind in den Tabellen 9 und 10 im Anhang aufgeführt.

Es fällt auf, dass das STTS nicht nur feinere Unterscheidungen von Wortarten macht, es modelliert auch die Finitheit bei Verben. Dies ist eine wichtige Unterscheidung für die syntaktische Analyse, aber es ist auch in bestimmten Fällen eine schwierige Aufgabe für den POS-Tagger, der nur einen sehr begrenzten lokalen Kontext verwendet: Deutsche Verbformen sind ambig in Bezug auf den Infinitiv und die Präsens-Plural-Form. Der Satz in (1) aus der TüBa-D/Z zeigt ein Beispiel hierfür: Das Verb `wirken` kann anhand des Kontexts von `wie Luken` nicht disambiguiert werden.

NOUN	Nomen
VERB	Verb
ADJ	Adjektiv
ADV	Adverb
PRON	Pronomen
DET	Determiner, Artikel
ADP	Preposition, Postposition
NUM	Numeral
CONJ	Konjunktion
PRT	Partikel
.	Interpunktion
X	alles andere

Tabelle 1: Die 12 Tags des Universal Tagsets.

ambig:	*
Genus:	maskulin (Masc), feminin (Fem), neutral (Neut)
Gradation:	Positiv (Pos), Komparativ (Comp), Superlativ (Sup)
Kasus:	Nominativ (Nom), Genitiv (Gen), Dativ (Dat), Akkusativ (Akk)
Modus:	Indikativ (Ind), Konjunktiv (Subj), Imperativ (Imp)
Numerus:	singular (Sg), plural (Pl)
Person:	1, 2, 3
Tempus:	Präsens (Pres), Präteritum (Past)

Tabelle 2: Die morphologischen Kategorien aus TiGer.

(1) es hat Passagen mit kleineren Fenstern , die aber nicht
PPER VAFIN NN APPR ADJA NN $, PRELS ADV PTKNEG
wie Luken wirken .
KOKOM NN VVFIN $.

Das STTS kann auch um eine morphologische Komponente erweitert werden. Dies ist bei den beiden Baumbanken TiGer und TüBa-D/Z geschehen, allerdings wurden unterschiedliche Entscheidungen getroffen. In der TiGer-Baumbank wird eine Menge von 585 verschiedenen morphologischen Merkmalskombinationen verwendet, die sich aus den in Tabelle 2 aufgelisteten Elementen zusammensetzen. Der Satz in (2) zeigt ein Beispiel der Kombination von STTS-Tags und -Morphologie, getrennt durch ein **%** Zeichen. Das Merkmal – bedeutet, dass keine morphologischen Merkmale vorliegen.

ambig:	*
Genus:	maskulin (m), feminin (f), neutral (n)
Kasus:	Nominativ (n), Genitiv (g), Dativ (d), Akkusativ (a)
Numerus:	singular (s), plural (p)
Person:	1, 2, 3
Tempus:	Präsens (s), Präteritum (t)
Modus:	Indikativ (i), Konjunktiv (k)

Tabelle 3: Die morphologischen Kategorien aus der TüBa-D/Z.

(2) Konzernchefs lehnen den
NN%Nom.Pl.Masc VVFIN%3.Pl.Pres.Ind ART%Acc.Sg.Masc
Milliardär als US-Präsidenten ab /
NN%Acc.Sg.Masc APPR%– NN%Acc.Sg.Masc PTKVZ%– $(%–

Aus der Menge der möglichen Kombinationen morphologischer Merkmale sind 271 verschiedene Kombinationen in der TiGer-Baumbank belegt. Daraus ergeben sich insgesamt 783 mögliche Kombinationen von STTS-Tags und Morphologie. Von diesen sind 761 im Trainingsset vorhanden. Wegen der hohen Anzahl von möglichen Kombinationen ist jedoch zu erwarten, dass Kombinationen, die im Development- oder Testset vorhanden sind, nicht im Trainingsset vorkommen, d.h. dass mit Datenknappheit zu rechnen ist. Deswegen haben wir ermittelt, wie viele der Kombinationen, die in den Development- und Testdaten erscheinen, auch in Trainingsset vorhanden sind. Es stellt sich heraus, dass 25% bzw. 30% nicht in den Trainingsdaten vorkommen. Man beachte dabei, dass die Anzahl der Kombinationen in den Test- und Developmentsets wesentlich geringer sind als im Trainingsset.

In der TüBa-D/Z gibt es insgesamt 132 mögliche morphologische Merkmalskombinationen. Sie setzen sich aus den in Tabelle 3 aufgelisteten Elementen zusammen. Der Satz in (3) zeigt ein Beispiel der Kombination von STTS-Tags und -Morphologie.

(3) Aber Bremerhavens AfB fordert jetzt
KON%– NE%gsn NE%nsf VVFIN%3sis ADV%–
Untersuchungsausschuß
NN%asm

Aus der Menge dieser möglichen Kombinationen morphologischer Merkmale sind 105 verschiedene Kombinationen in der TüBa-D/Z Baumbank belegt. Daraus ergeben sich insgesamt 524 verschiedene Kombinationen von STTS-Tags und Morphologie. Von diesen erscheinen 513 im Trainingsset; von den Kombinationen, die in unseren Development- und Testdaten erscheinen, sind 16% bzw. 18% in den Trainingsdaten nicht vorhanden.

Die Tatsache, dass die Kombinations-POS-Tagsets für beide Baumbanken mehrere hundert verschiedene Labels aufweisen, in Kombination mit der mangelnden Abdeckung

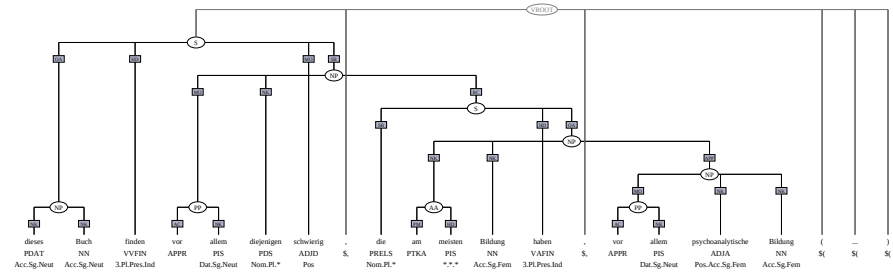


Abbildung 1: Ein Satz mit syntaktischer Annotation aus der TiGer-Baumbank.

des Trainingssets, lässt uns vermuten, dass die POS-Tagging-Ergebnisse für diese Variante weniger gut sein werden als für das Standard-STTS. Eine weitere Hypothese ist, dass das morphologische Tagset der TüBa-D/Z aufgrund seiner geringeren Größe besser zum Tagging geeignet ist als das TiGer-Tagset. Es ist jedoch offen, ob die morphologische Information beim Parsing gewinnbringend eingesetzt werden kann.

4 Aufbau der Experimente

4.1 Daten

Als Datensätze verwenden wir die zwei größten deutschen Baumbanken: TiGer (Brants et al., 2002) und die Tübinger Baumbank des Deutschen/Zeitungskorpus (TüBa-D/Z) (Telljohann et al., 2012). Beide Baumbanken basieren auf Zeitungstexten, die TiGer-Baumbank auf der Frankfurter Rundschau und die TüBa-D/Z auf der tageszeitung (taz). Beide Baumbanken verwenden das STTS mit minimalen Unterschieden. Aufbauend auf der POS-Ebene haben beide Baumbanken eine syntaktische Annotation bestehend aus einer Konstituentenstruktur, erweitert durch grammatische Funktionen. Die Baumbanken unterscheiden sich deutlich auf der Ebene der syntaktischen Annotationen, zum einen durch die unterschiedlichen Knoten- und Kantenlabels, zum anderen durch die Entscheidung in TiGer, kreuzende Kanten zu annotieren bzw. durch die Verwendung von topologischen Feldern (Höhle, 1986) in der TüBa-D/Z. Die Abbildungen 1 und 2 zeigen Beispielsätze aus den Baumbanken.

Für unsere Experimente verwenden wir Version 2.0 der TiGer-Baumbank, mit einem Umfang von 50 474 Sätzen, und Version 8.0 der TüBa-D/Z, mit einem Umfang von 75 408 Sätzen. Um unerwünschte Größeneffekte auszuschließen, verwenden wir die folgenden Größen für beide Baumbanken: die ersten 40 475 Sätze für das Training, die jeweils nächsten 5 000 Sätze für das Development- und das Testset. Dies entspricht der Aufteilung der TiGer-Baumbank durch Farkas und Schmid (2012).

Für beide Baumbanken müssen alle Interpunktionszeichen und alle anderen direkt an der Wurzel angehängte Elemente wie z.B. nicht angehängte Appositionen, in die

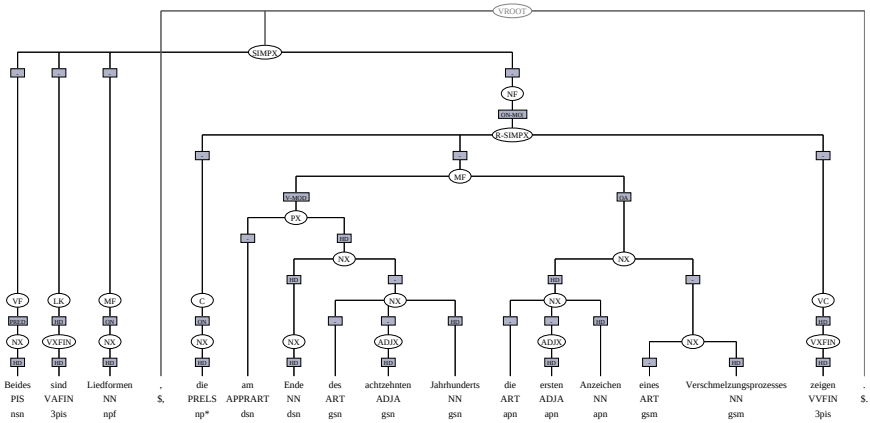


Abbildung 2: Ein Satz mit syntaktischer Annotation aus der TüBa-D/Z.

Konstituentenstruktur eingebaut werden. Hier folgen wir dem Ansatz von Maier et al. (2012).

Beide Baumbanken müssen außerdem in ein Format konvertiert werden, das von dem verwendeten Parser akzeptiert wird. Für die TiGer-Baumbank bedeutet dies, dass alle kreuzenden Kanten aufgelöst werden müssen, damit kontextfreie Regeln extrahiert werden können. Diese Auflösung wird in drei Schritten bewerkstelligt. In einem ersten Schritt wird für jede Phrase ein Kopf markiert. Dafür wird, soweit möglich, die vorhandene Kantenannotation („HD“) benutzt. Sollte ein Kopf so nicht zu bestimmen sein, kommen Heuristiken zum Einsatz. In einem zweiten Schritt wird die in Boyd (2007) beschriebene Transformation durchgeführt, d.h. für jeden kontinuierlichen Block einer diskontinuierlichen Konstituente wird ein separates Nichtterminal eingeführt; im gleichen Schritt wird aus der Menge bestehend aus dem ursprünglichen und den neu eingeführten Knoten derjenige Knoten markiert, unter dem letztendlich die als Kopf markierte Tochter verbleibt. In einem dritten Schritt erfolgt nun die eigentliche Auflösung der kreuzenden Kanten: Alle gesplitteten, aber nicht markierten Knoten werden entfernt und deren Töchter an den Mutterknoten des zu entfernenden Knotens gehängt. Für den Satz in Abbildung 1 bedeutet das, dass der Relativsatz S mit der grammatischen Funktion RC aus der NP `vor allem diejenigen` hochgereicht wird an die Mutter der NP, S.

4.2 POS-Tagger

Wir verwenden vier verschiedene POS-Tagger, die ein möglichst breites Spektrum an verschiedenen Ansätzen zum POS-Tagging abdecken, namentlich einen *Trigramm/Markov-*

Modell-Tagger, einen *Maximum-Entropy*-Tagger, einen *Conditional-Random-Field*-Tagger und einen *Support-Vector-Machine*-Tagger. Alle POS-Tagger werden mit den Standardeinstellungen verwendet.

TnT (Brants, 2000, 1998), kurz für *Trigrams and Tags*, ist ein Markov-Modell-POS-Tagger, der eine Interpolation zwischen Uni-, Bi- und Trigrammen als Wahrscheinlichkeitsmodell verwendet. Dieser POS-Tagger besitzt ein Modul zur Behandlung von unbekannten Wörtern, das einen *trie* aus Suffixen verwendet, der aus Hapax Legomena aus dem Trainingskorpus extrahiert wird. TnT ist auch 15 Jahre nach seiner Entwicklung das beste verfügbare Modul für diesen Zweck.

Der **Stanford log-linear POS Tagger** (Toutanova et al., 2003; Toutanova und Manning, 2000) basiert auf einem *Maximum-Entropy*-Modell. Dies bedeutet, dass dieser POS-Tagger, ebenso wie die beiden nachfolgend beschriebenen Tagger, im Hintergrund ein diskriminatives maschinelles Lernverfahren statt eines Markov-Modells verwendet. Dies hat den Vorteil, dass er mit Merkmalen arbeiten kann, die weit über den linken Kontext von 1-2 Wörtern hinausgehen. Der Nachteil des Verfahrens ist, dass keine globale Optimierung stattfindet. Wir verwenden den POS-Tagger mit dem bidirektionalen Modell basierend auf einem Kontext von 5 Wörtern.

Der dritte POS-Tagger basiert auf *Conditional Random Fields*. Diese sind zum Annotieren von Sequenzen besonders geeignet (Lafferty et al., 2001). Wir verwenden die Anwendung für *sequence tagging* in **MALLET** (McCallum, 2002).

Der letzte von uns verwendete Tagger, **SVMTool** (Giménez und Màrquez, 2004), basiert auf *Support Vector Machines* (SVMs), genauer gesagt auf der SVM-Implementierung SVM^{light} (Joachims, 1999). SVMs haben sich als extrem gut geeignet für Problemstellungen in der Computerlinguistik erwiesen. Wir verwenden das Modell, das einen Satz von links nach rechts in einem Durchgang analysiert, mit dem Standard-Merkmalsset, das Trigramme aus Wörtern und POS-Tags ebenso verwendet wie Wortlänge, Präfix- und Suffixinformation.

4.3 Parser

Als Parser verwenden wir den **Berkeley Parser** (Petrov et al., 2006), einen Konstituentenparser, der seine Grammatik dadurch verfeinert, dass er die syntaktischen Labels in Unterklassen aufteilt, bzw. Labels verschmilzt, die in ähnlichen Umgebungen vorkommen. Die im Berkeley Parser implementierte Technik für das automatische Aufteilen und Verschmelzen von Labels ist der derzeitige Stand der Technik für das Parsing des Deutschen, basierend auf einem individuellen Parser ohne Reranking. Wir beschränken uns hier auf grundlegendes Konstituentenparsing, d.h. wir verwenden keine grammatischen Funktionen beim Parsing. Der Parser wurde in 6 Iterationen trainiert.

4.4 Experimente

In diesem Beitrag betrachten wir 4 Fragestellungen:

1. STTS-Variationen: In diesem Satz von Experimenten untersuchen wir, wie sich die Granularität von STTS auf die Qualität der Ausgaben der verschiedenen POS-Tagger auswirkt. Wir betrachten das Universal Tagset (UTS), das Standard-STTS und das durch Morphologie erweiterte STTS.

2. STTS-Variationen und Parsing: Hier untersuchen wir, wie sich die unterschiedliche Granularität von STTS auf die Parsingqualität auswirkt. Um von der unterschiedlichen Qualität der POS-Tagger zu abstrahieren, verwenden wir für dieses Experiment erst Gold-POS-Tags und wiederholen die Experimente dann mit automatisch getaggten Texten.

3. Morphologische Variationen: In diesem Experiment untersuchen wir verschiedene Untermengen von morphologischen Merkmalen. Hierfür verwenden wir nur TnT.

4. Morphologische Variationen und Parsing: Hier untersuchen wir den Einfluss der morphologischen Merkmalsets auf die Qualität des Parsers. Hierzu verwenden wir erst Gold-POS-Tags, dann von TnT erzeugte Tags.

Der erste Satz von Experimenten soll zum einen untersuchen, in welchem Ausmaß sich die Art des POS-Taggers auf die Qualität der Ergebnisse auswirkt. Zum anderen wollen wir feststellen, wie sich die unterschiedliche Granularität der STTS-Varianten auf die Tagger-Qualität auswirkt. Unsere Hypothese ist: Je mehr Informationen im Tagset vorhanden sind, desto schwieriger ist die Disambiguierung von verschiedenen Tags. Es ist außerdem anzunehmen, dass unterschiedliche POS-Tagger unterschiedlich gut mit den verschiedenen Tagset-Größen zurechtkommen.

Der zweite Satz von Experimenten dient dazu, den Einfluss der im Tagset vorhandenen Informationen auf die Parsingergebnisse zu bestimmen. Die Frage hier ist, ob es ausreicht, dem Parser wenige Informationen zu geben, oder ob er von der hohen Informationsdichte im morphologisch angereicherten Tagset ebenfalls profitieren kann. Hier ist anzumerken, dass der Berkeley Parser zu feine Unterscheidungen in den Tags in seinem Lernverfahren zusammenfassen kann.

Der dritte Satz von Experimenten soll bestimmen, ob sich eine bestimmte Unterkategorie von morphologischen Merkmalen zuverlässig mit einem POS-Tagger ermitteln lässt. Die ausgewählten Unterkategorien gründen auf der Intuition der Autoren, welche Art von Merkmalen potentiell hilfreich für das Parsing sein könnte. Die erste Unterkategorie (Kongruenz) beschränkt sich auf die nominalen Kongruenzmerkmale, Genus, Numerus und Kasus. Die zweite Unterkategorie besteht nur aus Kasusmerkmalen. Hier ist nicht zu erwarten, dass ein POS-Tagger, mit seinem sehr eingeschränkten lokalen Kontext und ohne morphologische Analyse, diese Unterscheidung immer erfolgreich treffen kann. Vor allem wegen des Synkretismus von Nominativ/Akkusativ bzw. Genitiv/Dativ im Deutschen kann eine solche Unterscheidung nur auf einer syntaktischen Analyse basierend getroffen werden. Die dritte Unterkategorie besteht aus den Numerusmerkmalen, die vierte kombiniert Numerus mit Person. In der letzten Unterkategorie werden alle verbalen Merkmale verwendet.

Der vierte Satz von Experimenten untersucht, ob sich die Trends innerhalb des POS-Tagging mit morphologischen Unterkategorien auch auf das Parsing übertragen lassen. Die Frage ist hier, anders ausgedrückt, ob es wichtiger ist, dem Parser wichtige, aber u.U. unzuverlässige Information zur Verfügung zu stellen, oder ob es besser ist, ein grobkörnigeres Tagset zu verwenden, das aber zuverlässig automatisch annotiert werden kann.

Zu beachten ist, dass unsere Untersuchung task-basiert ist, d.h. wir untersuchen, wie sich die einzelnen Tagsets in Bezug zur Aufgabe des Taggings verhalten. Dies stellt keinen direkten Vergleich zwischen einzelnen Tagsets dar.

4.5 Evaluierung

Die Evaluierung auf der POS-Ebene berechnet die Korrektheit des POS-Taggers in Bezug auf *accuracy*. Wir verwenden das Evaluierungsskript von TnT, da dieses uns erlaubt, auch die Korrektheit von bekannten und unbekannten Wörtern zu berechnen.

Für die Evaluierung des Parsers geben wir *precision*, *recall*, und *F-Score* an. Wir verwenden die Implementierung `evalb-lcfrs`.[1] Diese Implementierung verhält sich auf kontinuierlichen Konstituenten, wie hier vorhanden, ebenso wie die Standard-Software Evalb (ohne Parameterdatei).[2] Bei der Parserevaluierung werden die POS-Tags nicht berücksichtigt.

Da wir keine Optimierung der POS-Tagger und des Parsers vornehmen und da sich die Development- und Testsets sehr unterscheiden, geben wir die Ergebnisse jeweils für beide Datensätze an.

Alle Experimente werden auf einem 3,16 GHz Intel Xeon mit jeweils 32 GB maximalem Speicher pro Experiment durchgeführt.

5 Ergebnisse

5.1 Drei Varianten von STTS

Hier untersuchen wir, welchen Effekt die unterschiedliche Granularität des POS-Tagsets auf die Qualität von verschiedenen POS-Taggern hat. Die Ergebnisse sind in Tabelle 4 zusammengefasst.

Das erste Resultat dieser Experimente ist, dass sich die beiden Baumbanken kaum in den Ergebnissen unterscheiden. Daraus können wir schließen, dass beide Textquellen gleich schwierig für die POS-Tagger sind. Es ist jedoch auffällig, dass das Developmentset und das Testset jeweils unterschiedliche Ergebnisse aufweisen. Es gibt also innerhalb der jeweiligen Baumbanken deutliche Unterschiede.

Wenn man die verschiedenen POS-Tagger vergleicht, dann fällt auf, dass SVMTool die besten Ergebnisse für das minimalistische Universal Tagset (UTS) liefert.[3]Für das

[1] Siehe `https://github.com/wmaier/evalb-lcfrs`.

[2] Siehe `http://nlp.cs.nyu.edu/evalb/`.

[3] Hier muss allerdings darauf hingewiesen werden, dass ein Vergleich über die verschiedenen Tagset-Varianten hinweg nur mit Vorsicht zu genießen ist, da die unterschiedlichen Tagset-Größen

	TiGer		TüBa-D/Z	
POS-Tagger	Development	Test	Development	Test
	UTS			
MALLET	91,67%	90,22%	93,54%	94,01%
Stanford	97,88%	96,83%	97,11%	97,26%
SVMTool	**98,54%**	**98,01%**	**98,09%**	**98,28%**
TnT	97,94%	97,48%	97,72%	97,92%
	STTS			
MALLET	92,45%	90,29%	88,81%	89,12%
Stanford	96,26%	95,15%	95,63%	95,79%
SVMTool	97,06%	96,22%	96,46%	96,69%
TnT	**97,15%**	**96,29%**	**96,92%**	**97,00%**
	STTSmorph			
MALLET	–	–	–	–
Stanford	–	–	–	–
SVMTool	82,47%	79,53%	80,33%	81,31%
TnT	**85,77%**	**82,77%**	**84,67%**	**85,45%**

Tabelle 4: Die Ergebnisse für verschiedene POS-Tagger mit verschiedenen Tagset-Varianten: das Universal Tagset (UTS), das STTS und das STTS mit Morphologie (STTSmorph).

Standard-STTS und für die Variante mit morphologischer Information erweist sich TnT, der bei weitem älteste POS-Tagger, als der zuverlässigste. Bei der morphologischen Variante liegt der Unterschied zwischen SVMTool und TnT bei mehr als 3% für TiGer und bei mehr als 4% für die TüBa-D/Z. Eine mögliche Erklärung für die guten Ergebnisse von TnT auf der morphologischen Variante ist, dass Markov-Modelle weniger Trainingsdaten benötigen als die anderen Tagger, die auf diskriminativen Lernverfahren basieren. Dies ist ein wichtiger Vorteil in Situationen, in denen viele POS-Tags nur selten in den Trainingsdaten vorkommen.

Bei einer Betrachtung der Ergebnisse des Stanford Taggers und von MALLET fällt zunächst auf, dass wir keine Ergebnisse für die morphologische Variante haben. Das liegt daran, dass für das Training von MALLET zwei Wochen nicht ausreichten. Zu diesem Zeitpunkt haben wir die Experimente abgebrochen. Der Stanford Tagger brach das Tagging mit einer Fehlermeldung ab, die uns vermuten lässt, dass der Tagger die hohe Anzahl von Merkmalen nicht verarbeiten kann, die bei dieser Variante anfällt. Bei den anderen Varianten schneidet der Stanford Tagger geringfügig schlechter ab als der jeweils zweitplatzierte POS-Tagger. Im Gegensatz dazu schneidet MALLET bedeutend schlechter ab. Der Unterschied zwischen MALLET und dem bestplatzierten POS-Tagger beträgt 7-8% für TiGer und 4-5% für die TüBa-D/Z. Diese Differenz

auch einen Einfluss auf die Schwierigkeit der Aufgabe haben können. Unser Interesse in allen Experimenten beschränkt sich darauf, welcher Prozentsatz von Wörtern korrekt annotiert wurde.

	TiGer				TüBa-D/Z			
	Dev.		Test		Dev.		Test	
POS-Tag.	bek.	unbek.	bek.	unbek.	bek.	unbek.	bek.	unbek.
	UTS							
MALLET	92,83	77,66	92,09	73,30	95,47	74,84	95,75	75,53
Stanford	**99,05**	91,85	**98,78**	87,70	**98,94**	79,30	**98,92**	79,69
SVMTool	98,81	95,26	98,41	94,45	98,63	92,89	98,66	94,27
TnT	98,06	**96,50**	97,67	**95,74**	98,07	**94,28**	98,25	**95,25**
	STTS							
MALLET	94,66	65,70	93,40	62,25	91,44	63,27	91,67	62,14
Stanford	**98,16**	73,56	**97,75**	71,60	**97,96**	73,04	**97,97**	72,64
SVMTool	97,86	87,41	97,26	86,82	97,50	86,47	97,60	87,05
TnT	97,80	**89,25**	97,21	**87,95**	97,65	**89,78**	97,72	**89,33**
	STTSmorph							
SVMTool	84,67	55,89	82,40	53,58	82,87	55,81	83,61	57,01
TnT	**87,62**	**63,41**	**85,55**	**57,65**	**86,91**	**62,95**	**87,61**	**62,55**

Tabelle 5: Die Ergebnisse für verschiedene POS-Tagger für bekannte und unbekannte Wörter.

liegt zum Teil sicher daran, dass MALLET kein designierter POS-Tagger sondern ein generelles Sequenzlernverfahren ist, d.h. es existieren z.B. keine Strategien, wie unbekannte Wörter zu behandeln sind.

Bei einem Vergleich der Tagsetvarianten bestätigt sich unsere Hypothese: Je mehr Informationen vorhanden sind, desto schwieriger ist die Disambiguierung für die POS-Tagger. Das UTS kann am zuverlässigsten annotiert werden, hier liegen die Ergebnisse bei über 97% für TnT. Das Standard-STTS verzeichnet nur minimale Einbußen. Dies bedeutet, dass ein größeres Tagset nicht unbedingt schwieriger zu taggen ist, wenn die Tags die tatsächliche Distribution modellieren, die durch die in den Taggern vorhandene Information abgebildet ist.

Die morphologische Variante von STTS dagegen resultiert in einem wesentlich größeren Tagset und daher auch in einem wesentlich schwierigeren Problem. Dies zeigt sich in den Ergebnissen, die durchgehend 12-14% schlechter sind als dieselben Ergebnisse für das Standard-STTS. Unsere Hypothese, dass das morphologische Tagset der TüBa-D/Z aufgrund seiner geringeren Größe zu besseren Ergebnissen führt, ist nicht bestätigt: die Unterschiede zwischen dem Developmentset und dem Testset von TiGer sind größer als die Unterschiede zwischen den Baumbanken.

In einer weiteren Auswertung unterscheiden wir zwischen bekannten und unbekannten Wörtern im Development- und Testset. Bekannte Wörter definieren wir als diejenigen Wörter, die im Trainingsset erscheinen, unbekannte Wörter sind solche, die im Trainingsset nicht vorhanden sind. In TiGer sind 7,64% der Wörter im Developmentset unbekannt, im Testset 9,96%. In TüBa-D/Z sind 9,36% der Wörter im Developmentset

unbekannt, im Testset 8,64%. Die Auswertung für bekannte und unbekannte Wörter findet sich in Tabelle 5. Diese Ergebnisse zeigen, dass der Stanford Tagger für das UTS und das STTS für bekannte Wörter die besten Ergebnisse erreicht, während TnT durchgehend die besten Ergebnisse für unbekannte Wörter erreicht. Für das morphologisch erweiterte STTS erreicht TnT außerdem die besten Ergebnisse auch für bekannte Wörter. Wir erinnern daran, dass diese Variante mit MALLET und dem Stanford Tagger nicht getaggt werden konnte.

Die Ergebnisse zeigen außerdem, dass MALLET, wie erwartet, bei unbekannten Wörtern deutlich schlechter abschneidet als die anderen POS-Tagger, was darauf zurückzuführen ist, dass MALLET keine separate Strategie zur Verarbeitung unbekannter Wörter besitzt. Darüber hinaus sind jedoch MALLETs Ergebnisse für bekannte Wörter auch deutlich schlechter als die der anderen Tagger, was bedeutet, dass das Sequenzmodell für das POS-Tagging nicht optimal ist, zumindest nicht ohne Anpassung an die vorliegende Aufgabe.

5.2 Parsing mit unterschiedlicher Tagset-Granularität

In diesem Satz von Experimenten untersuchen wir, wie sich die verschiedenen Ergebnisse beim POS-Tagging mit unterschiedlicher Tagset-Granularität auf das Parsing auswirken. Für diese Experimente verwenden wir die syntaktischen Annotationen der beiden Baumbanken, allerdings ohne grammatische Funktionen. Wir extrahieren dieselben Trainings-, Development-, und Testsets, die für die vorhergehenden Experimente verwendet wurden. Wir trainieren den Parser mit Gold-POS-Tags. Um die Leistung des Tagging-Mechanismus des Berkeley Parsers zu untersuchen, führen wir ein Experiment durch, in dem wir den Parser POS-Tags erzeugen lassen. In allen anderen Experimenten geben wir Wort/POS-Tag-Paare in die Parsereingabe ein, d.h. der Parser wird mit einer Option gestartet, die ihn POS-Tags erwarten lässt. In unserem Fall sind dies zum einen die Gold-POS-Tags, und zum anderen die von SVMTool und von TnT erzeugten Tags. Es ist anzumerken, dass der Berkeley Parser diese vorgegebenen POS-Tags ändert, falls er, basierend auf den Eingabe-Tags, keine syntaktische Analyse findet.

Die Ergebnisse dieser Experimente sind in Tabelle 6 aufgelistet. Ein erster Blick auf die Ergebnisse zeigt, dass es beträchtliche Unterschiede zwischen der TiGer- und der TüBa-D/Z-Baumbank gibt: Bei der TiGer-Baumbank bewegen sich die F-Scores zwischen 78,00 und 87,06 mit Gold-POS-Tags; bei der TüBa-D/Z liegen die F-Scores zwischen 91,91 und 94,57, ebenfalls basierend auf Gold-Tags. Dies steht im Kontrast zu den POS-Tagging-Ergebnissen in Tabelle 4, die zeigen, dass die Unterschiede zwischen den beiden Baumbanken in Bezug auf das POS-Tagging wesentlich geringer sind. Dies ist ein bekanntes Phänomen (Kübler et al., 2006; Rehbein und van Genabith, 2007; Kübler et al., 2008), das sich durch die Unterschiede in den syntaktischen Annotationen der beiden Baumbanken erklären lässt. Es bleibt jedoch ungeklärt, ob die Unterschiede zum Großteil dadurch entstehen, dass die Evaluierungsmetrik Annotationen mit einer großen Anzahl von Knoten, wie in der TüBa-D/Z, bevorzugt, oder ob die hierarchischeren

Parser-Eingabe	TiGer Dev. prec.	TiGer Dev. rec.	TiGer Dev. F	TiGer Test prec.	TiGer Test rec.	TiGer Test F	TüBa-D/Z Dev. prec.	TüBa-D/Z Dev. rec.	TüBa-D/Z Dev. F	TüBa-D/Z Test prec.	TüBa-D/Z Test rec.	TüBa-D/Z Test F
	UTS											
gold	87,43	85,85	86,63	83,62	81,42	82,51	92,24	91,59	91,91	92,63	91,92	92,27
SVM	85,74	84,47	85,10	81,06	79,26	80,15	89,13	88,86	89,00	89,69	89,42	89,55
TnT	84,48	83,39	83,93	79,50	78,05	78,77	88,55	88,11	88,32	89,19	88,69	88,94
Berkeley	82,99	81,83	82,41	78,87	77,09	77,97	90,51	89,72	90,11	91,08	90,25	90,67
	STTS											
gold	87,57	86,55	87,06	83,30	82,01	82,65	94,28	94,00	94,14	94,77	94,38	94,57
SVM	83,56	83,54	83,55	77,81	77,77	77,79	88,63	89,68	89,15	89,53	90,41	89,97
TnT	84,09	83,83	83,96	78,69	78,43	78,56	90,26	90,53	90,40	90,66	90,94	90,80
Berkeley	86,48	85,47	**85,97**	81,34	79,94	**80,64**	92,37	91,92	**92,15**	92,94	92,52	**92,73**
	STTSmorph											
gold	82,47	83,14	82,80	77,64	78,37	78,00	93,21	93,44	93,33	93,87	93,87	93,87
SVM	72,02	77,45	74,63	65,62	71,32	68,35	83,36	86,61	84,95	84,55	87,51	86,00
TnT	75,90	79,67	77,74	69,89	73,61	71,70	85,25	87,61	86,41	86,32	88,39	87,34
Berkeley	80,41	79,97	80,19	75,13	74,57	74,85	91,16	91,05	91,11	91,78	91,60	91,69

Tabelle 6: Die Ergebnisse des Berkeley Parsers auf verschiedenen Tagset-Varianten.

Annotationen in der TüBa-D/Z eine größere Generalisierung der Regeln und damit ein zuverlässigeres Parsing ermöglichen.

Ein Vergleich der Ergebnisse zwischen den Tags von SVMTool und TnT zeigt, dass die Tendenzen in beiden Baumbanken erhalten bleiben: Während SVMTool bei der Verwendung des Universal Tagsets die besten Ergebnisse erreicht (TiGer F-Scores: 85,10 und 80,15; TüBa-D/Z F-Scores: 89,00 und 89,55), erreicht TnT bessere Ergebnisse basierend auf STTS (TiGer F-Scores: 83,96 und 78,56; TüBa-D/Z F-Scores: 90,40 und 90,80) und auf STTSmorph (TiGer F-Scores: 77,74 und 71,70; TüBa-D/Z F-Scores: 86,41 und 87,34). Interessanterweise zeigt sich, dass die vom Parser selbst verteilten Tags zu besseren Parsingergebnissen führen (außer für TiGer mit UTS). Die entsprechenden F-Scores liegen i.d.R. zwischen den F-Scores für Gold-POS-Tags und den F-Scores für TnT, was zeigt, dass das Tagging von den Informationen im Split/Merge-Modell des Berkeley Parser profitiert – die Teilung und Verschmelzung von Tags hängt auch vom syntaktischen Kontext ab. Im Vergleich zu dem Experiment, in dem wir Gold-Standard POS-Tags verwenden zeigt sich, dass die Fehlerraten vom POS-Tagging zum Parsing größtenteils konstant bleiben. D.h. wenn der POS-Tagger eine Fehlerrate von 2% hat, verschlechtern sich die Parsingergebnisse von Gold-Tags zu automatischen Tags um etwa dieselbe Größenordnung. Das bedeutet, dass Fehler im POS-Tagging zwar einen negativen Einfluss auf die Qualität der Parses haben, dass aber nur in minimalem Umfang Nachfolgefehler auftreten. Beim morphologischen Tagset, STTSmorph, ist die Differenz zwischen den Parser-Ergebnissen mit Gold- und TnT-Tags sogar deutlich geringer als die Fehlerrate beim POS-Tagging, was bedeutet, dass Fehler in der Morphologie z.T. keine Auswirkung auf die Baumstruktur haben.

Vergleicht man die Performanz des Parsers basierend auf den verschiedenen Varianten des Tagsets, zeigen die beiden Baumbanken unterschiedliche Tendenzen: In TiGer sind die Unterschiede zwischen UTS und STTS minimal, d.h. TiGer kann nicht von der zusätzlichen Information im STTS profitieren. In der TüBa-D/Z dagegen bewirkt der Wechsel von UTS zu STTS eine Verbesserung von ca. 2%, d.h. die feinkörnigere Information im STTS ist hilfreich für die Disambiguierung von Konstituenten. Ein Grund dafür ist in der Tatsache zu finden, dass die TüBa-D/Z strukturelle Unterschiede zwischen Haupt- und untergeordneten Sätzen macht: nebenordnende Konjunktionen (KON) werden unter das Feld KOORD gruppiert, unterordnende Konjunktionen (KOUS und KOUI) unter C. Außerdem werden Relativsätze mit einem eigenen Knotenlabel (R-SIMPX) gekennzeichnet (siehe Abbildung 2 für ein Beispiel). Aus diesem Grund ist die Unterscheidung zwischen nebenordnenden und unterordnenden Konjunktionen wichtig. Diese werden im UTS unter einem gemeinsamen Tag (CONJ) gruppiert. In der TiGer-Baumbank dagegen werden alle Sätze unter S gruppiert (siehe Abbildung 1), da die Funktionsinformation beim Parsing nicht verwendet wurde. D.h. diese Unterscheidung im STTS ist nicht von Bedeutung. Eine andere für die TüBa-D/Z wichtige Unterscheidung ist diejenige zwischen finiten und infiniten Verben, da diese sich in der Verbalphrase widerspiegelt.

Die morphologischen Informationen in STTSmorph haben einen negativen Einfluss auf die Parsingergebnisse: Der F-Score für TiGer fällt um ca. 4,5% bei Gold-Tags und

	TiGer		TüBa-D/Z	
Morphologie	Dev.	Test	Dev.	Test
STTS	**97,15%**	**96,29%**	**96,92%**	**97,00%**
STTSmorph	85,77%	82,77%	84,67%	85,45%
Kongruenz	86,04%	83,08%	84,96%	85,77%
Kasus	88,10%	86,47%	87,48%	87,91%
Numerus	95,60%	94,19%	95,24%	95,41%
Numerus + Person	95,55%	94,11%	95,18%	95,24%
Verb-Merkmale	97,03%	96,02%	96,55%	96,44%

Tabelle 7: Die Ergebnisse für TnT mit verschiedenen morphologischen Varianten.

um ca. 6% bei POS-Tags von TnT. Bei der TüBa-D/Z ist die Differenz etwas geringer, sie liegt um 1% bei Gold-Tags und zwischen 2,5% und 3,5% bei TnT Tags. Daraus können wir schließen, dass der Parser die morphologische Information nicht sinnvoll verwenden kann, selbst wenn Gold-POS-Tags vorhanden sind.

Die insgesamt besten Parsingergebnisse erreichen wir bei TiGer und TüBa-D/Z mit dem internen POS-Tagging des Berkeley Parsers. Im „Pipeline“-Modell mit automatischen POS-Tags erreichen wir die besten Ergebnisse für TiGer mit der Kombination von SVMTool und STTS und für TüBa-D/Z mit TnT und STTS. Diese Ergebnisse bestätigen unsere Vermutung, dass das morphologische Tagset zu viel Information enthält, die automatisch nicht hochwertig genug annotiert werden kann, um für das Parsing hilfreich zu sein. Daraus ergibt sich die Frage, ob es Untermengen von morphologischen Merkmalen gibt, die zuverlässig mittels eines POS-Taggers annotiert werden können und die für das Parsing hilfreich sind. Diese Frage wird in den nächsten Abschnitten untersucht.

5.3 Morphologische Varianten

In diesem Satz von Experimenten untersuchen wir, ob es syntaktisch relevante Untermengen von morphologischen Merkmalen gibt, die sich zuverlässiger annotieren lassen als die komplette Menge morphologischer Merkmale. Diese Experimente wurden mit TnT durchgeführt, weil er sich in den vorhergehenden Experimenten als der zuverlässigste POS-Tagger erwiesen hat. Die Ergebnisse dieser Experimente sind in Tabelle 7 aufgelistet. Zu Vergleichszwecken wiederholen wir die Ergebnisse für das STTS und die STTS-Variante mit kompletter Morphologie (STTSmorph) aus Tabelle 4.

Die Ergebnisse zeigen, dass es morphologische Untermengen gibt, die zuverlässige Ergebnisse ermöglichen: Wenn nur Verb-Merkmale verwendet werden, erreichen wir Ergebnisse, die nur leicht unter denen der Standard-STTS-Variante liegen. Bei Numerus + Person und bei Numerus alleine beträgt die Differenz ca. 2%. Die Varianten, die nur Kasus oder alle Kongruenzmerkmale verwenden, sind leicht (Kasus) bzw. deutlich (Kongruenz) schlechter als die komplette Morphologie. Daraus ergibt sich die Frage,

inwieweit sich diese Ergebnisse auf das Parsing übertragen lassen. Dies wird im nächsten Abschnitt untersucht.

Es ist weiterhin auffällig, dass in den meisten Fällen die Unterschiede zwischen dem Development- und dem Testset in TiGer größer sind als die Unterschiede zwischen TiGer und der TüBa-D/Z. Ein Grund dafür ist sicher die hohe Anzahl der morphologischen Tags in TiGer, die zu einem hohen Maß von Tags rührt, die im Development- und Testset vorkommen, aber nicht im Trainingsset.

5.4 Parsing mit morphologischen Varianten

In diesem Satz von Experimenten untersuchen wir, wie sich die morphologischen Varianten des STTS aus den vorhergehenden Experimenten auf das Parsing auswirken. Wir verwenden dasselbe Setup wie in Abschnitt 5.2, d.h. wir verwenden keine grammatischen Funktionen beim Parsing, wir verwenden Gold-POS-Tags in der jeweiligen morphologischen Kombination fürs Training, und jeweils die Wörter in Kombination mit den Gold-POS-Tags oder den Ausgaben von TnT als Eingabe für den Parser.

Die Ergebnisse dieser Experimente finden sich in Tabelle 8. Zu Vergleichszwecken wiederholen wir die Ergebnisse für die STTS-Varianten mit und ohne komplette Morphologie (STTSmorph und STTS) aus Tabelle 6.

Die Ergebnisse zeigen, dass alle morphologischen Variationen basierend auf Gold-Tags bessere Parsingergebnisse erzielen als die Varianten mit kompletter Morphologie. Dies gilt für beide Baumbanken. Wenn die POS-Tags auf TnT-Ausgaben basieren, verbessern sich die Ergebnisse in allen Fällen, bis auf die Kombination von Kongruenzmerkmalen für die TiGer-Baumbank. In diesem Fall sind die Ergebnisse für die komplette Morphologie um ca. 2% besser. Des Weiteren fällt auf, dass die Ergebnisse für TüBa-D/Z mit kompletter Morphologie und mit Kongruenzmerkmalen extrem ähnlich (für das Developmentset) oder identisch (für das Testset) sind. Eine weitere Analyse zeigt, dass die Parser-Ergebnisse nur minimale Unterschiede aufweisen, obwohl die POS-Tagging-Ergebnisse sich deutlicher unterscheiden. Daraus können wir schließen, dass die Kongruenzmerkmale einen Großteil der Merkmale in der TüBa-D/Z ausmachen, die schwierig für POS-Tagger sind, die jedoch einen Einfluss auf den Parsebaum haben. D.h. wenn diese Merkmale mit größerer Genauigkeit annotiert werden könnten, hätten sie wohl einen positiven Einfluss auf die Parsequalität.

Bei einer näheren Betrachtung der Ergebnisse wird deutlich, dass für die TüBa-D/Z die Ergebnisse für die Merkmalskombinationen Numerus + Person und Verb-Merkmale mit den Ergebnissen für das Standard-STTS vergleichbar sind: Für das STTS erreicht der Parser einen F-Score von 90,40 bzw. 90,80 basierend auf TnT POS-Tags. Für Numerus + Person erreicht der Parser 90,03 und 90,37 und für Verb-Merkmale 90,41 und 90,88. Dies bedeutet, dass wir diese morphologischen Merkmale zuverlässig genug annotieren können, dass sie aber für das Konstituenzparsing nicht aussagekräftig genug sind. Im Gegensatz dazu sind die Ergebnisse für die TiGer-Baumbank durchgehend und deutlich schlechter als die Ergebnisse basierend auf dem UTS. D.h., in der TiGer-

Parser-Eingabe	TiGer Dev. prec.	TiGer Dev. rec.	TiGer Dev. F	TiGer Test prec.	TiGer Test rec.	TiGer Test F	TüBa-D/Z Dev. prec.	TüBa-D/Z Dev. rec.	TüBa-D/Z Dev. F	TüBa-D/Z Test prec.	TüBa-D/Z Test rec.	TüBa-D/Z Test F
	STTS											
gold	87,57	86,55	87,06	83,30	82,01	82,65	94,28	94,00	94,14	94,77	94,38	94,57
TnT	84,09	83,83	**83,96**	78,69	78,43	**78,56**	90,26	90,53	90,40	90,66	90,94	90,80
	STTSmorph											
gold	82,47	83,14	82,80	77,64	78,37	78,00	93,21	93,44	93,33	93,87	93,87	93,87
TnT	75,90	79,67	77,74	69,89	73,61	71,70	85,25	87,61	86,41	86,32	88,39	87,34
	Kongruenz											
gold	82,94	83,36	83,15	78,56	78,73	78,64	93,41	93,68	93,55	93,91	94,00	93,96
TnT	73,24	78,13	75,60	67,12	72,31	69,62	85,56	87,97	86,75	86,32	88,38	87,34
	Kasus											
gold	86,36	85,50	85,93	82,68	81,72	82,20	94,28	94,09	94,19	94,73	94,35	94,54
TnT	78,50	80,39	79,43	72,47	74,83	73,63	87,46	88,74	88,09	88,13	89,19	88,66
	Numerus											
gold	86,22	85,41	85,81	81,94	80,94	81,43	94,06	93,90	93,98	94,60	94,33	94,47
TnT	82,41	82,59	82,50	76,66	76,99	76,82	89,81	90,42	90,12	90,10	90,73	90,42
	Numerus + Person											
gold	86,25	85,52	85,88	81,95	80,93	81,44	94,09	93,94	94,01	94,48	94,17	94,33
TnT	82,58	82,79	82,68	76,59	76,90	76,74	89,73	90,34	90,03	90,11	90,64	90,37
	Verb-Merkmale											
gold	86,46	85,42	85,94	81,98	80,81	81,39	94,23	94,07	94,15	94,69	94,41	94,55
TnT	82,98	82,74	82,86	77,32	77,08	77,20	90,25	90,57	**90,41**	90,71	91,04	**90,88**

Tabelle 8: Die Ergebnisse des Berkeley Parsers auf verschiedenen morphologischen Varianten.

Baumbank können die morphologischen Merkmale keine nützlichen Informationen fürs Konstituenzparsing zur Verfügung stellen.

6 Diskussion

POS-Tags bzw. morphologische Merkmale bilden eine Schnittstelle zwischen der lexikalischen Ebene in einer Baumbank und den eigentlichen syntaktischen Bäumen. Wie schon eingangs erwähnt, kann ein einzelnes POS-Tag als eine Äquivalenzklasse von Wörtern gesehen werden (z.B. über alle finiten Verben mit einer bestimmten Kombination von Person und Numerus oder über alle Verben insgesamt). Ein Parser konstruiert nun den eigentlichen Syntax-Baum, ausgehend von den POS-Tags, nicht direkt auf den Wörtern. Die Haupterkenntnis, die wir aus unseren Experimenten mitnehmen, ist, dass der Erfolg einer Kombination von POS-Tagging und Parsing durch ein Zusammenspiel verschiedener Faktoren bestimmt wird. Faktoren sind die verwendeten Ansätze zur Kombination bzw. zur Disambiguierung von einzelnen Tags, die Charakteristika der Syntax-Bäume und die Granularität des POS-Tagsets.

Wir haben uns bei unserer Arbeit auf den „Pipeline"-Ansatz beschränkt, bei dem die Ausgabe eines POS-Taggers als Eingabe des Parsers dient. Wie in anderen Arbeiten (siehe Abschnitt 2) scheint auch bei uns die Qualität der Parserausgabe stark von der Qualität der Tag-Eingabe abzuhängen. Je komplexer das Tagset, desto mehr Informationen können dem Parser zur Verfügung gestellt werden. Allerdings sind nicht alle Informationen hilfreich. Und je komplexer das Tagset ist, desto schlechter können die Tagger zwischen einzelnen Tags für ein Wort disambiguieren. Fehler beim Tagging führen dann beim Parsen zu qualitativ schlechten Bäumen. Die „richtige" Granularität des POS-Tagsets hängt wiederum vom syntaktischen Annotationsschema ab. Während für die eher flach annotierte TiGer-Baumbank das Universal Tagset am besten abschneidet, so liegt bei der TüBa-D/Z das STTS vorne.

Es ist zu erwarten, dass sich ein anderes Bild ergäbe, würde man mit grammatischen Funktionen parsen und auswerten. In diesem Falle könnte der Parser von dem Parallelismus zwischen Kasus und grammatischer Funktion von Konstituenten profitieren. Für gute Ergebnisse sollte hier ein feineres POS-Tagset mit morphologischer Information effektiver sein. Allerdings muss diese morphologische Annotation mit einem hohen Korrektheitsgrad annotiert werden. Ansonsten leidet die Qualität des Parsers. Daher sollte erwogen werden, die morphologische Disambiguierung nicht durch einen POS-Tagger sondern durch eine spezialisierte morphologische Komponente, wie z.B. *Morfette* (Chrupala et al., 2008), vorzunehmen.

Ein für das Parsing ideales POS-Tagset hätte die Eigenschaft, dass es die einzelnen Wörter in Äquivalenzklassen einteilt, abhängig vom Potential einer Äquivalenzklasse, bei der Unterscheidung von Parse(unter)bäumen zu helfen. Diese Äquivalenzklassen müssen nicht notwendigerweise mit linguistisch motivierten Äquivalenzklassen zusammenfallen – ideal wäre ein Maß, mit dem dieser Informationsgehalt eines einzelnen Tags, bzw. einer Wort-Äquivalenzklasse, ausgedrückt werden könnte. Abhängig von der betrachteten Einzelsprache kommen verschiedene Techniken für die automatische Ausbildung solcher

Äquivalenzklassen durch Clustering in Betracht. Sollte man auf die Reproduktion eines bestimmten anderen POS-Tagsets angewiesen sein, so können diese Äquivalenzklassen auch nach dem Parsing durch die Tags aus dem entsprechenden POS-Tagset ersetzt werden.

Abgesehen davon scheint der „Pipeline"-Ansatz nicht unbedingt die ideale Kombination der beiden Aufgaben von Tagging und Parsing zu sein. Im „Pipeline"-Ansatz werden die POS-Tags bereits *vor* dem Parsing komplett disambiguiert. Dies hat den Vorteil, dass die Aufgabe des Parsers vereinfacht wird, was i.a. zu einer höheren Parsing-Geschwindigkeit führt. Allerdings ist der Parser an diese Entscheidungen gebunden, auch wenn sich herausstellen sollte, dass die POS-Sequenz nicht dem syntaktischen Modell des Parsers entspricht. Da bereits mehrere erfolgreiche kombinierte Modelle für gleichzeitige Tag-Disambiguierung und Dependenzparsing (oder kürzlich auch fürs Konstituenzparsing) vorgestellt wurden, sollten derartige Ansätze in Zukunft weiter verfolgt werden. Hierbei muss, wie schon vorher erwähnt, ein Maß gefunden werden, wie viel Information bzw. welcher Ambiguitätsgrad in den POS-Tags beim Parsing zu optimalen Ergebnissen führt. Des Weiteren muss abgeklärt werden, bis zu welcher morphologischen Granularität ein kombiniertes Modell erfolgreich arbeiten kann. In jedem Fall bleibt in kombinierten Modellen eine weitere Schwierigkeit bestehen. Für eine ideale Modellierung der Einzelkomponenten können verschiedene Modelle erforderlich sein, die schwierig zu kombinieren sind. So ist u.U. die Finite-State-Technologie am besten für eine morphologische Komponente geeignet, Markov-Modelle für das Part-of-Speech-Tagging und ein bestimmter Grammatikformalismus für das eigentliche Parsing.

7 Schlussbetrachtung

In unserer Arbeit haben wir den Einfluss von Part-of-Speech-Tagsets auf Parsingergebnisse untersucht. Dies geschah unter Benutzung des Stuttgart-Tübingen-Tagsets, morphologischer Varianten desselben, und des Universal Tagsets. Unsere Experimente haben wir in einem „Pipeline"-Ansatz durchgeführt, bei dem die Ausgabe eines Taggers als Eingabe des Parsers fungiert. Es kamen mehrere Tagger zum Einsatz, basierend auf einem Markov-Modell, auf einem Maximum-Entropy-Modell, auf Conditional Random Fields, und auf Support Vector Machines. Als Parser wurde der Berkeley Parser verwendet.

Unsere Ergebnisse zeigen, dass es deutliche Unterschiede in Bezug auf die Qualität der POS-Tagger gibt und dass die Auswahl des POS-Taggers auch von der Granularität des Tagsets abhängig gemacht werden muss. Für TiGer besteht die beste Kombination aus dem Universal Tagset und SVMTool, während für die TüBa-D/Z die Kombination aus STTS und TnT bessere Ergebnisse erbringt. Morphologische Information erweist sich beim Konstituentenparsing als wenig hilfreich, selbst wenn diese Information vollständig korrekt ist. Weitere Forschung ist nötig, um eine geeignete Repräsentation der morphologischen Information zu finden, die es dem Parser erlaubt, sie gewinnbringend einzusetzen.

Literatur

Apidianaki, M., Dagan, I., Foster, J., Marton, Y., Seddah, D. und Tsarfaty, R. (Hgg.) (2012). *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*. Association for Computational Linguistics, Jeju, Republik Korea.

Bangalore, S. und Joshi, A. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In: *Proceedings of the 23rd International Conference on Computational Linguistics (IJCNLP)*, Seiten 89–97, Peking, China.

Bohnet, B. und Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Seiten 1455–1465, Jeju, Republik Korea.

Boyd, A. (2007). Discontinuity revisited: An improved conversion to context-free representations. In: *Proceedings of The Linguistic Annotation Workshop (LAW) at ACL 2007*, Seiten 41–44, Prag, Tschechische Republik.

Brants, S., Dipper, S., Hansen, S., Lezius, W. und Smith, G. (2002). The TIGER treebank. In: *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT)*, Seiten 24–41, Sozopol, Bulgarien.

Brants, T. (1998). *TnT–A Statistical Part-of-Speech Tagger*. Universität des Saarlandes, Computational Linguistics, Saarbrücken, Deutschland.

Brants, T. (2000). TnT–a statistical part-of-speech tagger. In: *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (ANLP/NAACL)*, Seiten 224–231, Seattle, WA, USA.

Briscoe, T. und Carroll, J. (2002). Robust accurate statistical annotation of general text. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Seiten 1499–1504, Las Palmas, Spanien.

Candito, M. und Seddah, D. (2010). Parsing word clusters. In: *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Seiten 76–84, Los Angeles, CA, USA.

Charniak, E., Carroll, G., Adcock, J., Cassandra, A., Gotoh, Y., Katz, J., Littman, M. und McCann, J. (1996). Taggers for parsers. *Artificial Intelligence*, 85(1-2):45–57.

Chen, X. und Kit, C. (2011). Improving part-of-speech tagging for context-free parsing. In: *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*, Seiten 1260–1268, Chiang Mai, Thailand.

Chrupala, G., Dinu, G. und van Genabith, J. (2008). Learning morphology with Morfette. In: *Proceedings the Fifth International Conference on Language Resources and Evaluation (LREC)*, Marrakesch, Marokko.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Doktorarbeit, University of Pennsylvania, Philadelphia, PA, USA.

Csendes, D., Csirik, J., Gyimóthy, T. und Kocsor, A. (2005). The Szeged Treebank. In: Matoušek, V., Mautner, P. und Pavelka, T. (Hgg.), *Text, Speech and Dialogue: Proceedings of TSD 2005*, Seiten 123–131. Springer.

Curran, J. R., Clark, S. und Vadas, D. (2006). Multi-tagging for lexicalized-grammar parsing. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Seiten 697–704, Sydney, Australien.

Dalrymple, M. (2006). How much can part-of-speech tagging help parsing? *Natural Language Engineering*, 12(4):373–389.

Daum, M., Foth, K. und Menzel, W. (2003). Constraint based integration of deep and shallow parsing techniques. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, Ungarn.

Farkas, R. und Schmid, H. (2012). Forest reranking through subtree ranking. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*, Seiten 1038–1047, Jeju, Republik Korea.

Foth, K., Daum, M. und Menzel, W. (2005). Parsing unrestricted German text with defeasible constraints. In: Christiansen, H., Skadhauge, P. R. und Villadsen, J. (Hgg.), *Constraint Solving and Language Processing*, Seiten 140–157. Springer.

Giménez, J. und Màrquez, L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Seiten 43–46, Lissabon, Portugal.

Goldberg, Y. und Tsarfaty, R. (2008). A single generative model for joint morphological segmentation and syntactic parsing. In: *Proceedings of The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL:HLT)*, Seiten 371–379, Columbus, OH, USA.

Hajič, J., Böhmová, A., Hajičová, E. und Vidová-Hladká, B. (2000). The Prague Dependency Treebank: A three-level annotation scenario. In: Abeillé, A. (Hg.), *Treebanks: Building and Using Parsed Corpora*, Seiten 103–127. Kluwer, Amsterdam.

Hatori, J., Matsuzaki, T., Miyao, Y. und Tsujii, J. (2011). Incremental joint POS tagging and dependency parsing in Chinese. In: *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*, Seiten 1216–1224, Chiang Mai, Thailand.

Hinton, G. (1999). Products of experts. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks*, Seiten 1–6, Stockholm, Schweden.

Höhle, T. (1986). Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In: *Akten des Siebten Internationalen Germanistenkongresses 1985*, Seiten 329–340, Göttingen, Deutschland.

Joachims, T. (1999). Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C. und Smola, A. (Hgg.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Koo, T., Carreras, X. und Collins, M. (2008). Simple semi-supervised dependency parsing. In: *Proceedings of The 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL:HLT)*, Seiten 595–603, Columbus, OH, USA.

Kübler, S., Hinrichs, E. W. und Maier, W. (2006). Is it really that difficult to parse German? In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seiten 111–119, Sydney, Australien.

Kübler, S., Maier, W., Rehbein, I. und Versley, Y. (2008). How to compare treebanks. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Seiten 2322–2329, Marrakesch, Marokko.

Kübler, S. und Penn, G. (Hgg.) (2008). *Proceedings of the Workshop on Parsing German at ACL-08*. Association for Computational Linguistics, Columbus, OH, USA.

Lafferty, J. D., McCallum, A. und Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, Seiten 282–289, San Francisco, CA, USA.

Lakeland, C. (2005). *Lexical Approaches to Backoff in Statistical Parsing*. Doktorarbeit, University of Otago, Neuseeland.

Le Roux, J., Sagot, B. und Seddah, D. (2012). Statistical parsing of Spanish and data driven lemmatization. In: *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, Seiten 55–61, Jeju, Republik Korea.

Maier, W. (2006). Annotation schemes and their influence on parsing results. In: *Proceedings of the COLING/ACL 2006 Student Research Workshop*, Seiten 19–24, Sydney, Australien.

Maier, W., Kaeshammer, M. und Kallmeyer, L. (2012). Data-driven PLCFRS parsing revisited: Restricting the fan-out to two. In: *Proceedings of the Eleventh International Conference on Tree Adjoining Grammars and Related Formalisms (TAG+11)*, Paris, Frankreich.

Marcus, M. P., Santorini, B. und Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. Special Issue on Using Large Corpora: II.

McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit*. `http://mallet.cs.umass.edu`.

Petrov, S., Barrett, L., Thibaux, R. und Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Seiten 433–440, Sydney, Australien.

Petrov, S., Das, D. und McDonald, R. (2012). A universal part-of-speech tagset. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Türkei.

Prins, R. und van Noord, G. (2003). Reinforcing parser preferences through tagging. *Traitement Automatique des Langues, Special Issue on Evolutions in Parsing*, 44(3):121–139.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP)*, Seiten 133–142, Philadelphia, PA, USA.

Rehbein, I. und van Genabith, J. (2007). Treebank annotation schemes and parser evaluation for German. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Seiten 630–639, Prag, Tschechische Republik.

Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell, J. T. und Johnson, M. (2002). Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Seiten 271–278, Philadephia, PA, USA.

Schiller, A., Teufel, S., Stöckert, C. und Thielen, C. (1995). Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, und Seminar für Sprachwissenschaft, Universität Tübingen.

Seddah, D., Kübler, S. und Tsarfaty, R. (Hgg.) (2010). *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*. Association for Computational Linguistics, Los Angeles, CA, USA.

Seddah, D., Tsarfaty, R. und Foster, J. (Hgg.) (2011). *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*. Association for Computational Linguistics, Dublin, Irland.

Seeker, W. und Kuhn, J. (2013). Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics*, 39(1):23–55.

Skut, W., Krenn, B., Brants, T. und Uszkoreit, H. (1997). An annotation scheme for free word order languages. In: *Proceedings of the 5th Applied Natural Language Processing Conference (ANLP)*, Seiten 88–95, Washington, DC, USA.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H. und Beck, K. (2012). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Deutschland.

Thielen, C. und Schiller, A. (1994). Ein kleines und erweitertes Tagset fürs Deutsche. In: Feldweg, H. und Hinrichs, E. (Hgg.), *Lexikon & Text*, Seiten 215–226. Niemeyer, Tübingen.

Toutanova, K., Klein, D., Manning, C. und Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Seiten 252–259, Edmonton, Kanada.

Toutanova, K. und Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, Hong Kong.

Versley, Y. (2005). Parser evaluation across text types. In: *Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spanien.

Versley, Y. und Rehbein, I. (2009). Scalable discriminative parsing for German. In: *Proceedings of the 11th International Conference on Parsing Technologies (IWPT)*, Seiten 134–137, Paris, Frankreich.

Watson, R. (2006). Part-of-speech tagging models for parsing. In: *Proceedings of the 9th Annual CLUK Colloquium*, Milton Keynes, UK.

Xia, F. (2000). The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). IRCS Technical Report IRCS-00-07, University of Pennsylvania, Philadelphia, PA, USA.

Yoshida, K., Tsuruoka, Y., Miyao, Y. und Tsujii, J. (2007). Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers. In: *IJCAI'07: Proceedings of the 20th International Joint Conference on Artifical Intelligence*, Seiten 1783–1788, Hyderabad, Indien.

Anhang: Das STTS-Tagset

ADJA	attributives Adjektiv
ADJD	adverbiales oder prädikatives Adjektiv
ADV	Adverb
APPR	Präposition; Zirkumposition links
APPRART	Präposition mit Artikel
APPO	Postposition
APZR	Zirkumposition rechts
ART	Bestimmter oder unbestimmter Artikel
CARD	Kardinalzahl
FM	Fremdsprachliches Material
ITJ	Interjektion
KOUI	Unterordnende Konjunktion mit zu und Infinitiv
KOUS	Unterordnende Konjunktion mit Satz
KON	Nebenordnende Konjunktion
KOKOM	Vergleichspartikel, ohne Satz
NN	Normales Nomen
NE	Eigennamen
PDS	Substituierendes Demonstrativpronomen
PDAT	Attribuierendes Demonstrativpronomen
PIS	Substituierendes Indefinitpronomen
PIAT	Attribuierendes Indefinitpronomen
PPER	Ersetzbares Personalpronomen
PPOSS	Substituierendes Possessivpronomen
PPOSAT	Attribuierendes Possessivpronomen
PRELS	Substituierendes Relativpronomen
PRELAT	Attribuierendes Relativpronomen
PRF	Reflexives Personalpronomen
PWS	Substituierendes Interrogativpronomen
PWAT	Attribuierendes Interrogativpronomen

Tabelle 9: Die 54 Tags des STTS (Teil 1).

PWAV	Adverbiales Interrogativ- oder Relativpronomen
PROAV/PAV	Pronominaladverb
PTKZU	`zu` vor Infinitiv
PTKNEG	Negationspartikel
PTKVZ	Abgetrennter Verbzusatz
PTKANT	Antwortpartikel
PTKA	Partikel bei Adjektiv oder Adverb
TRUNC	Kompositions-Erstglied
VVFIN	Finites Verb, voll
VVIMP	Imperativ, voll
VVINF	Infinitiv, voll
VVIZU	Infinitiv mit `zu`, voll
VVPP	Partizip Perfekt, voll
VAFIN	Finites Verb, aux.
VAIMP	Imperativ, aux.
VAINF	Infinitiv, aux.
VAPP	Partizip Perfekt, aux.
VMFIN	Finites Verb, modal
VMINF	Infinitiv, modal
VMPP	Partizip Perfekt, modal
XY	Nichtwort, Sonderzeichen
$,	Komma
$.	Satzbeendende Interpunktion
$(	Sonstige Satzzeichen; satzintern
NNE	Kombination aus Nomen und Eigenname

Tabelle 10: Die 54 Tags des STTS (Teil 2).

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`

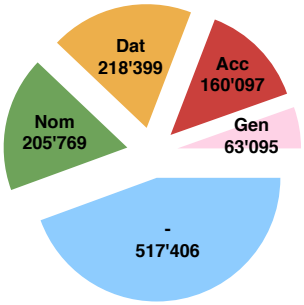


Abbildung 1: Verteilung der Kasus über allen 1.164.766 Tokens der TUEBA. Kasuslose Tokens („-“) dominieren stark. Der häufigste Kasus ist Dativ (Dat), gefolgt von Nominativ (Nom), Akkusativ (Acc) und Genitiv (Gen).

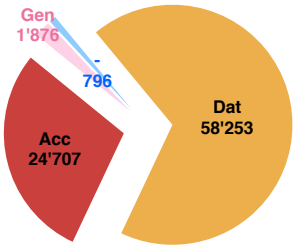


Abbildung 2: Verteilung der Kasus über allen 85.632 mit APPR klassifizierten Tokens der TUEBA (7,4% aller Tokens haben das STTS-Tag APPR). Der häufigste Kasus Dativ deckt 68% aller Fälle ab, Akkusativ 29% und Genitiv gut 2%. Nur 1% der Token, welche mit APPR getaggt sind, haben keine Kasusrektionsinformation.

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.

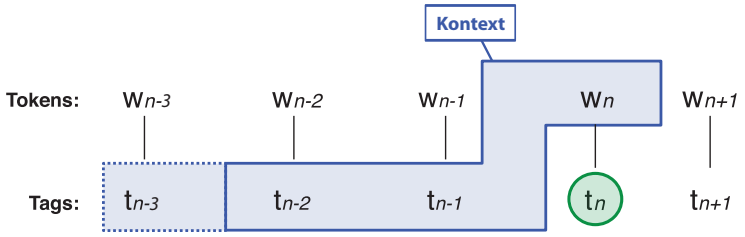


Abbildung 3: Kontextmodell eines Trigramm- bzw. Quadrigramm-Taggers

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

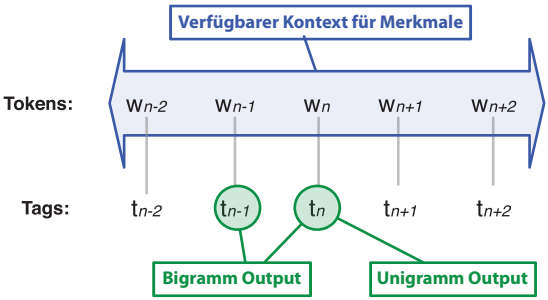


Abbildung 4: Kontextmodell eines sequentiellen CRF-Taggers

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	$_{\Delta}\mathrm{CI}_{low}$
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 $_{\Delta}\mathrm{CI}_{low}$ 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`).

Marc Reznicek, Heike Zinsmeister

STTS-Konfusionsklassen beim Tagging von Fremdsprachlernertexten

1 Motivation

Für viele aktuelle Fragestellungen der Zweit- und Fremdspracherwerbsforschung („L2-Erwerbsforschung“) sind Lernerkorpora unverzichtbar geworden. Sie stellen Texte von L2-Lernern[1] zur Verfügung, oftmals ergänzt durch vergleichbare Texte von Muttersprachlern der Zielsprache.

Beschränkten sich Analysen der Lernerkorpusforschung in den ersten Jahren hauptsächlich auf einzelne Wortformen (vgl. Granger, 1998), hat sich das Forschungsinteresse beständig hin zu komplexeren grammatischen Kategorien entwickelt. Dazu zählen u.A. die Untersuchung tiefer syntaktischer Analysen (Dickinson und Ragheb, 2009; Hirschmann et al., 2013, u.a.) oder die Strategien der Markierung von Kohärenzrelationen (z.B. Breckle und Zinsmeister, 2012). Derartige Analysen bauen dabei nur selten auf der Textoberfläche selbst auf, sondern setzen i.d.R. die Annotation von Wortarten für jedes Texttoken voraus und ggfs. weitere, darauf aufbauende Annotationsebenen.

Annotationen dienen generell immer der Suche nach Klassen in den Daten, die anhand der Oberflächenformen allein nicht leicht zugänglich wären (im Kontext von Lernerkorpora vgl. Díaz-Negrillo et al., 2010). Ist man z.B. an einer Analyse von Possessivpronomen interessiert, würde man bei einer Korpussuche, die nur Zugriff auf die Wortformen selbst hat, bei der ambigen Form *meinen* neben Beispielen für das Possessivpronomen (1) auch alle Belege für die gleichlautende Verbform (2) finden. Das Suchergebnis wäre also sehr ‘unsauber’, da die Wortform selbst keinen Aufschluss über ihre Interpretation gibt. Eine Annotation mit Wortarten würde die beiden Lesarten disambiguieren und damit die Rückgabe der Suchanfrage präziser machen. Die Rückgabe würde weniger ungewünschte Lesarten enthalten, die man andernfalls bei der Ergebnissichtung manuell ausschließen müsste. Kurz gesagt, eine Suchanfrage auf Wortarten-annotierten Daten ist für den Nutzer effizienter als eine Suche auf reinen Wortformen.

(1) Ich schäme mich noch immer für **meinen** Einsatz in Gorleben.
(tiger.release.dec05.0.302)[2]

(2) Viele Spanier **meinen** ohnehin, daß die „Transicion“ erst dann vollendet ist, wenn es einen reibungslosen Regierungswechsel gibt.
(tiger.release.dec05.1068)

[1] Der Konsistenz und Lesbarkeit halber verwendet dieser Text die maskuline Form von *Lerner, Sprecher* usw. Selbstverständlich schließen die Nennungen auch weibliche Lerner, Sprecher usw. ein.

[2] Alle Korpusbeispiele sind auffindbar über `https://korpling.german.hu-berlin.de/annis3`.

Für die Wortartenannotation des Deutschen werden von vielen Projekten die 54 Wortartenklassen des sog. kleinen Stuttgart-Tübingen-Tagsets STTS (Schiller et al., 1999) genutzt, welches sich dadurch als eine Art Standard etabliert hat. Da das STTS für die Analyse von geschriebener Standardsprache entwickelt wurde (z.B. für Zeitungstexte), stellt sich die Frage, in wie weit es auch für die Analyse von lernersprachlichen Texten einsetzbar ist.

Beispiel (3) illustriert typische Probleme, mit denen man sich bei der Wortartenanalyse von Lernertexten auseinandersetzen muss. Zum Beispiel existiert der Ausdruck *Kriminal* zielsprachlich nur als gebundenes Morphem, wird hier aber selbstständig genutzt. Das Wort *heutzutage* ist zielsprachlich ein Adverb, das nicht flektiert und nicht attribuierend wie ein Adjektiv zusammen mit einem Nomen verwendet werden kann, hier aber genau in dieser Verwendung erscheint und zudem großgeschrieben ist, was zielsprachlich außer am Satzanfang nur Substantiven und Substantivierungen vorbehalten ist.[3]

(3) Jeden Tag viele **Kriminal Aktivitäten** passiert in der **Heutzutager** Gesellschaft. (FalkoEssayL2v2.4)

In dieser Studie soll nun untersucht werden, inwieweit sich die Besonderheiten geschriebener Lernersprache wie in Beispiel (3) illustriert auf das STTS abbilden lassen, um dann Vorschläge dafür zu präsentieren, wie man mit problematischen Fällen umgehen kann.

Um beispielhaft zu illustrieren, welche Wortartentags ein automatischer Tagger nichtzielsprachlichen Strukturen in Lernertexten zuweist, zeigt Beispiel (4) den Satz aus Beispiel (3) erneut diesmal erweitert mit automatisch zugewiesenen STTS-Tags.

Für das Token *Heutzutager* deutet die Information des lexikalischen Stammes *heutzutage* eindeutig auf ADV (Adverb), Großschreibung und *-er* -Suffix hingegen legen ein NN (normales Nomen) nahe. Bezogen auf das Auftreten eines einzelnen Tokens ist allerdings zwischen ART (Artikel) *der* und dem Nomen *Gesellschaft* in der Zielsprache distributionell sehr stark ADJA (attribuierendes Adjektiv) bevorzugt.

(4) Jeden Tag viele Kriminal/NN Aktivitäten passiert/VVPP in der/ART Heutzutager/NN Gesellschaft/NN. (TreeTagger@FalkoEssayL2v2.4)

Für die Anwendbarkeit des STTS auf Lernervarietäten ergeben sich aus diesen Beobachtungen drei Fragestellungen.

- Welche Strukturen in Texten deutscher Lerner werden durch das STTS nicht adäquat abgedeckt?

- Wie lassen sich Lücken in der Beschreibbarkeit lernersprachlicher Strukturen durch das zielsprachliche System aufdecken?

[3] Beispiel (3) weist weitere Eigenheiten auf z.B. dass das Vorfeld vor dem finiten Verb *passiert* mit zwei Konstituenten besetzt ist und dass *passiert* und das Subjekt im Numerus nicht kongruieren.

- Kann das STTS so erweitert oder anders angewendet werden, dass die quantitative Vergleichbarkeit zwischen Ziel- und Lernersprache aufrechterhalten wird, ohne die Information sich widersprechender POS-Tag-Hinweise zu eliminieren?

Im weiteren Verlauf dieses Artikels werden wir zuerst weiter auf die allgemeine Problematik der Wortartenklassifizierung von Lernersprache eingehen (Abschnitt 2). In Abschnitt 3 stellen wir unsere Untersuchung zum automatischen Tagging von Lernersprache vor, die ermittelt, inwieweit aktuelle auf Zeitungssprache trainierte Wortarten-Tagger in der Lage sind, die Kategorien des STTS korrekt auf Lernersprache abzubilden. Hierbei führen wir eine quantitative Methode ein, besonders interessante Fälle lernersprachlicher Strukturen zu ermitteln. Auf der Basis einer qualitative Analyse dieser Fälle (Abschnitt 5) diskutieren wir anschließend ausführlich verschiedene Herangehensweisen, den Herausforderungen des Wortartentaggings von Lernersprache gerecht zu werden (Abschnitt 6). Abschnitt 7 fasst den Artikel abschließend kurz zusammen.

2 Wortartentagging

Das Konzept von Wortarten hat einen heterogenen Charakter – ganz unabhängig von der Klassifikationen im Stuttgart-Tübingen-Tagset. Im Folgenden skizzieren wir Faktoren dieser Heterogenität und setzen sie mit den Anforderungen des automatischen Taggings und den deskriptiven Anforderungen von Lernersprache in Beziehung.

2.1 Wortarten als Merkmalsbündel

Wortarten können auf Informationen unterschiedlicher linguistischer Ebenen beruhen. Die gängigen Wortartenklassifikationen in beschreibenden Grammatiken berufen sich auf „das syntaktische Prinzip als primäres Kriterium“ (Helbig und Buscha, 2007, S. 19). Hierbei werden über die Darstellung von Prototypen (Eisenberg, 2004, S. 36) und Tests (vgl. u.a. Helbig und Buscha, 2007, S. 19) Klassen auf der Grundlage von Informationen zumindest dreier Ebenen definiert: lexikalische Information, morphologische Markierung und Distribution im Satz (vgl. Díaz-Negrillo et al., 2010, S.3). Die Gewichtung der unterschiedlichen Ebenen ist dabei nicht immer gleich. Anders als Helbig und Buscha (2007) geht das STTS bei seiner Klassifikation zunächst von einer morphologischen Unterscheidung aus (flektierbar vs. nicht-flektierbar). Distributionelle Eigenschaften werden erst im zweiten Schritt berücksichtigt, aus dem sich dann die Aufteilung der elf Hauptwortarten des STTS ableitet, vgl. Tabelle 1.
Diese Gewichtung wird beispw. bei der Unterteilung der Adjektive und Adverben deutlich. So werden sowohl *schnelle* in (5) als auch *schnell* in (6) als Adjektiv klassifiziert (attributiv vs. prädikativ), da das abstrakte lexikalische Element *schnell* prinzipiell flektierbar ist, während das nicht flektierbare aber mit *schnell* distributionell identische *gestern* in Beispiel (7) zu den Adverben zählt.[4]

[4] Für eine ausführliche Diskussion der problematischen Kategorisierung von Adverben vgl. Hirschmann (2013).

1.	Nomina (N)	7.	Adverben (ADV)
2.	Verben (V)	8.	Konjunktionen (KO)
3.	Artikel (ART)	9.	Adpositionen (AP)
4.	Adjektive (ADJ)	10.	Interjektionen (ITJ)
5.	Pronomina (P)	11.	Partikeln (PTK)
6.	Kardinalzahlen (CARD)		

Tabelle 1: Hauptwortarten des STTS (Schiller et al., 1999, S. 4)

> *Als Adverbien werden nur reine, nicht von Adjektiven abgeleitete, nicht flektierbare Modifizierer von Verben, Adjektiven, Adverbien und ganzen Sätzen verstanden.* (Schiller et al., 1999, S. 56)

(5) Die **schnelle/ADJA** Bearbeitung war erfreulich.

(6) Der Beamte arbeitete **schnell/ADJD**.

(7) Der Beamte arbeitete **gestern/ADV**.

2.2 Automatische Wortartenzuweisung

Automatische Wortartentagger (POS-Tagger)[5] besitzen oft zwei Komponenten: eine, die unmittelbar auf die lexikalische Information im Lexikon des Taggers zugreift, und eine, die ggfs. anschließend aus den alternativen Analysen auswählt. Die lexikalisch-morphologische Komponente ordnet dabei jeder Wortform (bzw. jedem Token) alle möglichen Wortartentags zu, die im Lexikon für diese Wortform aufgelistet sind. Ist eine Form im Lexikon nicht vorhanden, weisen einige Systeme mögliche Tags anhand einer morphologischen Analyse zu.[6] Die Disambiguierungskomponente nutzt, je nach System, entweder Regeln oder Wahrscheinlichkeiten von Tag-Abfolgen ('syntaktische Information')[7] oder auch komplexe Eigenschaftsbündel[8]. Führt keine dieser Methoden zu einem eindeutigen Tag, weichen Tagger auf robuste Lösungen aus, indem sie bsplw. das frequenteste Tag im Kontext übernehmen.

Lexikalische und syntaktische Informationen ermitteln viele Tagger statistisch aus manuell annotierten, sog. Trainingsdaten, bei welchen es sich aus praktischen Gründen oftmals um Zeitungsartikel handelt. Als Konsequenz sind die Tagger im Normalfall

[5] „POS" für Englisch *part-of-speech*

[6] Einer der Gutachter wies darauf hin, dass die hier angedeutete reduzierte morphologische Analyse von vollständigen morphologischen Analysen abgegrenzt werden sollte, wie sie von Systemen wie z.B. Morfette (Chrupala et al., 2008) oder SMOR (Schmid et al., 2004) durchgeführt wird.

[7] Wahrscheinlichkeiten von Tag-Abfolgen (v.a. in Hidden Markov Modellen, vgl. Jurafsky und Martin, 2009, Kapitel 5 & 6)

[8] Conditional Random Fields (Lafferty et al., 2001) erlauben es, Merkmale voneinander unabhängiger Ebenen zu nutzen wie z.B. orthographische und distributionelle Information, sowie ein unterschiedlich weites Fenster an Token vor oder nach dem zu taggenden Token auf den entsprechenden Ebenen zu betrachten.

für Standardsprache in der Form, wie sie in überregionalen Zeitschriften vertreten ist, optimiert.

In prototypischen Fällen sind sowohl lexikalische Information als auch morphologische Markierung und syntaktische Verteilung miteinander kompatibel; in Einzelfällen widersprechen sich diese allerdings wie in Beispiel (8). Während die lexikalische Information für *Wenn* und *Aber* auf eine Konjunktion bzw. Adverb verweist, erlaubt der Kontext zunächst nur (Pro)Nomen (oder Substantivierungen).[9]

(8) Schröder ist überzeugt, daß die SPD „ohne **Wenn** und **Aber** mit der Union um ökonomische Kompetenz konkurriert ..." (tiger_release_dec05_616)

In Fällen wie diesem müsste ein Hierarchisierungsmechanismus angewandt werden, der entscheidet, welcher Informationstyp Vorrang haben soll. In automatischen POS-Taggern ist dies normalerweise nur eingeschränkt möglich, da die lexikalische Information, wie oben beschrieben, oft als erster Filter dient, dessen Ausgabe von den anderen Informationstypen nur disambiguiert, aber nicht vollkommen überschrieben werden kann.

2.3 Wortarten in Lernersprache

Kann die Entscheidung für eines der STTS-Tags in Zeitungssprache (zumindest von menschlichen Experten) noch in hohem Maße eindeutig getroffen werden, ist diese Situation in Lernersprache völlig anders. Lernersprache weicht in vielen Aspekten systematisch von der Zielsprache ab (Selinker, 1972; Corder, 1986), wobei ein Teil dieser Abweichungen, wie in Beispiel (3) in Abschnitt 1 illustriert, auch die Entscheidungen des Wortartentaggings betrifft.

Aus computerlinguistischer Perspektive kann das Tagging von Lernersprache als der Versuch beschrieben werden, trotz fehlerhafter Daten korrekte Tagging-Ergebnisse im Sinne der Zielsprache zu erzielen, d.h. eine "robuste" Analyse auf "verrauschtem" Input zu erzeugen (vgl. van Rooy und Schäfer, 2002).

Studien zur Lernersyntax verfolgen hauptsächlich zwei unterschiedliche Ansätze: Fehleranalyse und kontrastive Interlanguage-Analyse. In der kontrastiven Interlanguage-Analyse (u.a. Granger, 1996, 2008) werden Subkorpora miteinander verglichen um signifikante Frequenzunterschiede auszumachen, meistens werden dabei Lernertexte mit Muttersprachlertexten verglichen.[10] Dieser Vergleich ist nur möglich, wenn über die gleichen Kategorien hinweg verglichen wird, d.h. wenn die Tagsets für Lernersprache und Zielsprache identisch oder zumindest aufeinander abbildbar sind. Vor diesem Hintergrund wird verständlich, weshalb der Versuch, eine eigene Lernersprachengrammatik mit einer eigenen Wortartenklassifizierung zu schreiben, keine großen Verfechter gefunden hat.[11]

[9] Unter "Kontext" ist hier die syntaktische Präpositionalphrase mit Koordination zu verstehen, nicht nur die lineare POS-Abfolge.

[10] Andere Vergleichsgruppen sind z.B. Texte von Lernern unterschiedlicher Muttersprachen, mit denen Transfereffekte nachgewiesen werden können. Vergleiche unterschiedlicher Kompetenzniveaus wiederum erlauben eine entwicklungsbezogene Analyse.

[11] Für andere Nichtstandard-Varietäten, beisplw. gesprochene Sprache, ist dies dagegen geschehen (Hennig und Bücker, 2008).

Im nächsten Abschnitt stellen wir eine Methode vor, um die Frage zu beantworten, wie sich sich Lücken in der Beschreibbarkeit lernersprachlicher Strukturen durch das zielsprachliche System aufdecken lassen.

3 Experiment

Um problematische Fälle der Analyse von Lernersprache mit STTS-Tags zu identifizieren, verwenden wir eine halbautomatische Methode, die auf der Idee des *Ensembletaggings* (van Halteren et al., 2001) beruht.

3.1 Tagging

Ein einzelner Tagger macht bestimmte Fehler beim Tagging. Verschiedene Tagger unterscheiden sich potenziell bei den Fehlern, die sie machen. Diese Beobachtung kann für das automatische Tagging und dessen manuelle Korrektur fruchtbar gemacht werden: Annotiert man denselben Text parallel mit mehreren Taggern, kann man davon ausgehen, dass die Einheitsentscheidung der Tagger wahrscheinlich korrekt ist, Unterschiede hingegen auf potenziell problematische Instanzen hinweisen. Wenn es darum geht, möglichst effizient eine gute Annotation zu erreichen, beschränkt man sich bei der manuellen Nachannotation auf diese Unterschiedsfälle.

In der vorliegenden Untersuchung gehen wir davon aus, dass die Unterschiedsfälle neben reinen Taggingfehlern zusätzlich auf potenzielle Abweichungen in der Lernersprache hinweisen. Um diese beiden Typen von Unterschiedsfällen zu trennen, gleichen wir die Ergebnisse der Lernertexte mit Annotationen vergleichbarer muttersprachlicher Texte ab und betrachten in der weiteren Analyse nur solche Unterschiedsfälle, die für die Lernertexte im Vergleich zu den muttersprachlichen Texten markant waren.

Für das Ensembletagging verwendeten wir drei Wortarten-Tagger, die frei verfügbar und gut dokumentiert vorliegen: den TreeTagger (Schmid, 1994, 1995), den RFTagger (Schmid und Laws, 2008) und den Stanford Tagger (Toutanova und Manning, 2000). Anstelle die Tagger auf die Lernerdaten durch Re-Training und anderen Methoden der Domänenadaptation optimal anzupassen, sollten in der Untersuchung gerade Standardmodule zum Einsatz kommen, die auf zielsprachlichen Daten trainiert worden waren. Die Abweichungen des Tagger-Ensembles an Stellen, bei denen sich die Lernersprache nicht zielsprachlich verhält, soll uns auf Probleme in der Beschreibung durch das STTS aufmerksam machen, siehe dazu die Ergebnisse in Abschnitt 5 und die Diskussion in Abschnitt 6.[12]

[12] Wir danken einem der Gutachter für den Hinweis, dass der TNT-Tagger sehr gute Ergebnisse auf STTS-annotierten Daten liefert (vgl. Giesbrecht und Evert, 2009). In zukünftigen Experimenten sollte dieser Tagger mit berücksichtigt werden. Für die vorliegende qualitative Untersuchung beschränken wir uns jedoch auf die im Text genannten Tagger.

3.2 Datengrundlage

Als Datengrundlage dienen die Texte des Kobalt-Korpus (`www.kobalt-daf.de`). Es handelt sich hierbei um Aufsätze von fortgeschrittenen Deutschlernenden, die die Frage diskutieren: „Geht es der Jugend heute besser als früher?“ Tabelle 2 fasst die Zusammenstellung des Korpus anhand der Erstsprachen der Autoren zusammen.[13]

Texttyp	Erstsprache	ISO 639-3	Textanzahl	Tokenanzahl
L2	Weißrussisch	BEL	20	14.401
L2	Chinesisch (Mandarin)	CMN	20	11.724
L2	Schwedisch	SWE	10	5.537
L1	Deutsch	DEU	20	12.410

Tabelle 2: Zusammenstellung des Kobalt-Korpus nach Erstsprachen (Release 1.4)

3.3 Normalisierung

Das Kobalt-Korpus beinhaltet neben den Originaltexten mehrere Normalisierungsebenen, sog. Zielhypothesen (vgl. Lüdeling et al., 2008; Reznicek et al., 2013). Für die vorliegende Untersuchung ist die Ebene der *grammatischen Zielhypothese* (ZH1) relevant, für die jeder Lernersatz systematisch mit einer morpho-syntaktisch grammatischen Entsprechung annotiert wird. Semantische und pragmatische Abweichungen bleiben bei dieser Ebene unberücksichtigt. Für die eigentliche Datenauswertung verwendeten wir eine vereinfachte Variante der Zielhypothese, bei der Bewegungen ignoriert wurde (ZH0)

Tabelle 3 illustriert die ZH1 und ZH0 für den Satz aus Beispiel (4). Die rechte Spalte zeigt automatisch generierte Differenztags, die auf unterschiedliche Abweichungstypen im Lernertext hinweisen.[14]

In der tokenbasierten Korrektur für die ZH1 werden unter Beibehaltung des finiten Verbs sowohl Wortstellung und Kongruenzbedingungen korrigiert (*viele Kriminal Aktivitäten*) als auch lexikalische Ersetzungen durchgeführt (*Heutzutager*). Es sei denn, man findet einen Kontext und eine Lesart, bei denen keine Korrektur notwendig wäre. *Kriminalaktivität* ist grammatisch möglich und verlangt die geringste Korrektur: Die Anzahl der mechanischen Veränderungen (edit distance) im Vergleich zu *kriminelle Aktivität* ist zwar gleich, die phonologische Abweichung ist allerdings geringer. Daher wird es nicht durch das vielleicht gängigere *kriminelle Aktivität* ersetzt.[15]

[13] Die OnDaF-Testergebnisse (`www.ondaf.de`) der Autoren entsprachen etwa dem Kompetenzniveau B2 nach dem Europäischen Referenzrahmen („oberes Mittelmaß“).

[14] Nach einer Konvention aus der Falko-Annotation, wird bei dem Tag MERGE, wie bei *Kriminal Aktivitäten* im Beispiel, kein zusätzliches CHANGE markiert. Die Differenztags werden in der vorliegenden Studie nicht betrachtet, sollten aber in zukünftigen Studien genauer untersucht werden.

[15] Eine detaillierte Operationalisierung des Konzepts *geringste Korrektur* muss noch geleistet werden.

LT	ZH1	**ZH0**	ZH0-Differenztag
Jeden	Jeden	Jeden	
Tag	Tag	Tag	
viele		**viel**	**CHA**NGE
Kriminal Aktivitäten		**Kriminalaktivität**	**MERGE**
passiert	passiert	passiert	
	viel Kriminalaktivität		
in	in	in	
der	der	der	
Heutzutager	**heutigen**	**heutigen**	**CHA**NGE
Gesellschaft	Gesellschaft	Gesellschaft	
.	.	.	

Tabelle 3: Normalisierung von Lernertext (LT) im Sinne der grammatischen Zielhypothese (ZH1). 'CHANGE'-Tag markieren eine Änderung der Buchstabenkette, 'MERGE' das Verschmelzen von Token. ZH0 entsteht durch die Wiederherstellung der urprünglichen Wortstellung aus ZH1 und liegt der aktuellen Untersuchung zugrunde.

4 Quantitative Analyse

Für die Untersuchung wurden beide Ebenen, der Lernertext als auch die Zielhypothese, mittels Ensembletagging mit STTS annotiert. Bei Nicht-Übereinstimmung der Tagger wurde die Mehrheitsentscheidung ausgegeben. Wenn alle drei Tagger unterschiedliche Tags vorschlugen, wurde auf die TreeTagger-Ausgabe als Defaultlösung zurückgegriffen, weil sie unabhängig die höchste Akkuratheit der drei Tagger auf ZH1 erreichte. Da es für den Lernertext keine „richtige" Annotation gibt, kann es auch keinen unmittelbaren Goldstandard geben, für die Normalisierung (ZH1), die der Standardgrammatik entspricht, aber schon. Daher wurden alle Fälle, bei denen die drei Tagger auf der ZH1 nicht übereinstimmten, manuell von zwei Annotatoren überprüft und ggf. korrigiert. Diese korrigierte Fassung nutzten wir auch als Goldstandard für die Lernertexte. Neben der im folgenden vorgestellten Akkuratheitsbestimmung, dient der Vergleich in erster Linie dazu, durch Abweichungen in den Annotationen potenziell nicht-zielsprachliche Strukturen in den Lernertexten zu markieren und auffindbar zu machen.

4.1 Taggingergebnisse

Tabelle 4 fasst die Akkuratheit der Tagger in Bezug auf diesen manuell erstellten (Quasi-)Goldstandard zusammen. In einem Kontrollexperiment wurden in jeweils drei Texten pro Erstsprache alle Token manuell überprüft. Die Differenz zum eigentlichen Experiment deutet an, in welchem Umfang die Taggerleistung zu optimistisch eingeschätzt wird, wenn nur die nicht-übereinstimmenden Tags korrigiert, Taggerfehler, die sich hinter einer Taggerübereinstimmung verbergen, aber ignoriert werden. Für das Weißrussische (BEL) z.B. wird als durchschnittliche Akkuratheit des Tagger-Ensembles auf den Lernertexten 96,8 % gemessen. Auf den drei Texten des Kontrollexperiments, bei denen alle Tags

manuell korrigiert wurden, auch die, bei denen die drei Tagger sich einig waren, liegt die durchschnittliche Akkuratheit bei etwa 96,2 %, also 0,6 %-Punkte niedriger.

Die Taggingergebnisse auf der Zielhypothese sind erwartungsgemäß besser als auf den Lernertexten selbst und auch die Varianz zwischen den Texten wird geringer, wie die Standardabweichungen in den Klammern zeigen. Die Akkuratheit der ZH1 für das Weißrussische liegt z.B. bei 98,0 %, also um 1,2 %-Punkte besser als das Ergebnis auf der Lernertextebene (LT).[16]

	Experiment		Kontrolle	
	LT	ZH1	LT	ZH1
BEL	96,8 (±1,2)	98,0 (±0,8)	96,2 (±1,0)	97,2 (±1,0)
CMN	97,1 (±1,5)	98,3 (±0,8)	96,9 (±11,2)	97,8 (±0,6)
SWE	95,0 (±2,0)	97,7 (±0,9)	94,8 (±2,3)	97,0 (±0,7)
DEU	95,8 (±1,6)	97,8 (±0,9)	95,6 (±1,6)	96,5 (±0,8)

Tabelle 4: Durchschnittliche Tagging-Akkuratheit auf dem Kobalt-Korpus, welche manuell nur für Token, bei denen die Tagger nicht übereinstimmten, korrigiert wurde (links), und auf einer Kontrollgruppe von je drei Texten pro Sprache, die manuell für alle Token korrigiert wurde (rechts); Standardabweichung in Klammern.

Es fällt auf, dass die Tagging-Ergebnisse auf den muttersprachlichen DEU-Texten nicht die besten sind, sondern hinter den Ergebnissen auf den BEL- und CMN-Texten zurückbleiben. Dies lässt sich damit erklären, dass die deutschen Texte ebenfalls Tippfehler und andere für die Tagger unbekannte Wörter enthalten und potenziell komplexere Strukturen verwenden. Zudem besteht eine schwache negative Korrelation zwischen durchschnittlicher Satzlänge und Taggingakkuratheit (Spearmans Rangkorrelationskoeffizient, rho = -0.26, $p<0.05$):[17] Je länger die Satzlänge desto niedriger die Taggingakkuratheit.

Sprache	Min.	Median	Mean	Max.
BEL	10,80	14,00	15,23	23,75
CMN	11,57	15,08	15,00	18,00
SWE	12,16	17,64	17,30	26,58
DEU	12,80	18,83	19,97	32,35

Tabelle 5: Durchschittliche Satzlänge pro Text im Kobalt-Korpus (in Token pro Satz; ergänzt um die minimale und maximale beobachtete Satzlänge).

[16] Uns ist bewusst, dass auch gerade die Fälle interessant sein können, in denen die Tagger sich zwar einig, der Goldstandard aber abweichend ist. Für diese Untersuchung war der erstellte Goldstandard allerdings zu klein.

[17] Wir verwendeten den nicht-parametrischen Rangkorrelationstest nach Spearman, da nach dem Shapiro-Wilk-Test weder die durchschnittlichen Satzlängen noch die Akkuratheitswerte annähernd normalverteilt sind.

In der folgenden Untersuchung vergleichen wir die automatischen Tags der LT-Ebene mit den Tags der ZH0-Ebene und kontrastieren die Lernerdaten mit den muttersprachlichen DEU-Daten.

4.2 Konfusionsklassen

Wenn Lernersprache Strukturen beinhaltet, in denen sich die unterschiedlichen Hinweise in Hinblick auf Distribution, morphologische Markierung und lexikalische Information widersprechen, sollten die Tagger bei der Vergabe dieser Tags besonders häufig falsch liegen. Der Abgleich der vergebenen Tags mit einem Goldstandard wird klassischerweise in einer Konfusionsmatrix (Abbildung 1) dargestellt. In einer solchen Tabelle wird der Taggeroutput (vertikal) zu den Tags im Goldstandard (horizontal) in Beziehung gesetzt. Bei einem perfekten Tagger würden alle Ergebnisse auf der Diagonalen liegen. In Abbildung 1 werden der Übersichtlichkeit halber nur Verwechslungen angezeigt und die Diagonale ausgespart.

Nicht alle Verwechslungen können auf Besonderheiten der Lernertexte zurückgeführt werden. So können auch andere Variablen (z.B. Textsorteneffekt) den auf Zeitungssprache optimierten Taggern Schwierigkeiten bereiten. Um diesen Einfluss möglichst auszuschließen, betrachten wir ähnlich wie in klassischen *Under*- und *Overuse*-Studien (Granger und Tyson, 1996; Hirschmann et al., 2013) nur die signifikanten Abweichungen zwischen L1- und L2-Texten. In Abbildung 1 werden daher nur die in Lernersprache häufiger auftretenden Wortartenkonfusionen aufgezeigt. Größe zeigt dabei die logarithmische Häufigkeit, Schwärze den Grad der Signifikanz an. So lässt sich aus der Graphik ablesen, dass die Verwechslungen von satzinterner (**$(**) mit satzexterner (**$.**) Interpunktion sowie von Partizipien Perfekt von Vollverben (**VVPP**) mit entsprechenden Infinitiven (**VVINF**) jene sind, die sich in den L2-Texten am signifikantesten häufiger finden als in den L1-Texten (dunkelste Kreise). Gleichzeit kann man sehen, dass insgesamt am häufigsten Infinitive von Modalverben (**VMINF**) mit finiten Modalverben (**VMFIN**) verwechselt werden (größter Kreis), dass diese aber nicht besonders lernerspezifisch ist (heller Kreis). Sortiert man die Konfusionsklassen zuerst nach Signifikanz und dann nach Häufigkeit, erhält man die Liste in Tabelle 6.

Anhand der rein quantitativen Analyse lässt sich allerdings noch nicht entscheiden, wie die Konfusionen zu interpretieren sind. Um diese Frage zu klären, wollen wir im nächsten Abschnitt einige Beispiele genauer betrachten.

5 Qualitative Analyse

In diesem Abschnitt analysieren wir beispielhaft Vertreter der in Abschnitt 4.2 eingeführten Konfusionsklassen, um festzustellen, ob sie tatsächlich auf Lücken im Tagset hinweisen oder von Experten hätten korrekt getaggt werden können.

In den Beispielen (9)-(16) werden für jede der gefundenen Konfusionsklassen zwei Beispiele präsentiert. Für jeden Fall wird explizit gemacht, ob Lexik, Morphologie und Distribution dem Standard entsprechen oder nicht. So hat das Tagger-Ensemble für

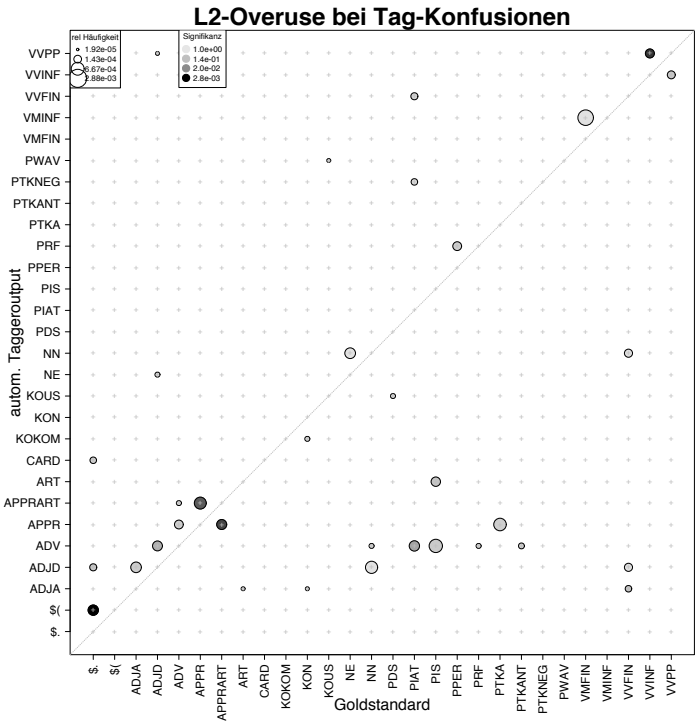


Abbildung 1: L2-Overuse bei Tag-Konfusionen auf dem Goldstandard (70 Texte: 43.285)
größer = häufiger in L1 und L2, dunkler = Kontrast L1 vs. L2 signifikanter

das Token *besprochen* im Beispiel (9) als Partizip Perfekt (**VVPP**) getaggt. Im Goldstandard wurde das in eckigen Klammern angegebene Goldtoken *besprechen* allerdings als Infinitiv des Vollverbs (**VVINF**) getaggt. Die Tatsache, dass *besprochen* an dieser Stelle nichtstandardsprachlich verwendet wird, wird erst im Kontext des vorangehenden Modalverbs erkennbar. Dieser relevante Kontext ist in den Beispielen durch Unterstreichung markiert. Am Goldtoken *besprechen* lässt sich ableiten, dass das Lemma und die distributionelle Position dem Standard entsprechen, während die morphologische Markierung nicht dem Standard folgt. Im Beispiel (13) weicht die Zielhypothese nicht für das fälschlicherweise als Artikel (ART) statt als attributives Indefinitpronomen (PIAT) getaggte Token *mehr* ab, sondern für das benachbarte *günstige*. Hier wurde das *mehr* also in einem nichtstandardsprachlichen Kontext verwendet, seine Distribution ist somit nichtstandardsprachlich.

Die Beispiele zeigen, dass eine Konfusionsklasse durch unterschiedliche Widersprüche

L2		GOLD	Signifikanz L2 vs. L1	$\sum$ L2	$\sum$ L1
$(	$\neq$	$.	0.003	14	1
VVPP	$\neq$	VVINF	0.004	11	0
APPR	$\neq$	APPRART	0.004	15	0
APPRART	$\neq$	APPR	0.006	2	2
ADV	$\neq$	PIAT	0.048	14	1
ADV	$\neq$	ADJD	0.075	12	1
ART	$\neq$	PIS	0.190	10	1
ADJD	$\neq$	ADJA	0.259	13	2

Tabelle 6: L2-spezifische Tagging-Fehler nach Signifikanz (L2 vs. L1) sortiert und Häufigkeiten

	Lexik	Morphologie	Distribution
9)	**VVPP**$_{LT}$ $\neq$ **VVINF**$_{Gold}$		
	Standard	**Nichtstandard**	**Standard**
	Wenn bei mir etwas passierte, kann ich dass mit meinen Eltern **besprochen**[⇒ **besprechen**]. (kobalt_BEL_018_2011_03)		
10)	**APPR**$_{LT}$ $\neq$ **APPRART**$_{Gold}$		
	Nichtstandard	**Standard**	**Standard**
	Junge Menschen sind oft **nach**[⇒ **zur**] Universität oder Arbeit gezogen. (kobalt_SWE_011_2012_03)		
11)	**APPRART**$_{LT}$ $\neq$ **APPR**$_{Gold}$		
	Nichtstandard	**Standard**	**Nichtstandard**
	früher hatte man einfach nicht die Moglichkeit herumzusitzen, fern zu schauen, **vom**[⇒ **vor** dem] Computer stundenlang zu sitzen. (kobalt_SWE_006_2011_12)		
12)	**VVINF**$_{LT}$ $\neq$ **VVPP**$_{Gold}$		
	Standard	**Nichtstandard**	**Standard**
	In den meisten Fällen machten sie nur das, was von ihnen **erwarten**[⇒ **erwartet**] wurde. (kobalt_BEL_012_2011_03)		

	Lexik	Morphologie	Distribution
13)		**ADV**$_{LT}$ ≠ **PIAT**$_{Gold}$	
	Standard	**Standard**	**Nichtstandard**
	Wir haben heute **mehr** günstigen[⇒ günstige] Möglichkeiten, um unseren richtigen Platz in der Welt zu finden (kobalt_BEL_015_2011_03)		
14)		**ADV**$_{LT}$ ≠ **ADJD**$_{Gold}$	
	Nichtstandard	**Standard**	**Standard**
	Wir machen es aber für uns einfach und nehmen Schweden, das kleine Land ganz **viel**[⇒ **weit**] oben auf dem Erdball. (kobalt_SWE_007_2011_12)		
15)		**ART**$_{LT}$ ≠ **PIS**$_{Gold}$	
	Standard	**Nichtstandard**	**Standard**
	Zuvor wurden das Familiensleben und das Lernen des Jugendes wegen Kriege und Reformen, die **ein** [⇒ **einer**] nach dem anderen kam, zerstört. (kobalt_CM_020_2011_03)		
16)		**ADJD**$_{LT}$ ≠ **ADJA**$_{Gold}$	
	Standard	**Nichtstandard**	**Standard**
	Andererseits, so argumentieren sie, hat die jüngere Generation zum Glück eine Welt mit **relativ**[⇒ **relativem**] Frieden und Freiheit. (kobalt_CM_008_2011_03)		

Lexik	Morphologie	Distribution	Beispiele
S	S	**NS**	13
S	**NS**	S	9, 12, 15, 16
NS	S	S	10, 14
NS	S	**NS**	11

Tabelle 7: Informationen auf linguistischen Ebenen
S: Standard, NS: Nichtstandard

der linguistischen Ebenen bedingt sein kann. Tabelle 7 fasst die einzelnen Kombinationen zusammen. Aufgrund dieser Beispiele lässt sich jetzt auch erkennen, dass die Verwechslungen tatsächlich Lernerstrukturen ermitteln, die durch das Tagset nicht beschrieben werden können, anstatt lediglich Schwächen der Tagger aufzuzeigen. Wäre dem nicht so, müsste man für die Beispiele Lesarten finden können, in denen sich die Ebenen nicht widersprechen. Dies ist aber nicht der Fall.

6 Diskussion

Wie im Abschnitt 5 deutlich geworden ist, können die Tags des STTS eine Reihe von Strukturen in Lernersprache nicht abdecken. Wir wollen drei Ansätze besprechen, mit diesem Problem umzugehen: Mehrdimensionale Tags (Abschnitt 6.1), Portemanteau-Tags (Abschnitt 6.2) und unterspezifizierte Tags (Abschnitt 6.3).

6.1 Mehrdimensionale Tags

Die konsequenteste Lösung wurde bereits von Díaz-Negrillo et al. (2010) für das Tagging von englischen Lernertexten vorgeschlagen. Um Lernersprache und Zielsprache einheitlich beschreiben zu können, wäre es sinnvoll, die Einzelinformationen auf den drei Ebenen (Lexik, Morphologie und Distribution) getrennt in einem mehrdimensionalen Tags (*tripartite POS*) anzugeben. Die Schwäche dieses Ansatzes liegt allerdings darin, dass die einzelnen Ebenen unabhängig voneinander beschrieben werden. Prinzipiell könnte man zwar sowohl die lexikalische als auch die morphologische Information aus einer Liste ziehen. Die Ambiguitäten jeder einzelnen Ebene werden normalerweise allerdings erst durch die Schnittmenge mit den anderen beiden Ebenen beherrschbar. Will man die Ebenen aber gerade unabhängig voneinander machen, so müsste das neue POS-Tags alle möglichen Kombinationen beinhalten. Dies lässt sich gut anhand von *erwarten* im Beispiel (12) zeigen, hier als (17) wiederholt.

(17) In den meisten Fällen machten sie nur das, was von ihnen erwarten [⇒ erwartet] wurde.

Rein lexikalisch handelt es sich um ein Vollverb (**VVFIN**, **VVINF**, **VVPP** oder **VVIMP** (letzteres mit abweichender Morphologie). Morphologisch im Sinne von Paradigmen können kleingeschriebene Wortformen mit Endung auf *en* attribuierende Adjektive (**ADJA**), finite oder nicht-finite Vollverben sein (**VVINF**, **VVIZU**, **VVFIN**, **VVPP**). Bezieht man geschlossene Wortformen mit ein, erhöht sich die Anzahl der möglichen Lesarten um Artikel (**ART**), Auxiliar- und Modalverben (**VAFIN**, **VAINF**, **VAPP**, **VMFIN**, **VMINF**), Pronomen (**PIAT**, **PPOSAT**, **PRELS**, **PDS**, **PIS**, **PRELAT**, **PDAT**, **PPER**, **PWAT**, **PWS**, **PPOSS**) und Kardinalzahlen (**CARD**). Operationalisiert man die morphologische Analyse einfach nur als Bedingung auf der reinen Buchstabenkette, erhöht sich die Anzahl der möglichen Tags erneut (z.B. **APPR**, **ADV**, **ADJD**, **PTKVZ**), da es insgesamt nur wenige Tags gibt, die keine kleingeschriebene Wortformen mit der Endung *en* beschreiben (z.B. **PTKNEG**, **KOKOM**,

KOUI). Insgesamt würde es sich anbieten, aus einem Korpus graduelle Erwartbarkeiten für die verschiedenen Tags zu ermitteln.[18]

Für die Distributionsebene stellt sich genauso wie bei der Morphologie die Frage nach der Operationalisierung, d.h. mit welchen Methoden die möglichen Tags ermittelt werden. Díaz-Negrillo et al. (2010) beziehen sich bei Beispielerklärungen auf den grammatischen Kontext (grammatical context) im Sinne von Phrasenstruktur und Selektionsbeschränkungen. Wenn man für das Beispiel (17) den umgebenden Teilsatz als distributionelle Suchmaske annimmt wie in (18a), dann sollte die Variable *X* mit **VVPP** gefüllt werden – gegeben, dass alle anderen Wörter korrekt formuliert sind. Wechselt man auf eine abstraktere Ebene und überlegt, welche Wortarten zwischen einer Präpositionalphrase mit Personalpronomen APPR PPER und einem finiten Auxiliarverb stehen können wie in (18b), wird die Liste der möglichen Tags natürlich länger z.B. **ADJD** wie in *...(dass der Vorsprung) vor ihm größer wurde.*

(18) a. ...was von ihnen X wurde.

b. ...APPR PPER X VAFIN

Aufgrund der oben genannten Schwierigkeiten, sehen wir in dieser Variante der Mehrebenenbeschreibung noch keinen gangbaren Weg. Weitere Forschung könnte hier einen Ausweg zeigen.

6.2 Portemanteau-Tags

In einer zweiten Variante werden Original- und Zielhypothesen-Information verbunden, indem das auf dem Originaltext vom Tagger-Ensemble zugewiesene Tag um das Tag im Goldstandard ergänzt wird. So trägt das Token *erwarten* in Beispiel (12) das Tag **VVINF_VVPP** vgl. (19).

(19) In den meisten Fällen machten sie nur das, was von ihnen **erwarten/VVFIN_VVPP** wurde.

Dieses Tag bringt zum Ausdruck, dass oberflächennähere Faktoren für einen Infinitiv sprechen, während unter Einbezug aller kontextuellen Informationen ein **VVPP** in der ZH1 stehen würde. Das Hauptproblem dieses Ansatzes liegt im fehlenden universellen Goldstandard für eine grammatische Zielhypothese. Wie bereits mehrfach erwähnt wurde, ist die Festlegung auf ein einziges Tag im Goldstandard nur möglich, indem man die Information auf einer der linguistischen Ebenen höher gewichtet als die einer anderen. Durch solch eine Hierarchisierung, geht die zugrundeliegende Information für die weitere Verarbeitung verloren. Die Kombination beider Ebenen kann dies jedoch zum Teil auffangen. Die Formulierung der grammatischen Zielhypothese folgt dabei stets den Anforderungen der jeweiligen Forschungsfrage (vgl. Reznicek et al., 2013).

[18] Zum Beispiel ergibt die Anfrage `#1:word=[a-zäöü.*en/]!` auf dem TiGer-Korpus (v2.1) 34 unterschiedliche STTS-Tags für die beschriebene Wortform, wobei sich die Erwartbarkeiten von 27% für ADJA bis hin zu unter einer Promille für z.B. NE, PWAT und APPO verteilen.

Diese Lösung ist in Mehrebenen-Korpora wie Kobalt oder Falko bereits umgesetzt. Hierbei werden die Tags allerdings nicht in einer Ebene verschmolzen, sondern sind parallel durchsuchbar, da neben dem Originaltext auch die Zielhypothese(n) und deren Annotationen ins Korpus integriert sind.

Dickinson und Ragheb (2013) verfolgen der Terminologie nach einen Mehrebenenansatz und beschreiben Lernersprache in einer morphologischen und einer distributionellen Ebene.[19] Die beiden Tags können sich dabei voneinander unterscheiden. Bei genauerer Betrachtung wird allerdings deutlich, dass es sich bei diesem Ansatz nicht um „echte" unabhängige Ebenen handelt, sondern dass die Tags auf der distributionellen Ebene eher der ZH1-Ebene entsprechen, auch wenn es sich nicht um explizite Zielhypothesen handelt. Damit gleicht die Annotation der beiden Ebenen eher der hier vorgestellten Portemanteau-Variante als den in Abschnitt 6.1 eingeführten mehrdimensionalen Tags.

> *We define a distributional slot as a position where a token with particular properties (e.g., singular noun) is predicted to occur, on the (syntactic) basis of its surrounding tokens.* (Dickinson und Ragheb, 2013, S. 27)

6.3 Unterspezifizierte Tags

Eine dritte Möglichkeit, Widersprüche zwischen den Ebenen in Lernersprache in der Beschreibung durch das STTS abzubilden, könnte darin bestehen, neue unterspezifizierte Tags zuzulassen. Innerhalb der Hauptwortarten des STTS ist dies durch den hierarchischen Aufbau der Tags schon heute leicht realisierbar. So könnten Überschneidungen aus **ADJD** und **ADJA** als **ADJ** getaggt werden. Das Tag **ADJ** ist dabei gegenüber beiden ursprünglichen Tag unterspezifiziert, da es ein Adjektiv beschreibt, aber keine Aussage darüber trifft, ob es attributiv oder prädikativ verwendet wird. Ein Blick in die relevanten Konfusionstypen in Tabelle 8 zeigt, dass sich ein Großteil der problematischen Fälle auf diese Weise abdecken lassen. Nur für Fälle, in denen Hauptwortarten überschritten werden (**ART** $\neq$ **PIS** & **ADV** $\neq$ **PIAT**), müssten neue Oberklassen im Sinne von unterspezifizierten Tags über zwei oder mehrere Hauptwortarten hinweg eingefügt werden.

Um die im Kobalt-Korpus untersuchten Strukturen adäquat abzubilden, würde es ausreichen, zwei solche Oberklassen einzuführen: **AD** als Vereinigung von Adjektiven und Adverben, sowie **ADPART**, die zusätzlich auch die Pronomen miteinschließt. Abbildung 2 zeigt die nötigen Unterspezifizierungen. Hierbei sind alle Hauptwortklassen bereits enthalten.

Inwieweit diese Oberklassen auch in anderen Varietäten genutzt werden können, muss untersucht werden. Gerade die Verschmelzung von Pronomen mit Adjektiven/Adverben scheint linguistisch unintuitiv, allerdings zeigen auch Clusteranalysen (vgl. Rapp, 2007) Ähnlichkeiten zwischen Wortformen, die nicht der linguistischen Intuition entsprechen.

[19] SALLE: `http://cl.indiana.edu/~salle/`

L2		GOLD		unterspez. Tag
VVPP	≠	VVINF	→	VV
VVINF	≠	VVPP		
APPR	≠	APPRART	→	AP
APPRART		APPR		
ADV	≠	ADJD	→	AD
ADJD	≠	ADJA		
ART	≠	PIS	→	PART
ADV	≠	PIAT	→	ADPART

Tabelle 8: Unterspezifikation für L2-relevante Konfusionsklassen

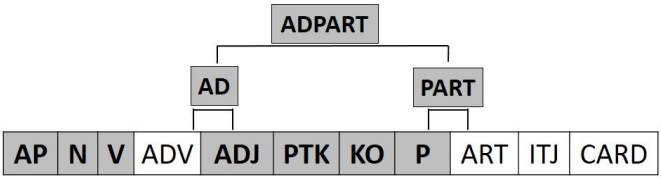


Abbildung 2: Vorschlag für Oberklassen (grau markiert) zu STTS-Hauptwortarten

6.4 Anwendung

Zwar sehen wir wie unter 6.1 für mehrdimensionale Tags noch keine Anwendungsmöglichkeiten, dagegen haben aber sowohl die Einbeziehung der ZH1 als auch die Unterspezifizierung von Tags anwendungsbezogene Stärken und Schwächen. Der Vorteil der Portemanteau-Tags liegt in der ausdifferenzierten Beschreibung der Form und der Funktion über die Tags der konkurrierenden Ebenen. Dies ist besonders für spezielle Suchen oder beispielsweise der Untersuchung von Frequenzkontrasten von POS-Ketten aussagekräftiger als die Reduktion der Tags auf Oberklassen. Für die weitere Verarbeitung (bspw. durch einen Parser) führt dieser Ansatz durch die Vielzahl möglicher Kombinationen der einzelnen Tags allerdings schnell zu *data sparseness* und verschlechtert die Ergebnisse des automatischen Trainings. Für diesen Fall scheinen die Oberklassen die bessere Lösung darzustellen, da durch die Reduktion auf weniger Tags die Anzahl der Instanzen in jeder Kategorie steigt. Darüber hinaus konnten Rehbein et al. (2012) zeigen, dass beim automatischen syntaktischen Parsing von Lernersprache nur bestimmte POS-Tag-Abweichungen die Ergebnisse verschlechtern, da sich eine Reihe von Tags in bestimmten Kontexten syntaktisch identisch verhalten.

> *This clearly shows that the overall accuracy is not enough to predict parsing scores, but that particular error types are more harmful for parser performance than others.* (Rehbein et al., 2012, S. 13)

In diesen Fällen könnte das Trainieren mit Oberklassen also sogar zu besseren Ergebnissen führen.

6.5 Einfügungen, Löschungen

Die hier vorgestellte Pilotstudie ignoriert bewusst einen wichtigen Bereich von Lernersprache: fehlende (1,6% aller Tokens) und überflüssige (0,78%) Wortformen, sowie falsche Auseinander-(0,27%) und Zusammenschreibungen (0,17%). Für all diese Fälle (insges. 2,82% aller Tokens) gibt es keine 1:1-Beziehung zwischen den Token im Lernertext und in der Zielhypothesenebene. So entspricht den beiden Tokens *Computer*/[**NN**] und *Spiele*/[**NN**] des Lernertextes im Beispiel (9) das Token *Computerspiele*/[**NN**] in der ZH1. Das Token *ein*/[**ART**] hat gar keine Entsprechung im Lernertext.

tok	**Computer**	**Spiele**	,	Online-Chatting		soziale	Netzwerk
pos	**NN**	**NN**	$(	NN		ADJA	NN
ZH1	**Computerspiele**		,	Online-Chatting	**ein**	soziales	Netzwerk
Npos	**NN**		$(	NN	**ART**	ADJA	NN
Diff	**MERGE**				**INS**		

Tabelle 9: (Kobalt_CMN_006_2011_03) tok: Lernertext, pos: automatische POS-Annotation des Lernertextes, ZH1: grammatische Zielhypothese, Npos: Goldtags auf der ZH1, Diff: Annotation der Abweichungen der ZH1 von tok. INS: im Lernertext nicht enthaltenes Token, MERGE: im Lernertext fehlerhaft zweigeteiltes Token

Während nicht vorhandene Tags (*ein*) für das Tagset kein Problem darstellen, sind Fälle von Zusammenschreibungen wie in Tabelle 9 deshalb problematisch, weil sie in der Kombination distributionell nicht an dieser **NN**-Positionen auftauchen können. Durch den Ausschluss dieser Fälle wurde sicherlich spannende Phänomene der Lernersprache für die Diskussion des STTS ignoriert, die hier vorgestellte Methode auf diese Fälle anzupassen, ist daher eine Aufgabe zukünftiger Forschung.

7 Zusammenfassung

In diesem Artikel haben wir gezeigt, dass das STTS in seiner derzeitigen Form und Verwendung (ein Token – ein Tag) eine Reihe sprachlicher Strukturen, die für Texte fortgeschrittener Deutschlerner typisch sind, nicht erfolgreich abbilden kann. Für die Aufdeckung problematischer Bereiche haben wir kontrastive Konfusionsmatrizen verwendet. Wir haben diskutiert, dass in bestimmten Anwendungskontexten sowohl Portemanteau als auch unterspezifierte Tags zu interessanten Verbesserungen führen

können. Einige lernersprachliche Phänomene lassen sich nicht über 1:1-Beziehungen zwischen Token im Lernertext und der Normalisierung darstellen. Diese Phänomene sollen aber im Zentrum zukünftiger Forschung stehen.

Danksagung

Wir danken den drei anonymen Gutachtern ganz herzlich für ihre konstruktiven Kommentare. Den anderen Mitgliedern des Netzwerks Kobalt-DaF möchten wir ebenfalls danken, da ohne sie unsere Datengrundlage, das Kobalt-Korpus, nicht existieren würde. Felix Golcher hat uns wiederholt bei der Erstellung von R-Skripts geholfen und für uns die Grafik programmiert. Ihm gilt unser ganz besonderer Dank.

Literatur

Breckle, M. und Zinsmeister, H. (2012). A corpus-based contrastive analysis of local coherence in L1 and L2 German. In: Karabalić, V., Varga, M. und Pon, L. (Hgg.), *Discourse and Dialogue / Diskurs- und Dialog*, Seiten 235–250. Peter Lang Verlag, Frankfurt am Main [u.a.].

Chrupala, G., Dinu, G. und van Genabith, J. (2008). Learning morphology with Morfette. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Marrakesch, Marokko.

Corder, S. P. (Hg.) (1986). *Error Analysis and Interlanguage*. Oxford University Press, Oxford, 4. Auflage.

Díaz-Negrillo, A., Meurers, W., Valera, S. und Wunsch, H. (2010). Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154. Special Issue on Corpus Linguistics for Teaching and Learning, in Honour of John Sinclair.

Dickinson, M. und Ragheb, M. (2009). Dependency Annotation for Learner Corpora. In: Passarotti, M., Przepiórkowski, A., Raynaud, S. und van Eynde, F. (Hgg.), *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories*, Seiten 59–70, Mailand, Italien.

Dickinson, M. und Ragheb, M. (2013). Annotation for Learner English Guidelines: v. 0.1. Technischer Bericht, Indiana University, Bloomington, IN.

Eisenberg, P. (2004). *Das Wort: Grundriß der deutschen Grammatik*. Metzler, Stuttgart [u.a.].

Giesbrecht, E. und Evert, S. (2009). Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In: Alegria, I., Leturia, I. und Sharoff, S. (Hgg.), *Proceedings of the 5th Web as Corpus Workshop (WAC5)*.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In: Aijmer, K. (Hg.), *Languages in contrast*, Band 88 von *Lund studies in English*, Seiten 37–51. Lund University Press [u.a.], Lund.

Granger, S. (Hg.) (1998). *Learner English on Computer*. Studies in language and linguistics. Longman Publishers, London [u.a.].

Granger, S. (2008). Learner Corpora. In: Lüdeling, A. und Kytö, M. (Hgg.), *Corpus linguistics*, Band 1 von *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science*, Seiten 259–275. Mouton de Gruyter, Berlin und New York.

Granger, S. und Tyson, S. (1996). Connector usage in the English Essay Writing of Native and Non-native EFL Speakers of English. *World Englishes*, 15(1):17–27.

Helbig, G. und Buscha, J. (2007). *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Langenscheidt, Berlin [u.a.].

Hennig, M. und Bücker, J. (2008). Grammatik der gesprochenen Sprache in Theorie und Praxis. *Deutsch als Fremdsprache*, 45(2):115–116.

Hirschmann, H. (2013). *Modifikatoren im Deutschen*. Doktorarbeit, Humboldt-Universität zu Berlin, Berlin.

Hirschmann, H., Lüdeling, A., Rehbein, I., Reznicek, M. und Zeldes, A. (2013). Underuse of Syntactic Categories in Falko: A Case Study on Modification. In: Granger, S., Gilquin, G. und Meunier, F. (Hgg.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*, Seiten 223–234, Louvain-la-Neuve. Presses universitaires de Louvain.

Jurafsky, D. S. und Martin, J. H. (2009). *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Pearson Education Internat., Upper Saddle River, NJ, 2. Auflage.

Lafferty, J., McCallum, A. und Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *International Conference on Machine Learning (ICML)*.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. und Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45(2):67–73.

Rapp, R. (2007). Part-of-Speech Discovery by Clustering Contextual Features. In: Decker, R. und Lenz, H.-J. (Hgg.), *Advances in Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, Seiten 627–634. Springer, Berlin und Heidelberg.

Rehbein, I., Hirschmann, H., Lüdeling, A. und Reznicek, M. (2012). Better Tags Give Better Trees or do they? In: *Proceedings of Treebanks and Linguistic Theory (TLT-10)*.

Reznicek, M., Lüdeling, A. und Hirschmann, H. (2013). Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In: Ballier, N., Díaz Negrillo, A. und Thompson, P. (Hgg.), *Automatic Treatment and Analysis of Learner Corpus Data*, Band 59 von *Studies in Corpus Linguistics*, Seiten 101–124.

Schiller, A., Teufel, S., Stöckert, C. und Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technischer Bericht, Universitäten Stuttgart und Tübingen, `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf`.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, Großbritannien.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In: *Proceedings of the ACL SIGDAT-Workshop*.

Schmid, H., Fitschen, A. und Heid, U. (2004). SMOR:A German computational morphology covering derivation, composition and inflection. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*.

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, Großbritannien.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(3):209–231.

Toutanova, K. und Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In: *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Seiten 63–70, Hong Kong, China.

van Halteren, H., Daelemans, W. und Zavrel, J. (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics*, 27(2):199–229.

van Rooy, B. und Schäfer, L. (2002). The Effect of Learner Errors on POS Tag Errors during Automatic POS Tagging. *Southern African Linguistics & Applied Language Studies*, 20(4):325–335.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller, Klaus-Peter Wegera

HiTS: ein Tagset für historische Sprachstufen des Deutschen

1 Einleitung[1]

Im Kontext der Projekte zur Erstellung historischer Sprachkorpora „Referenzkorpus Altdeutsch“ und „Referenzkorpus Mittelhochdeutsch“ entstand ein Tagset für die Wortartanalyse, *HiTS* („Historisches Tagset“). Im Projekt „Referenzkorpus Frühneuhochdeutsch“ wird eine vereinfachte Version davon angewendet, da v.a. die späteren Zeiträume schon nahe am neuhochdeutschen (nhd.) Stand sind.[2]

HiTS orientiert sich am „Stuttgart-Tübingen Tagset“ (STTS, Schiller et al., 1999), dem Standardtagset für nhd. Korpora, und übernimmt — neben einer ganzen Reihe von Tags — auch das hierarchische Design der Tagnamen. Ursprünglich sollte das Tagset komplett auf STTS aufbauen und dieses lediglich um einige neue Tags erweitern. Es stellte sich jedoch heraus, dass neben einigen notwendigen feineren Unterscheidungen (z.B. bei den Pronominaladverbien) auch die Tagnamen des STTS nicht immer geeignet schienen. Z.B. sind in HiTS der definite und indefinite Artikel eine Unterklasse der Determinativa — die Sonderstellung des Artikels, die im STTS durch ein eigenes Tag ‘ART’ betont wird, ist eine neuere Entwicklung.

Das Tagset dient zur Annotation diachroner Daten. Einige wenige Tags finden dabei nur in den alt(hoch)deutschen Daten Anwendung. Um diachrone Untersuchungen bis in die jetzige Zeit zu ermöglichen, werden im vorliegenden Artikel den HiTS-Tags die jeweils entsprechenden STTS-Tags gegenübergestellt. Allerdings ist nicht in jedem Fall eine eindeutige Abbildung möglich. So wird z.B. in HiTS zwischen attributivisch und substantivisch verwendeten Zahlen unterschieden, nicht aber im STTS. Umgekehrt unterscheidet HiTS nicht zwischen elliptischen („kopflosen“) Nominalphrasen und substantivierten Adjektiven (s. Abschnitt 5), was das STTS aber tut.

[1] Die vorliegende Arbeit wurde finanziell unterstützt durch die Deutsche Forschungsgemeinschaft (DFG), DO 544/5-1/2 (Projektgruppe Altdeutsch), und KL 472/3-3/5, WE 1318/14-1/2, WE 1318/16-1, DI 1558/1-1/2 (Projektgruppe Mittelhochdeutsch). Außerdem möchten wir den anonymen Reviewern für ihre hilfreichen Kommentare danken.

[2] Das Referenzkorpus Altdeutsch (`http://www.deutschdiachrondigital.de`) deckt den Zeitraum 750–1050 ab, das Referenzkorpus Mittelhochdeutsch (`http://referenzkorpus-mhd.uni-bonn.de`, `http://www.rub.de/wegera/rem`) den Zeitraum 1050–1350 und das Referenzkorpus Frühneuhochdeutsch (`http://www.rub.de/wegera/ref`) den Zeitraum 1350–1650. Das Referenzkorpus Altdeutsch erfasst dabei die gesamte althochdeutsche und altniederdeutsche Textüberlieferung mit 650 000 Belegen. Die Referenzkorpora Mittel- und Frühneuhochdeutsch enthalten jeweils eine strukturierte Auswahl der Textüberlieferung mit etwa 2,1 Mio (mittelhochdeutsch) bzw. 4,4 Mio (frühneuhochdeutsch) annotierten Wortformen. Alle Korpora sind bzw. werden annotiert mit Metainformation sowie Information zum Lemma, zur Morphologie und zur Wortart.

Da die muttersprachliche Intuition für die älteren Sprachstufen grundsätzlich fehlt, können die Annotatoren z.B. keine linguistischen Tests durchführen. An die Stelle von Tests tritt daher in der historischen Sprachwissenschaft die Synopse und statistische Analyse möglichst vieler Belege entsprechender oder alternativer Konstruktionen. Bei den verbleibenden nicht auflösbaren ambigen Strukturen muss festgelegt werden, welche der möglichen Analysen annotiert werden soll. Prinzipiell wird dabei die historisch ältere bevorzugt. Beispielsweise wird bei Konstruktionen, die (noch) als pränominaler Genitiv oder (schon) als Kompositumserstglied analysiert werden könnten, die Genitivlesart annotiert (s. Abschnitt 2).

Den Wortartannotationen liegen als Basiseinheit Wörter zu Grunde. Der Bestimmung der Wortgrenzen („Tokenisierung“) ist ein eigener Abschnitt gewidmet (Abschnitt 2). Im STTS wird dieser Punkt nicht weiter thematisiert (nur der Sonderfall Mehrwortlexeme wird kurz angesprochen (Schiller et al., 1999, S. 9)). In modernen Korpora werden üblicherweise Wortgrenzen anhand von Leerzeichen bestimmt; zusätzlich zählen Satzzeichen als eigene Wörter. In historischen Daten finden sich jedoch zahlreiche Beispiele, bei denen Leerzeichen als Kriterium nicht genügen, um (nach heutigem Verständnis sinnvolle) Worteinheiten zu bestimmen. Die unterschiedlichen Schreibungen können wichtige Hinweise auf den Entwicklungsstand der Sprache liefern, z.B.: werden Komposita (noch) in zwei Wörtern oder (schon) in einem Wort geschrieben? Daher enthält HiTS auch ein Tagset für das Tagging von historischen Wortgrenzenauszeichnungen.

Eine weitere Besonderheit von HiTS ist die konsequente Unterscheidung von Lemma- vs. Beleg-spezifischer Annotation (Abschnitt 3). Beispielsweise kann ein Adjektiv (Lemma) in der Funktion eines Adverbs verwendet werden (Beleg); beide Wortarten werden in den Referenzkorpora annotiert. Diese Doppeltannotation eignet sich insbesondere für die Untersuchung von Sprachwandelprozessen, die mit einem Wortartwechsel einhergehen. Eine ähnliche Unterscheidung findet sich im STTS unter den „lexikalischen Kategorien“: z.B. wird hier angegeben, dass das Nomen *Alter* lexikalisch gesehen ein Adjektiv ist; in STTS-Darstellung: *der Alte*/NN<ADJ (Schiller et al., 1999, S. 13).

Für die Beispielsätze wurde für die althochdeutschen (ahd.) und altsächsischen (as.) Beispiele eine editionsnahe, für die mittelhochdeutschen (mhd.) Beispiele eine handschriftennahe Transkription gewählt.[3] Vereinzelt sind auch konstruierte Beispiele darunter; diese sind normalisiert wiedergegeben und ohne Quellenangabe. Handschriftennah ist z.B. die Form *vf* ‘auf’, die normalisiert als *ūf* wiedergegeben wird. Die Quellenangaben

[3] Die ahd. Beispiele werden nach den jeweiligen Editionen der Texte zitiert. Das Referenzkorpus Altdeutsch hat hierbei, wenn immer möglich, die handschriftengetreueste Ausgabe verwendet. Aus diesem Grund differieren die Schreibweisen in den ahd. Beispielen zum Teil sehr stark. Eine Anpassung der einzelnen Schreibungen an eine gewissermaßen künstlich konstruierte „normal-althochdeutsche“ Schreibung ist schon aus Gründen der Authentizität und der phonologischen dialektalen Differenzen, deren Abbildung wünschenswert ist, nicht sinnvoll.
Die mhd. Beispiele folgen in der Regel der Handschrift, insbesondere dort, wo sie dem MiGraKo (s.u.) entnommen sind; korpusexterne Beispiele werden dagegen meist nach normalisierenden Editionen zitiert.
Das MiGraKo ist das Korpus der entstehenden neuen Mhd. Grammatik; es entspricht in seiner Textkonstitution und Struktur dem „Bochumer Mittelhochdeutsch-Korpus“ und ist in Bonn komplett tokenisiert, lemmatisiert und morphologisch annotiert worden. Es wird einen separaten Teil des Referenzkorpus Mittelhochdeutsch bilden.

referieren auf Siglen, die im Anhang aufgelöst werden. Wir haben uns entschieden, weitgehend auf die Angabe von Sekundärliteratur zu verzichten, und verweisen stattdessen auf die Standardlexika und -wörterbücher.

Der Artikel ist wie folgt aufgebaut: Abschnitt 2 thematisiert die Wortgrenzenbestimmung, Abschnitt 3 die Unterscheidung Lemma- vs. Belegannotation. Die beiden nachfolgenden Abschnitte 4 und 5 beschreiben einzelne Wortarten im Detail: Determinative und Adjektive; bei diesen beiden Wortarten ergeben sich die meisten Unterschiede zwischen HiTS und STTS. Den Abschnitten (oder auch Unterabschnitten) geht jeweils eine Überblickstabelle voraus, die die Tags der vorgestellten Wortart auflistet. In Abschnitt 6 wird das Tagging lateinischer Passagen thematisiert. Im Anhang finden sich Überblickstabellen aller Tags in HiTS.

2 Token (Worteinheiten)

Beispiel		**HiTS**
original	**modernisiert**	
indaz	*in daz*	MS (*Multiverbierung mit Spatium*)
Schutz Gott	*Schutzgott*	US (*Univerbierung mit Spatium*)
Liebes=Ohnmachten	*Liebesohnmachten*	UH (*Univerbierung mit Hyphen*)
LandGraff	*Landgraff*	UB (*Univerbierung mit Binnenmajuskel*)

Token, d.h. die Einheiten, die der Wortartannotation zu Grunde liegen, werden in modernen Korpora weitgehend aufgrund von Leerzeichen (Spatien) bestimmt. Zusätzlich werden Satzzeichen als eigene Token aufgefasst, und häufig werden klitisierte Formen ebenfalls als eigenständige Token analysiert, so z.B. im TIGER-Korpus: *gibt 's*.[4]

In älteren Handschriften und Drucken gibt es einige wichtige Unterschiede zu modernen bzw. gegenwärtigen Texten, die bei der Tokenisierung zu beachten sind und die diese beeinflussen. Die Zusammen- und Getrenntschreibung sowie die Worttrennung am Zeilenende weichen vielfach von dem ab, was aus heutiger Sicht vornehmlich aus syntaktischen Gründen als Worteinheit und damit als Grundlage für die Tokenisierung zu betrachten ist. Interpunktionen bestehen nur in Ansätzen und unterscheiden sich hinsichtlich ihrer Funktion von moderner Interpunktion.

Die Wortgestalt unterliegt diachron verschiedenen Wandelprozessen. In HiTS gilt daher der Grundsatz, dass der handschriftliche Befund erkennbar bleiben soll. Daher unterscheiden wir zwei Ebenen der Textrepräsentation: eine diplomatische, d.h. handschriftennahe Ebene und eine Ebene mit moderner Tokenisierung. Im Folgenden beziehen wir uns auf die beiden Ebenen mit den Begriffen „Text-Ebene“ und „Token(isierungs)-Ebene“ und stellen die beiden Ebenen durch einen senkrechten Strich getrennt dar,

[4] TIGER-Korpus: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html`

gefolgt von einem „Token(isierungs)-Tag“, das die Art der Differenz zwischen den beiden Ebenen benennt:

(1) a. Schema: Text-Ebene | Token-Ebene | Token-Tag
b. Beispiel: *Schutz Gott* | *Schutzgott* | US

Wichtigstes Indiz für Getrenntschreibung ist die relative Größe eines Spatiums: Ist ein Spatium zwischen zwei Buchstabenfolgen in Relation deutlich größer als der „normale“ Abstand zwischen zwei Buchstaben, kann Getrenntschreibung angenommen werden. Hier kann es auch innerhalb derselben Handschrift oder auch desselben Drucks erhebliche Schwankungen geben, so dass sich oft nicht sicher entscheiden lässt, welches Spatium (schon) als Getrenntschreibung zu werten ist. Weitere Indizien können sein: Majuskeln am Wortanfang, doch können sie (bes. *J, N, R*) auch im Wortinneren vorkommen; ab der späteren mhd. Zeit steht außerdem meist <s> statt <ſ> (Schaft-s) am Wortende.

Die getrennt geschriebenen Einheiten entsprechen zwar mehrheitlich „Wörtern“ im modernen Verständnis, doch wird häufig davon abgewichen: Einerseits können auch Kompositionsglieder oder gebundene Morphe wie Präfixe getrennt geschrieben werden, z.B. *golt vaz, ge lebet, un verzaget* (‘Goldgefäß, gelebt, unverzagt’). Andererseits werden oft Wörter zusammengeschrieben, vor allem bei Pro- oder Enklise, z.B. *enwoltern = en wolte er in* (‘er wollte ihn nicht’); häufig ist Zusammenschreibung von Präpositionen mit dem folgenden Wort, z.B. in Nib *inir = in ir* ‘in ihr’, *indaz = in daz* ‘in das’, *inscharpfen = in scharpfen* ‘in scharfen’, *zeiungest = ze iungest* ‘jüngst/zuletzt’, *zeminne = ze minne* ‘zu(r) Minne’, *zeritter = ze ritter* ‘zu(m) Ritter’.

2.1 Univerbierung getrennt geschriebener Worteinheiten

Zusammen- und Getrenntschreibung von Worteinheiten, die heute univerbiert sind, schwanken stark bis in die Neuzeit hinein. Gelegentlich wird hier die Schreiberintention sichtbar. Dies betrifft insbesondere Komposita, bei denen die Entwicklung von einer Kontaktstellung (etwa in einer Genitivkonstruktion) über eine lose Verbindung (graphisch als Bindestrich oder Doppelbindestrich umgesetzt) oder durch Zusammenrückung schließlich zur Zusammenschreibung führt. Dabei können sich Binnenmajuskeln ergeben.

Auf der Token-Ebene werden die zusammengehörigen Wortteile gemäß heutiger Konventionen zusammengeschrieben. Die Tokenisierungstags (s. Tabelle S. 87) kennzeichnen die Zusammenfügungen mit einem ‘U’ (für „Univerbierung“) sowie mit weiteren Tags ‘S’ (durch „Spatium“ getrennte Wortbestandteile), ‘H’ (durch „Hyphen“ getrennt) und ‘B’ (durch „Binnenmajuskel“ markiert).

(2) a. Kontaktstellung (getrennte Schreibung): *Schutz Gott* | *Schutzgott* | US
b. (Doppel-)Bindestrich: *Liebes=Ohnmachten* | *Liebesohnmachten* | UH
c. Binnenmajuskel durch Zusammenrückung: *LandGraff* | *Landgraff* | UB

Zusammengeschriebene Einheiten, die bereits als Komposita angesehen werden, werden als solche übernommen. Bei Getrenntschreibung ist jedoch häufig kaum zu entscheiden, ob es sich um eine Genitivfügung oder ein Kompositum handelt. Lediglich bei eindeutiger Disambiguierung durch Substantivbegleiter kann zwischen Komposita und Genitivfügungen unterschieden werden, s. Bsp. (3a). Lässt der Substantivbegleiter nicht eindeutig erkennen, ob es sich um eine Genitivfügung oder um ein Kompositum handelt, so wird Getrenntschreibung beibehalten, s. (3b):[5]

(3) a. *die*[**Fem**] *herzen*[**Neut**] *küniginne*[**Fem**] | *herzenküniginne* | US
'die Herzenskönigin' (Kompositum)

b. *der*[**Fem/Masc**] *sunnen*[**Fem**] *schîn*[**Masc**] | |
'der Schein der Sonne/der Sonnenschein' (ambig; in HiTS analysiert als Genitivfügung)

2.2 Multiverbierung komplexer Worteinheiten

Vielfach finden sich Einheiten aufgrund eines individuellen Schreibstils („fehlerhaft") oder bedingt durch Klitisierung bis hin zur Krasis (Verschmelzung, s.u.), die eine Trennung in zwei oder mehrere Einheiten nahelegen. Solche Trennungen werden durch das Token-Tag 'MS' (für „Multiverbierung mit Spatium") markiert. (4b) zeigt ein Beispiel, das sowohl Uni- als auch Multiverbierung involviert. Hierfür werden zwei Token-Tags in der Reihenfolge ihres Auftretens verwendet.

(4) a. *indaz* | *in daz* | MS 'in das'
enwolde | *en wolde* | MS 'nicht wollte'
lebeter | *lebet er* | MS 'lebt er'

b. *be durfeter* | *bedurfet er* | US-MS 'bedürft ihr' (mhd., GRud βb 12)

Werden direkt aufeinanderfolgende Wörter nicht nur ohne Spatium geschrieben, sondern miteinander verschmolzen (Krasis), so gelten für die Trennung gesonderte Regeln. Beispielhaft seien hier zwei Fälle angeführt: zum einen die Verschmelzung der Verbendung der 2.Pers.Sg mit nachfolgendem Personalpronomen, bei der aufgrund der unterschiedlichen Verbstämme einmal nach und einmal vor dem Dental getrennt wird (5); zum andern die Verschmelzung von Präposition und nachfolgendem Artikel (6), bei der ausschlaggebend ist, ob der Vokal im Bereich der Krasis eindeutig (auch) zum zweiten Wort gehört, dem er dann zugeschlagen wird (6a).

(5) a. *wildu* | *wil du* | MS 'willst du'

b. *mahtu* | *maht u* | MS '(ver)magst du'

(6) a. *zen* | *z en* (< *ze den*) | MS 'zu den'

b. *zun* | *zu n* (< *zuo den*) | MS 'zu den'

[5] In diesen Fällen bleibt die Token-Ebene und das entsprechende Tokentag-Feld leer, da die modernen Wortgrenzen mit dem Original übereinstimmen.

2.3 Worttrennung am Zeilenende

Beispiel		HiTS
original	**modernisiert**	
in–↩ dem lande	*in dem lande*	MLH (*Multiverbierung am Zeilenende mit Hyphen*)
ou=↩ gen	*ougen*	ULH (*Univerbierung am Zeilenende mit Hyphen*)
ou↩ gen	*ougen*	ULS (*Univerbierung am Zeilenende mit Spatium*)

Die Worttrennung am Zeilenende kann in älteren Handschriften an jeder Stelle eines Wortes erfolgen (*ou-gen, Johan-nes, na-ch*), wenn sie auch überwiegend schon den modernen Trennregeln entspricht. Sie wird in den Handschriften und frühen Drucken entweder durch verschiedene Trennzeichen wie „–“ oder „=“ markiert oder bleibt unmarkiert; auch Doppelmarkierung am Ende einer und am Anfang der nächsten Zeile kommt vor.

In allen Fällen wird auf der Token-Ebene ein einziges Token angesetzt. Als Token-Tags dienen ‘ULH’ (für „Univerbierung am Zeilenende, mit Hyphen“) bzw. ‘ULS’ (für „Univerbierung am Zeilenende, mit Spatium (d.h. ohne Hyphen)“; der Zeilenumbruch wird im Folgenden durch „ ↩ “ repräsentiert.)

(7) a. *ou=↩ gen* | *ougen* | ULH

b. *ou↩ gen* | *ougen* | ULS

Bei „falsch“ gesetzten Trennzeichen in der Handschrift, mit denen über die Zeilengrenze hinweg Wörter als zusammengehörig gekennzeichnet werden, die (modern) nicht als ein Wort aufgefasst werden, werden auf der Token-Ebene mehrere Token angesetzt. Zur Markierung dient ‘MLH’ (für „Multiverbierung am Zeilenende mit Hyphen“; s. auch nächsten Abschnitt).

(8) *in–↩ dem lande* | *in dem lande* | MLH

2.4 Partikelverben (trennbare Präfixverben)

Im STTS werden separat stehende Partikeln von Partikelverben mit ‘PTKVZ’ („Partikel, Verbzusatz“) annotiert. Mit welchem Verb im Satz die Partikel zusammen eine Einheit bildet, wird nicht explizit markiert. Im Gegensatz dazu werden in HiTS beide Wortbestandteile als zusammengehörig ausgezeichnet.

Für historische Texte wie etwa mhd. werden die Grenzen für Partikelverben in HiTS enger gezogen als im STTS für das Nhd.: Nur Präpositionen und Adverbien können Ausgangspunkt für Partikelverben sein, nicht jedoch Adjektive oder Substantive. Dabei sind als Partikeln alle Adverbien möglich, zu denen es eine gleichlautende

Präposition gibt: insbesondere *abe, ane, bī, nāch, ūf, umbe, ūz, vor(e), vür(e), zuo* ('ab, an, bei, nach, auf, um, aus, vor, für, zu'), außerdem *īn* 'ein' (zur Präp. *in*), *nider* 'nieder' und *wider* 'zurück'. Wie im Nhd. können (Pronominal-)Adverbien dem Verb in Kontaktstellung unmittelbar vorangehen oder in Fernstellung dem Verb nachfolgen (s. auch Abschnitt 3.2).

Um die Partikelverben von festen Präfixbildungen zu unterscheiden und um der möglichen Fernstellung von Basisverb und Partikel Rechnung zu tragen, werden beide Bestandteile — auch bei Kontaktstellung oder Zusammenschreibung — als zwei Token analysiert (im Gegensatz zu Präfixverben). Für die Markierung der Zusammengehörigkeit von Partikel und Verb werden Indizes verwendet (hier als Subskripte auf der Token-Ebene dargestellt).

(9) a. *als er den brief* ***ane sach*** | *...* ***ane***$_1$ ***sach***$_1$
'als er den Brief ansah' (mhd., GrR βb 29)

b. *er* ***sach*** *den brief* ***ane*** | ***sach***$_1$ *...* ***ane***$_1$ |
'er sah den Brief an'

Das folgende Beispiel zeigt das Verb *zurückziehen* einmal als Partikel- und einmal als Präfixverb.

(10) a. *sī* ***zōch***$_1$ *in* ***wider***$_1$
'sie zog ihn zurück' (mhd., Iw 1484), Partikelverb

b. *hieten si dic reht erchant, si hieten* ***biderzogen*** *doh an dier ier hant*
'hätten sie dich recht erkannt, sie hätten ihre Hand auch von dir zurückgezogen' (mhd., AlexiusA 991f), Präfixverb

Da nicht wie im Nhd. in der Regel die Zusammen- bzw. Getrenntschreibung als Unterscheidungsmerkmal dienen kann, ergeben sich Abgrenzungsprobleme zwischen Verbpartikeln einerseits und Adverbien (etwa bei *nider*) oder Pronominaladverbien andererseits. Der Ansatz eines Pronominaladverbs und nicht eines Partikelverbs lässt sich oft darauf stützen, dass die Lesart mit Pronominaladverb kontextuell sinnvoll, die Lesart mit Partikelverb dagegen sinnwidrig ist:

(11) a. *ein teppet ...* ***da*** *sie* ***vf*** *solde sizzen*
'ein Teppich, darauf sie sitzen sollte' (mhd., GRud αb 16), mit dem Pronominaladverb *dār ūf*
und nicht: '*ein Teppich, wo sie aufsitzen sollte', mit dem Partikelverb *ūfsitzen* (auch mhd. meist vom Besteigen eines Pferdes)

b. *wilt vnde zam man da* ***vure***$_1$ ***truch***$_1$
'(Fleisch von) wilden und domestizierten/zahmen (Tieren) trug man da auf' (mhd., GRud A 19), mit dem Partikelverb *vüre tragen*
und nicht: '*dafür trug man Fleisch von wilden und zahmen Tieren'

2.5 Interpunktionszeichen

Die Interpunktion historischer Handschriften und Drucke folgt in der Regel (noch) nicht modernen satzlogischen Gesichtspunkten, sondern ist weitgehend — wenn auch keinesfalls ausschließlich — noch rhetorisch (nach dem sog. rhythmisch-intonatorischen Prinzip) organisiert.

Im Altdeutschen herrscht der auf antike Vorbilder zurückgehende Punkt vor; andere Zeichen sind dagegen selten. Die Punkte können relativ zur Zeile verschieden positioniert sein: *comma* (*subdistinctio* = tiefgestellter Punkt), *colon* (*media distinctio* = Punkt auf der mittleren Höhe der Zeile) und *periodus* (*distinctio* = hochgestellter Punkt). Dieses System wird im Mittelalter durch Striche und Linien, verbunden mit Punkten, erweitert und dient bis in die Frühneuzeit hinein zur Markierung von (Vor)Lesepausen bzw. Sprechpausen bei Sinneinheiten.

Im Spätmittelalter und in der Frühneuzeit setzt sich ein detaillierteres System von Interpunktionszeichen durch: neben dem Punkt und der Virgel („/“) findet sich zunehmend die Kleinform der Virgel, das Komma, das sich gegenüber der Langform zur Kennzeichnung von Teilsätzen zum Nhd. hin durchsetzt, zunehmend auch in Konkurrenz zum Semikolon (Strichpunkt).

Der Doppelpunkt mit ursprünglich (so bereits im 9. Jh.) allgemeiner (teil-)satztrennender Funktion erscheint häufiger erst im 16. Jh. Das Fragezeichen ist bereits im Ahd. bekannt, findet sich aber ebenfalls erst seit dem 16. Jh. häufiger. Das Ausrufezeichen ist in seiner heutigen Form seit dem 16. Jh. in Grammatiken erwähnt und im 17. Jh. zur Kennzeichnung von Ausrufen, Wünschen, Verwunderung und des Nachdrucks als „Ausrufungszeichen“ verwendet. Ebenfalls zunehmend finden sich die Klammer „( )“ zur Markierung von Parenthesen (Einschlusszeichen) und — seit Schottelius (1641) — der aus Poetiken entlehnte Apostroph (Hinterstrich) (vgl. dazu bes. Höchli, 1981).

Für Satzzeichen (in der Originalhandschrift bzw. -druck) wie *. : , / ? !* verwenden wir das Tag ‘$_’, für sonstige satzinterne Zeichen wie ‘(’ das Tag ‘$(’.

3 Lemma- vs. belegspezifische Annotation

In HiTS wird die Wortart einer jeden Wortform zweifach annotiert, und zwar zum einen mit Blick auf das Lemma und zum anderen mit Blick auf den konkreten Beleg, also der Verwendung einer Wortform in einem spezifischen Kontext. Natürlich sind Lemma- und Belegwortart in sehr vielen Fällen identisch. Sie können aber auch divergieren, z.B. wenn ein appellatives Nomen (NA)[6] als Eigenname (NE) verwendet wird: so etwa das bair. Appellativum *Hollerbirl* (‘Holunderbeere’) als Name eines Bioladens. Die Annotation von *Hollerbirl* nur als NA nach seiner Lemmawortart wäre in diesem Verwendungskontext definitiv irreführend. Und auch bei der Annotation nur nach der Belegwortart geht wertvolle Information verloren, nämlich die, dass es sich hier um keine Eigenschaft des Lemmas handelt.

[6] Im STTS: NN, „normales Nomen“.

Abgesehen davon ist die auf den ersten Blick sehr aufwändige doppelte Wortartannotation speziell für historische Korpora in besonderer Weise attraktiv, weil sie einen unmittelbaren Zugriff auf die Gebrauchsdynamik von Wörtern bietet und auf Sprachwandelprozesse, die mit einer Veränderung der Wortart einhergehen. Dabei werden Verwendungsbereiche und Zeiträume recherchierbar, in denen sich diese Veränderung vollzieht. Dazu im Folgenden zwei einschlägige Beispiele.

Die beiden POS-Ebenen werden dabei wie in (12b,c) gezeigt dargestellt. Steht nur eine Wortartkategorie hinter der Wortform, handelt es sich immer um die Belegwortart, vgl. (12a).

(12) a. Schema: Wortform / Beleg-POS

b. Schema: Wortform / Lemma-POS > Beleg-POS[7]

c. Beispiel: *Hollerbierl* / NA > NE

3.1 Entwicklung der Negationspartikel *nicht*

Die Negationspartikel nhd. *nicht* entwickelt sich aus der Akkusativform des negativen Indefinitpronomens bzw. Pronominalsubstantivs ahd. *niowiht*, das ursprünglich folgende vier Kasusformen aufweist: *niowiht* (Nom./Acc.), *niowihtes* (Gen.), *niowihti/e* (Dat.), *niowihtu* (Instrumental). Anders als das nhd. Indefinitpronomen *nichts* ist ahd. *niowiht* zunächst nicht selbst Träger der Satznegation, sondern die Kennzeichnung der Negation erfolgt in der Regel durch die Negationspartikel ahd. *ni*, die in frühen althochdeutschen Texten allem Anschein nach im negativen Satz obligatorisch ist (13). Entsprechend annotieren wir *ni* als Negationspartikel mit PTK > PTKNEG und *niowiht* als negatives Indefinitpronomen mit PI > PNEG.[8]

(13) *inti sie* ***ni*/PTK > PTKNEG** *quedent imo* ***niouuiht*/PI > PNEG:Sg.Acc**
'und sie nicht sagen ihm Nichts' = 'sie sagen ihm nichts' (ahd., T 104, 7)

Im späten Althochdeutschen wird die Partikel *ni* lautlich zu *ne* abgeschwächt. Dabei treten in Verbindung mit *ni* bzw. *ne* zunehmend adverbiale Verstärker auf, darunter der adverbial gebrauchte Akkusativ des negativen Indefinitpronomens *niowiht* > *niwiht* > *niht*. Letzteres findet sich in spätalthochdeutschen und frühmittelhochdeutschen Texten dann auch ohne *ne* in der Funktion als alleiniger Träger der Negation:

(14) *swer* ***niht*/PI > PTKNEG** *gloubet, der ist iu verteilet*
'wer nicht glaubt, der ist schon verurteilt' (ahd., BGB3)

[7] Das Zeichen > verwenden wir in zweifacher Bedeutung: In Verbindung mit Wortformen wie im Beispiel unten *niowiht* > *niwiht* > *niht* bedeutet es, dass sich die zweite Wortform aus der ersten diachron entwickelt hat. In Verbindung mit POS-Tags bezeichnet das erste POS-Tag die Kategorie des Lemma, das zweite die des Belegs. Wie im Text argumentiert, stellt die Beziehung zwischen der Lemma- und der Belegkategorie auch (oft) eine diachrone Entwicklung dar.

[8] PTK: Partikel, PTKNEG: Negationspartikel; PI: Pronomen, indefinit; PNEG: Pronomen, negativ.

Die Annotation trägt dem Rechnung, indem sie *niht* in dieser Verwendung auf der Belegebene mit dem POS-Tag PTKNEG versieht. Die Wortartzuweisung auf Lemmaebene (PI) bleibt unverändert mit dem Effekt, dass Belege wie diese diachron aus verschiedenen Perspektiven recherchierbar sind: aus der Perspektive von PI mit Blick auf die Entwicklung der negativen Indefinitpronomina des Deutschen ebenso wie aus der Perspektive von PTKNEG, d.h. unter dem Aspekt, wie der Wechsel von ahd. *ni* zu nhd. *nicht* von statten geht. Die Interpretation der Daten bleibt so auf ein Minimum beschränkt.

3.2 Partikelverben im Althochdeutschen

Das zweite Beispiel, das wir in diesem Zusammenhang diskutieren wollen, ist im Grundsatz vergleichbar, aber im Detail etwas anders gelagert. Auch hier handelt es sich um einen Sprachwandelprozess, der eine Veränderung des POS-Tagging (AVD[9] > PTKVZ) involviert. D.h. es geht um die Herausbildung von Partikelverben im Deutschen, bei deren Entstehung selbständige Morpheme — nämlich Adverbien — als Verbzusätze in das verbale Lexem integriert werden. Anders als im Falle der Entstehung der Negationspartikel *nicht* geht es hier allerdings nicht um einen Wandelprozess, der am Ende der althochdeutschen Zeit einsetzt, sondern um Veränderungen, die schon in althochdeutscher Zeit in vollem Gange sind. Jedenfalls verzeichnen die Referenzwörterbücher des Althochdeutschen (insbesondere Splett, 1993) bereits eine Vielzahl von Partikelverben, gleichzeitig aber immer auch noch die formgleichen Adverbien, die weiterhin zum Lexembestand des Althochdeutschen gehören. Dazu kommt eine weitere Erschwernis, die Annotationsentscheidungen in diesem Bereich problematisch macht, nämlich eine relativ hohe Zahl von Belegen, die nicht eindeutig interpretierbar sind, bzw. deren Interpretation umfangreiche formale und inhaltliche Analysen notwendig macht.

Wir bleiben hier nichtsdestotrotz bei der gleichen Grundstrategie und weisen alle betroffenen Wortformen auf der Lemmaebene durchgehend als AVD aus. Bei der Bestimmung der Wortart auf Belegebene orientieren wir uns hier im Sinne einer für Nutzer jederzeit nachvollziehbaren und transparenten Strategie an dem Referenzwörterbuch von Splett. D.h. wenn dieser — wie im Falle des folgenden Beispiels für das Verb ahd. *duon* einer Partikelverbvariante *uz-duon* (Splett, 1993, S. 1032) verzeichnet, wird *uz* in Belegen wie dem folgenden auf der Belegebene als PTKVZ ausgezeichnet:

[9] In HiTS lautet der Tag für Adverbien AVD (nicht ADV). Daneben gibt es interrogative (AVW), generalisierende (AVG, s. Abschnitt 4) und negative Adverbien (AVNEG).

(15)	AHD	***Uz***	*fon*	*iro*	*samanungu*
		aus	von	sie:gen.pl.masc	Versammlung
	Lemma-POS	**AVD**	AP	PPER	NA
	Beleg-POS	**PTKVZ**	APPR	DPOSGEN	NA
	AHD	*duont*	*sie*	*iuuuih*	
		tun	sie	euch	
	Lemma-POS	VV	PPER	PPER	
	Beleg-POS	VVFIN	PPER	PPER	

'Sie schließen euch aus ihrer Gemeinde aus' (ahd., T 171,3)

Das führt dazu, dass die hier auch wieder aus beiden Perspektiven (AVD) und (PTKVZ) möglichen Recherchen zu Ergebnissen mit einem relativ hohen Überlappungsbereich führen und nicht ohne weitere Analysen interpretierbar sind. Dies ist im Sinne einer Minimierung nicht reversibler und intransparenter interpretatorischer Vorentscheidungen durch die Annotatoren aber nicht zu vermeiden. In jedem Fall erhält der Recherchierende so eine zuverlässige und kalkulierbare Ausgangsbasis für seine Untersuchungen. Die interpretatorische Endentscheidung bleibt ihm hier aus guten Gründen im Detail selbst überlassen.

3.3 Synchrone Schwankungen

Neben der Darstellung von Sprachwandelprozessen in einem historischen Korpus ermöglicht die Trennung der POS-Annotation synchronen Erscheinungen gerecht zu werden, bei denen ein Lemma unter bestimmten Umständen die Funktion einer anderen Wortart als der ursprünglichen einnimmt. Die uns bekannten STTS-basierten Korpora bieten für derartige Fälle keine adäquate Recherchemöglichkeit.[10] Entweder würde bei einer lemmabezogenen Annotation ausschließlich die grundlegende Wortart eines Lemmas angegeben, und davon abweichende Verwendungsweisen müssten so ignoriert werden. Oder die POS-Kennzeichnung erfolgte grundsätzlich kontextbezogen, was bedeuten würde, dass u.U. keine klaren kategorialen Zuweisungen mehr möglich sind und zudem quantitative Recherchen verfälscht werden würden. Hingegen kann die Unterscheidung zwischen einer lemma- und einer belegspezifischen Annotation kontextabhängigen Schwankungen bei der Wortart gerecht werden.

Im Nhd. treten besonders in der mündlichen Sprache Konstruktionen auf, in denen ein Indefinitum durch ein Fragepronomen wiedergegeben wird, vgl. (16).

(16) *Ich habe **wen** gesehen.*

Hier entspricht das w-Pronomen semantisch einem Indefinitum. Diese Möglichkeit, indefinite Elemente durch w-Pronomina darzustellen, kennt bereits das Altdeutsche. In

[10] Wie in Abschnitt 1 erwähnt, sieht das STTS mit den „lexikalischen Kategorien" eine ähnliche Unterscheidung vor, die unseres Wissens aber aktuell von keinem Korpusprojekt genutzt wird.

HiTS-annotierten Korpora erhalten indefinit verwendete w-Pronomina das lemmabezogene Tag PW, als belegbezogenes Tag wird PI vergeben.[11]

(17) *ef man* ***huuemu*/PW > PI:Masc.Sg.Dat** *saca sokea* 'falls man mit jemandem Streit suche' (Hel 1521)

4 Determinativa, Pronomen und Pronominaladverbien

4.1 Artikel

Beispiel	HiTS	STTS
der *liehte tac* 'der helle Tag'	DD > DDA (ad.), DDART (mhd.) *Determinativ, definit, (artikelartig)*	ART *bestimmter Artikel*
ein *tîer* 'ein Tier'	DI > DIA (ad.), DIART (mhd.) *Determinativ, indefinit, (artikelartig)*	ART *unbestimmter Artikel*

Der definite Artikel hat sich seit germanischer Zeit aus dem Demonstrativpronomen, der indefinite Artikel erst einzelsprachlich aus dem Kardinalzahlwort *ein* entwickelt. Beide Prozesse sind im Altdeutschen noch im Gange, wobei der erste schon deutlich weiter gediehen ist als der zunächst noch in ersten Anfängen steckende zweite. Daher werden im altdeutschen Referenzkorpus *der* und *ein* noch als 'Determinativ, definit' (DDA) bzw. 'Determinativ, indefinit' (DIA) getaggt.

Im Mhd. wird dagegen regelmäßig DDART als Beleg-POS für den Definitartikel und DIART für den Indefinitartikel vergeben. Die Wortartabgrenzungen innerhalb der „Wortfamilie" *ein* (mit Verwendungen als Artikel, Indefinitpronomen, unbestimmtes Zahladjektiv oder Kardinalzahl, vgl. Dudenredaktion (2009, § 446)) bereiten hier allerdings noch größere Probleme als im Nhd.

(18) a. ***that*/DD > DDA** *helaga barn* 'das heilige Kind' (as., Hel 804)

b. ***der*/DD > DDART** *liehte tac* 'der helle Tag' (mhd., Iw 644)

(19) a. *tho gisaz er múader bi* ***einemo*/DI > DIA** *brunnen* 'da saß er müde bei einem Brunnen' (ahd., O II,14,8-9)

b. *Ave ist* ***ein*/DI > DIART** *tîer, daz heizit helphant* 'weiter gibt es ein Tier, das heißt Elefant' (mhd., Phys 138v,9 [13,11])

[11] PW: Pronomen, interrogativ; PI: Pronomen, indefinit.

4.2 Reflexiv verwendete Formen des anaphorischen Pronomens

Beispiel	HiTS	STTS
im(o) (masc) 'ihm'	PPER > PRF:3.Sg.Masc.Dat	*sich* PRF:3.Sg.Dat
im(o) (neut) 'ihm'	PPER > PRF:3.Sg.Neut.Dat	*sich* PRF:3.Sg.Dat
ir(u) (fem) 'ihr'	PPER > PRF:3.Sg.Fem.Dat	*sich* PRF:3.Sg.Dat
in (pl) 'ihnen'	PPER > PRF:3.Pl.*.Dat	*sich* PRF:3.Pl.Dat
	Prononem der 3. Person in reflexivem Gebrauch	*Reflexivpronomen*

Nur das Gotische und Altnordische verfügen noch über alle drei Formen des germ. Reflexivpronomens: got.: Gen. *seina*, Dat. *sis*, Acc. *sik*; an.: Gen. *sín*, Dat. *sér*, Acc. *sik*. Das Hochdeutsche hat dagegen lediglich die Akkusativform ahd. *sih*, mhd. *sich* in reflexiver Funktion bewahrt. Der Dativ **sir* < germ. **siz* ist untergegangen. Die Genitivform *sīn* des Reflexivs hat sich zwar formal erhalten, ist aber schon ahd. auch für den Gen.Sg.Masc *is* des anaphorischen Pronomens eingetreten und ersetzt im Mhd. zunehmend auch *es* Gen.Sg.Neut: *der eselære wart sin*/PPER:3.Sg.Masc.Gen *gewar* 'der Eseltreiber wurde seiner gewahr' (Kchr 1716); *wi wol ich dir sin*/PPER:3.Sg.Neut.Gen *gan!* 'wie sehr gönne ich es dir!' (Kchr 3186).

Das bedeutet, dass im Ahd. und Mhd. im Gen., vor allem aber im Dativ ausdrucksseitig zwischen reflexiver und nicht reflexiver Verwendung nicht unterschieden werden kann: *ir(e)* 'ihr' Dat.Sg.Fem, *im(e)* 'ihm' Dat.Sg.Masc/Neut und *in* 'ihnen' Dat.Pl können daher, wenn sie numeruskongruent und im Sg. auch genuskongruent mit dem Subjekt sind, neben der nicht-reflexiven auch eine reflexive Bedeutung haben; reflexiv ist z.B. *im* 'ihm/sich' in *er enwolde in vor* ***im*** *lāzen niht komen in den strît* 'er wollte ihn nicht vor **sich** in den Kampf kommen lassen' (Nib 2274,3), doch wäre in anderem Kontext auch die nicht-reflexive Lesart möglich: 'er wollte ihn nicht vor **ihm** [d.h. einem andern, dritten] in den Kampf lassen'. Nur sehr vereinzelt tritt in mhd. Zeit schon *sich* als Reflexivpronomen auch im Dativ ein und ersetzt die reflexiven *ir(e), im(e)* und *in*; vergleichsweise häufig geschieht dies allerdings mit bereits ahd. Anfängen nach der Präposition *ze* 'zu'.

Im HiTS erhalten ahd. *imo, iru, in* und mhd. *im, ir, in* in reflexiver Verwendung als allgemeines POS-Tag PPER, als POS-Tag des Belegs dagegen PRF im Unterschied zu PPER bei nicht-reflexivem Gebrauch.

(20) a. *than beginnid he* ***imu*****/PPER > PRF:3.Sg.Masc.Dat** *uuiti andreden*
'dann beginnt er, **sich** vor Strafe zu fürchten' (as., Hel 3495)

b. *si erwarp* ***ir*****/PPER > PRF:3.Sg.Fem.Dat** *lop vil grozen*
'sie erwarb **sich** sehr großen Ruhm' (mhd., Nib 1417,4)

4.3 Das reziproke Pronomen *einander*

Das Pronomen *einander* hat stets nur reziproke Bedeutung ('sich gegenseitig, wechselseitig'), während *sich* neben der reflexiven auch eine reziproke Lesart haben kann, z.B. *sie zogen sich die Stiefel aus*: 'sich selbst' bzw. 'sich gegenseitig'. Es erscheint dennoch angemessen, wenn im STTS *sich* und *einander* gleichermaßen als PRF:3.Pl... getaggt werden (Schiller et al., 1999, S. 35).

Im Alt- und Mittelhochdeutschen ist die Situation komplizierter, da das Pronomen *einander* wohl erst seit der späteren ahd. Zeit durch Zusammenrückung von *ein* und flektiertem *ander* entstanden ist. Vorläufer sind Fügungen wie *zimit íú ...thaz ein ándremo fúazi wasge* 'es geziemt sich für euch, dass einer dem anderen die Füße wasche' (O IV 11,50; s. AWB 1,476). Im Unterschied dazu ist mhd. (wie nhd.) *einander* zwar subjektbezüglich, kann jedoch nie zugleich auch das Subjekt bilden. Im Ahd. verbindet sich in aller Regel das endungslose *ein* mit dem flektierten *ander*. Wo *ein* nicht Subjekt ist wie im Otfrid-Beispiel (O), ist schwer entscheidbar, seit wann mit Univerbierung zu *einander-* zu rechnen ist; jedoch ist dies sicher noch nicht in den seltenen spätahd.–mhd. Belegen der Fall, in denen auch *ein* flektiert ist, z.B. *dar beualen Constantinis man Einen*/DIS:Dat.Pl *anderen*/DIS:Dat.Pl *die kint vnde uviph* 'da vertrauten Konstantins Männer einander ihre Kinder und Frauen an' (Roth 2655).

Auch wenn in den Handschriften meist getrennt geschrieben, werden *ein ander, ein andere, ein anderen* (mit unflektiertem *ein*) ansonsten als univerbiertes *einander* lemmatisiert und sie erhalten einheitlich PRF als POS-Tag. Anders als im Nhd. muss jedoch zwischen unflektiertem *einander*/PRF:3.Pl.* und flektiertem *einanderen*/PRF:3.Pl.Dat unterschieden werden.

(21) a. *Síu*/PPER:3.Pl.Neut.Nom *tîlegônt ouh* ***einánderíu***[12]**/PPER > PRF: 3.Pl.Neut.Acc** 'sie vernichten auch einander' (ahd., N Cat.77,8)

b. *dc erzentům ...dc die gůten* ***ainanderen***[13]**/PPER > PRF:3.Pl.Dat** *gebent* 'die Arznei ..., die die Guten einander geben' (mhd., TrHL 109r,17)

4.4 Possessiv verwendete Formen des anaphorischen Pronomens

Im Altdeutschen (wie schon im Germanischen) gab es für die 3. Person kein Possessivpronomen im Sg. des Femininums und im Pl. aller Genera. Stattdessen wurden die Genitivformen des anaphorischen Pronomens *ira* Gen.Sg.Fem bzw. *iro* Gen.Pl in possessivischer Funktion verwendet. Daran änderte sich zunächst auch nichts, als beide Formen in spätahd. Zeit in der Form des Gen.Pl *iro* zusammenfielen (wie auch sonst in der pronominalen Deklination) und *iro* später zu frühmhd. *ire* > *ir* reduziert wurde.

Im Laufe des 12. Jh. muss dieses *ire* (> *ir*) von der Genitivform des Personalpronomens zu einem zunächst unflektierbaren Possessivpronomen umgedeutet worden sein; diese Reanalyse war die Voraussetzung für die Neubildung flektierter Formen wie *iren* (Dat.Pl),

[12] *ein ánderíu* | *einánderíu* | US
[13] *ain anderen* | *ainanderen* | US

die vom 3. Viertel des 12. Jh. im Mittelfränkischen belegt sind: *iren swein sunnen* 'ihren zwei Söhnen', *irs lives* 'ihres Lebens' (Kölner Schreinskarten der Laurenzpfarre, 1159–1172); noch älter sind einige flektierte Belege im altniederländischen Leidener Willeram (um 1100), z.B. *uz heran lando* 'aus ihrem Land'.

Es sind also drei Entwicklungsphasen zu unterscheiden:

Phase 1: possessiv verwendeter Genitiv des anaphorischen Pronomens

Phase 2: ≫ indeklinables Possessivprononem

Phase 3: ≫ a) flektierbares ≫ b) flektiertes Possessivpronomen

Wie sind diese unterschiedlichen Verhältnisse jeweils angemessen zu taggen? Für die altdeutsche 1. Phase empfiehlt sich eine Regelung, die einerseits die unveränderte Zugehörigkeit zu PPER, anderseits aber auch die possessive Funktion berücksichtigt, also PPER als allgemeinen POS-Tag, DPOSGEN als POS-Tag des Belegs. Die Grenzziehung zwischen Phase 1 und 2 lässt sich nicht anhand klarer formaler Kriterien vornehmen: Vor allem spätahd.-frühmhd. Texte wie WNot oder Will können ebensogut der 1. wie der 2. Phase zugewiesen werden. So ist es eine letztlich willkürliche Entscheidung, gemäß der üblichen Epochengrenze alle bis zur Mitte des 11. Jh. datierbaren (altdeutschen) Texte der 1. Phase und alle danach datierbaren (mhd.) Texte der 2. bzw. 3. Phase zuzuweisen. Phase 2 und 3 unterscheiden sich nur noch in den morphologischen Merkmalen, z.B.:

Phase 1: *in* ***iro*/PPER** > **DPOSGEN:*.Gen.Pl.0** *lande*

Phase 2: *in* ***ir(e)*/PPER** > **DPOSA:Neut.Dat.Sg.0** *lande*
(oder: ***ir(e)*/PPER** > **DPOSA:*.*.*.0** *lande*)

Phase 3: *in* ***irem*/DPOS** > **DPOSA:Neut.Dat.Sg.st** *lande*

Der Unterschied zwischen den Phrasen 3a und 3b besteht darin, dass *ir* in 3b regelmäßig wie die anderen Possessivpronomen vom Typ *mīn* 'mein' flektiert wird, während in 3a-Texten flektierte und unflektierte Formen in weithin regellosem Wechsel vorkommen, so z.B. in den Predigtfragmenten PrF (Ende 12. Jh.) einerseits *ire gebetes* 'ihres Gebetes' (3,6), andererseits *ires gebetes* (4,19). Wo — wie in diesem Text — unflektiertes *ire* noch nicht zu *ir* verkürzt ist, lässt sich im Acc.Sg.Fem und Nom./Acc.Pl.Neut allenfalls einer Statistik des gesamten Formengebrauchs entnehmen, ob *ire* eher als flektiert oder als unflektiert anzusehen ist: *ire*/DPOSA:*.*.*.0 *afterkumelinge* 'ihre Nachkommen' (PrF 4,14) oder *ire*/DPOSA:Masc.Nom.Pl.st *afterkumelinge*.

4.5 Die generalisierenden Pronomen und Adverbien

Beispiel	HiTS	STTS
(s)waz 'was immer'	PG > PG *Relativpronomen, generalisierend*	PWS *subst. Int.pron.*
(s)wanne 'wann (immer)'	AVG > AVG *Relativadverb, generalisierend*	PWAV *adv. Int.- o. Rel.pron.*

Das STTS sieht vor (Schiller et al., 1999, S. 49, 51), dass *wer* in den beiden folgenden Beispielen gleichermaßen als substituierendes Interrogativpronomen (PWS) getaggt wird:

(22) a. ***wer*/PWS** *das sagt, weiß nicht, was los ist* (nhd.)

b. *er will wissen,* ***wer*/PWS** *mit welchem Zug kommt* (nhd.)

Aus sprachhistorischer Sicht und für die älteren Sprachstufen des Deutschen müssen beide Fälle jedoch kategorial unterschieden werden: Während beim indirekten Fragesatz in (22b) auch im Mhd./Ahd. das Interrogativpronomen *wer* vorausgeht, beruht *wer* im freien Relativsatz in (22a) auf dem „verallgemeinernden Relativpronomen" ahd. *sō hwer sō* bzw. mhd. *swer*:

Im Ahd. (wie im übrigen Westgermanischen) konnten verallgemeinernde Ausdrücke durch die Verbindung von Indefinitpronomina (*hwer* 'jemand', *hwelih* '(irgend)einer') oder Adverbien (*hwanne* 'irgendwann', *hwār* '(irgend)wo', *hwio* 'wie' etc.) mit zweifachem *sō* gebildet werden, z.B. *sō hwer sō* 'wenn jemand; wer auch immer', *sō hwār sō* 'wo auch immer'. Bereits in ahd. Zeit beginnt das zweite *sō* zu entfallen und im Mhd. ist es mit seltenen Ausnahmen verschwunden. Außerdem verschmilzt das zu *s* reduzierte erste *sō* im Großteil des Mhd. mit dem Folgewort, so dass sich die Formen *swer, swaz, swelich, swanne, swie* usw. ergeben. Nur im Mittelfränkischen bleibt *sō* erhalten, sofern es nicht ganz schwindet, z.B. *du bit mir, so wat du wolt* 'tu mit mir, was immer du willst' (RhMl 419). Im Laufe der mhd. Zeit beginnt das anlautende *s* der *sw*-Formen zu schwinden und dieser Schwund hat sich im Spätmhd. weitgehend durchgesetzt. Damit ist beim verallgemeinernden Prononem *wer* < *swer* im Grunde schon der nhd. Stand erreicht, denn die freie Relativsätze einleitenden nhd. *wer, was* stehen nicht nur formal in der direkten Nachfolge von mhd. *swer, swaz*, sie haben auch die generalisierende Bedeutung bewahrt, die durch optional beigefügtes *(auch) immer* nur hervorgehoben, aber nicht erst hergestellt wird; und dasselbe gilt teils auch für Relativadverbien, vgl. z.B. nhd. (23d) und den entsprechenden mhd. Satz in (23b) unten.

Zur Kennzeichnung der generalisierenden Bedeutung dieser Ausdrücke dient im HiTS das Tag-Element 'G': PG für die generalisierenden Pronomina *sō hwer sō, sō (h)welih sō* (ahd.), *swer, swelich* (mhd.), AVG für die generalisierenden Adverbien *sō (h)wanne sō, sō (h)wā sō* (ahd.) (23a), *swanne, swār, swar(e), swie* (mhd.) etc. (23b). Und dasselbe gilt konsequenterweise auch für die spätmhd. Formen mit *s*-Schwund *wer, welich, wanne*

usw. (23c). Ob und wann im Frühnhd. ggf. stattdessen dem STTS entsprechend 'PW…' als POS-Tag vergeben wird, bleibt noch zu klären.

(23) a. *sie tuont,* ***sōwazsō*/PG**[14] *sie wellen, inti* ***sōwārsō*/AVG, *sōwannesō*/AVG** *inti* ***sōwiesō*/AVG** *sie wellen* (ahd.)

b. *sie tuont,* ***swaz*/PG** *sie wellen, unde* ***swā*/AVG, *swanne*/AVG** *unde* ***swie*/AVG** *sie wellen* (mhd.)

c. *sie tuont,* ***waz*/PG** *sie wellen, unde* ***wā*/AVG, *wanne*/AVG** *unde* ***wie*/AVG** *sie wellen* (spätmhd.)

d. *sie tun,* ***was*/PWS** *(immer) sie wollen, und* ***wo*/PWAV** *(immer),* ***wann*/PWAV** *(immer) und* ***wie*/PWAV** *(immer) sie es wollen* (nhd.)

Ein Problem stellen die Formen dar, die zwischen ahd. *sō (h)wer sō* und mhd. *swęr* liegen: zum einen das im hochdeutschen Raum nahezu ausschließlich mittelfränkische *sowe*, z.B. *sowe*[15] *on beruret, he is selich* 'wer immer ihn berührt, er ist glücklich' (Rhein. Marienlob 224); obgleich meist getrennt geschrieben, wird *sowe* in HiTS lediglich als Variante von *swęr* aufgefasst und entsprechend lemmatisiert.

Zum anderen kann oberdeutsch wie mitteldeutsch bis zur ersten Hälfte des 13. Jh., danach nur noch vereinzelt das zweite *so* erhalten bleiben: *swer so*. In diesen Fällen taggen wir *so* als Partikel (PTK), z.B. *Swer*/PG *so*/PTK *wil sin der erste, der wirt der alleriungiste*[16] 'wer (immer) der Erste sein will, der wird der Allerletzte sein' (WMEv 32,2). Wo *sō* dagegen die adverbiale Bedeutung 'so, auf solche Weise' hat, wird es als Adverb (AVD) getaggt, z.B.: *wer*/PG *so*/AVD *horit, dastit, sihit, ruchit ader smackeit, …* 'wer auf diese Weise hört, tastet, sieht, riecht oder schmeckt' (Erlös 6926).

[14] *sō waz sō* | *sōwazsō* | US; *sō wārā sō* | *sōwārā* | US etc.
[15] *so we* | *sowe* | US
[16] *aller iungiste* | *alleriungiste* | US

4.6 Pronominaladverbien

Beispiel	HiTS (zweiteilig) pronominal	HiTS (zweiteilig) präpositional	STTS
dâ *pâgant siu* ***umpi*** 'darum streiten sie' (ahd.); ***dā*** *was vil volkes* ***inne*** 'darin waren viele Leute' (mhd.)	AVD > PAVD *Pron.adverb*	+ AP > PAVAP	PAV *Pron.adverb*
thár *wir* ***ana*** *lágun* 'woran wir erlagen' (ahd.); *der palas,* ***dā*** *...* ***inne*** *was* 'der Palast, in dem' (mhd.)	AVD > PAVREL *Pron.adverb, relativisch*	+ AP > PAVAP	—
wā ... umbe 'warum' (mhd.)	AVW > PAVW *Pron.adverb, interrogativ*	+ AP > PAVAP	PWAV *adv. Int.- o. Rel.pron.*
swā ... umbe 'warum (auch immer)' (mhd.)	AVG > PAVG *Pron.adverb, generalisierend*	+ AP > PAVAP	—

Im STTS werden Pronominaladverbien den Verhältnissen in der nhd. Standardsprache gemäß als stets univerbiert aufgefasst und als solche mit dem POS-Tag PAV versehen. In den älteren Sprachstufen findet sich dagegen nicht selten Distanzstellung des zweiten (präpositionalen) Teils, und zwar als Folge von Topikalisierung nur des ersten (pronominalen) Teils im Hauptsatz, z.B. *dā was vil volkes inne = vil volkes was dārinne*, und durch Nachstellung des zweiten Teils bei relativischer Verwendung des Pronominaladverbs, z.B. *der palas, dā vil volkes inne was*. Beim häufigsten Typ mit pronominalem *dā(r)* belaufen sich die Belege mit Distanzstellung im Korpus der mhd. Grammatik (MiGraKo)[17] auf immerhin 10,6% (= 783 Belege).

Pronominaladverbien werden daher in HiTS (mhd.) grundsätzlich als „zweiwortig" aufgefasst, auch wo sie in der Handschrift zusammengeschrieben sind: Ein handschriftliches *darinne* wird in *dar* und *inne* aufgespalten, die getrennt getaggt werden (s. Abschnitt 2). Entsprechend wird auch mit den relativischen, interrogativen und generalisierenden Pronominaladverbien verfahren:

(24) a. *dā*/AVD > PAVD *umbe*/AP > PAVAP
b. *dā*/AVD > PAVREL *umbe*/AP > PAVAP
c. *wā*/AVW > PAVW *umbe*/AP > PAVAP
d. *swā*/AVG > PAVG *umbe*/AP > PAVAP

[17] Vgl. Fußnote 3.

4.7 Interrogativadverbien

Beispiel	HiTS	STTS
(h)wār 'wo'	AVW > AVW *Interrogativadverb*	PWAV *Inter.pron., adverbial*
sō (h)wār sō 'wo auch immer'	AVG > AVG *Relativadverb, generalisierend*	PWAV *Rel.pron., adverbial*

Im STTS werden *wo, wann, wie* etc. nicht als Interrogativadverbien, sondern als adverbiale Interrogativ- bzw. Relativpronomen (PWAV) aufgefasst (Schiller et al., 1999, S. 53):

(25) a. ***Wo*/PWAV** *wohnt er?* (nhd.)
b. *er fragt,* ***wo*/PWAV** *er wohnt* (nhd.)
c. *der Ort,* ***wo*/PWAV** *er wohnt* (nhd.)

Schon hinsichtlich des Nhd. kann man sich aber fragen, warum *wo, wann* und *warum* Pronomen, die korrespondierenden *da, dann* und *darum* aber Adverb sein sollen (Schiller et al., 1999, S. 56). Für das Mhd. jedenfalls wird man diese Frage wohl verneinen müssen, denn hier gibt es ein — auch formal — sehr ausgeprägtes System von zusammengehörigen phorisch-deiktischen, interrogativen und generalisierenden Wörtern, die üblicherweise (z.B. Splett (1993); Lexer (1878)) als Adverbien klassifiziert werden. Dem schließt sich HiTS an:

Adverb (AVD)			Interrogativadverb (AVW)		
ahd	mhd		ahd	mhd	
dār	*dā(r)*	'da, dort'	*(h)wār*	*wā(r)*	'wo'
dara	*dar(e)*	'dorthin'	*(h)wara*	*war(e)*	'wohin'
danān	*dannen*	'von da weg'	*(h)wanān*	*wannen*	'woher'
dō	*(dō)*	'da(mals)'	*(h)wanne*	*wanne*	'wann'
sō	*(sō)*	'so'	*(h)wio*	*wie*	'wie'

generalisierendes Relativadverb (AVG)		
ahd	mhd	
sō (h)wār sō	*swā(r)*	'wo auch immer'
sō (h)wara sō	*swar(e)*	'wohin auch immer'
sō (h)wanān sō	*swannen*	'woher auch immer'
sō (h)wanne sō	*swanne*	'wann auch immer '
sō (h)wie sō	*swie*	'wie auch immer'

Noch weniger akzeptabel wäre es für die historischen Sprachstufen, wie im STTS zu den PWAV unterschiedslos auch die interrogativen Pronominaladverbien des Typs mhd.

wār umbe (nhd. *warum*) zu zählen. Da auch sie im Ahd. und Mhd. trennbar sind, müssen ihre beiden Teile entsprechend den deiktisch-determinativen Pronominaladverbien im Beleg das POS-Tag PAVW bzw. PAVAP erhalten (s. Abschnitt 4.6). Dasselbe gilt für die relativischen generalisierenden Pronominaladverbien des Typs *swār umbe*; als allgemeine POS-Tags fungieren also PAV bzw. AP, als POS-Tags des Belegs PAVG bzw. PAVAP.

5 Adjektive und Adjektivadverbien

5.1 Prädikative Adjektive und Adjektivadverbien

Beispiel	HiTS	STTS
daz almuoſen iſt ***reht*** 'das Almosen ist gerecht'	ADJ > ADJD *Adjektiv, prädikativ*	ADJA *präd. oder adv. Adjektiv*
lebet ***rehto*** 'lebt gerecht!'	ADJ > AVD *Adjektivadverb*	ADJD; ADV; PIS (s.u.)

Wie die Beispiele in der Tabelle[18] zeigen, werden prädikativ verwendete Adjektive im HiTS so wie im STTS bestimmt,[19] adverbial verwendete dagegen anders, nämlich in der Regel wie die prädikativen und nicht als Adjektivadverbien. Das ist für das Nhd. inzwischen üblich: Statt einer anderen Wortart wird eine weitere Verwendungsweise angenommen und neben die attributive und prädikative gestellt. Denn es ist kaum zu rechtfertigen, zwar zwischen der attributiven und prädikativen Form keinen Wortartwechsel anzunehmen, wohl aber zwischen der prädikativen und adverbialen, die fast immer gleich lauten (eine Ausnahme wird unten genannt). Früher jedoch wurde für die adverbiale Verwendung ein anderes Wort genommen als für die prädikative. Der Wortunterschied kam meist durch Derivation, ausnahmsweise durch Suppletion zustande. Das zweite Beispiel in der Tabelle zeigt die Derivation mit dem Adjektivadverbsuffix ahd. *-o* > mhd. *-e*. Ein Beispiel für Suppletion ist ahd. *wola* > mhd. *wol(e)* zum Adjektiv ahd., mhd. *guot*: *ſuîe ſer gilébet hábe* ***wôla***/ADJ > AVD *alder ûbelo*/ADJ > AVD 'wie er gelebt habe, gut oder schlecht' (frühmhd., BaGB 139,21).[20] Ein stärkerer Formunterschied zwischen prädikativ und adverbial verwendeten Adjektiven bestand auch insofern, als erstere im Ahd. und Mhd. gelegentlich noch mit Flexionsendung auftraten, vgl. *uuanda er ſuoz****er*** *unde reht****ir*** *iſt* 'dass er liebenswert und gerecht ist'

[18] WNot 6va und 22va. Die Belege in den Tabellen stammen mit einer Ausnahme alle aus dem frühmhd. WNot und enthalten alle den Stamm *reht* als Beispiel, damit die Beispiele möglichst vergleichbar sind und alt- wie mittelhochdeutsch annähernd gleich berücksichtigen.

[19] Allerdings wäre *er hält* ***geheim*** nach dem HiTS nicht als PTKVZ („adjektivische abgetrennte Verbzusätze", Schiller et al., 1999, S. 23) zu bestimmen, sondern als ADJD, was sich wohl aus einer anderen Tokenisierung ergibt.

[20] Beide Mittel, Derivation und Suppletion, waren aus dem Germanischen ererbt. Denn sie finden sich nicht nur im Westgermanischen (zu dem das Deutsche gehört), sondern auch im Nord- und Ostgermanischen, und zwar zum großen Teil mit ursprungsgleichen Zeichen ausgedrückt.

(frühmhd., WNot 40va), hier vielleicht von lat. *dulcis et rectus dominus* beeinflusst, was kurz vorher zitiert wird.

Es läge nun nahe zu sagen, die Adjektivadverbien seien nach dem HiTS deswegen anders als nach dem STTS zu bestimmen, weil sie im Ahd. und Mhd. noch durch eine eigene Wortbildungsweise von Adjektiven unterschieden wurden, im Nhd. nicht mehr. Aber wo man solch scharfe Grenzen zwischen Sprachstufen und Wortarten sucht, findet man in der Regel fließende Übergänge. Auch die Unterscheidungen zwischen Adjektiv und zugehörigem Adjektivadverb wurden allmählich, nicht plötzlich abgebaut: Noch heute kann das alte Adjektivadverb *lange* nur in adverbialer Verwendung, *lang* dagegen in adverbialer und prädikativer Verwendung gebraucht werden. Vgl. *das dauert ihr zu lang(e)*, kaum aber *das Kleid ist ihr zu lange*. Umgekehrt setzt schon in mhd. Zeit der Schwund des Adverbsuffixes *-e* ein, und zwar so allmählich, dass Formen mit und ohne *-e* in ein- und demselben Text, in selber Bedeutung und sogar in selber lautlicher Umgebung vorkommen:

(26) a. *d' quā* ***rechte***/ADJ > AVD *als eín helt gerant vf nampotenifen*
'der rannte **recht** als ein Held gegen Nampotenis an' (mhd., HTri 6236)

b. *h' triſtan lac ab'* ***recht***/ADJ > AVD *als eín ron*
'Tristan lag wieder **ganz** wie ein gefällter Stamm (da)' (mhd., HTri 3714)

Zudem gab es auch abgesehen vom Schwund des *-e* schon zu mhd. und sogar ahd. Zeit den Fall, dass Adverb- und Adjektivform ausdrucksseitig nicht unterschieden waren, vgl. ahd. *filu* > mhd. *vil(e)* 'sehr, viel, gänzlich', ein alter Acc.Sg.Neutr.

Im HiTS werden solche adverbial verwendeten Fälle, die wie im Nhd. mit prädikativen gleich lauteten, ebenfalls als Adverb angesetzt. Es wird also letztlich nicht nach der Wortbildungs-, sondern nach der Verwendungsweise unterschieden: Trägt die fragliche Einheit nichts zum Inhalt einer Nominalphrase, sondern einer Verbal-, Adjektiv- oder Adverbphrase bei, wird sie als Adjektivadverb bestimmt.

Auch im STTS gibt es Fälle von Formen, deren Stamm wie ein Adjektivstamm lautet, die aber als ADV bestimmt werden: wenn keine Kopulakonstruktion üblich ist oder ein erheblicher Bedeutungsunterschied gegenüber der Verwendung in der Kopulakonstruktion besteht (vgl. Schiller et al., 1999, S. 57).[21]

[21] Die Bindung an Beleglage, Verwendungsweise und an einen mehr oder minder großen Bedeutungsunterschied lässt absehen, dass die Unterscheidung nicht immer jedem nachvollziehbar ist, die Anzeichnung erschwert und ihr Ergebnis uneinheitlicher wird als nötig. Sogar schon anhand der ausgewählten und von Schiller et al. selbst analysierten Beispiele in den STTS-Richtlinien wird das deutlich. Zwei Beispiele:

1. In *Er ist heute früher gekommen* wird *früher* als ADJD bestimmt, *reichlich* in *er hat reichlich gelacht* dagegen als ADV (vgl. Schiller et al., 1999, S. 45, 57, 58), obwohl die Bedeutung kaum abweicht und auch eine Kopulakonstruktion nicht unüblich ist, vgl. *Das Essen war gut und reichlich.* In *Wir haben reichlich gegessen* wird es als PIS (substituierendes Pronomen) bestimmt (vgl. Schiller et al., 1999, S. 58), also vermutlich als Akkusativobjekt gedeutet. Man kann hier aber sicher eine Deutung als Adverbiale vorziehen, vgl. *Wir haben gut und reichlich gegessen*, wo *reichlich* wohl von gleicher Satzgliedart ist wie *gut*. Nach dem HiTS würden *früher* und *reichlich* in allen genannten Beispielen außer in der Kopulakonstruktion als AVD zu bestimmen sein.

5.2 Attributive Adjektive, vorangestellt und nachgestellt

Beispiel	HiTS	STTS
der ***reht*** *rihtari* 'der gerechte Richter'	ADJ > ADJA *Adjektiv, attributiv, vorangestellt*	ADJA *attr. Adjektiv*
die brugge ***reht*** 'die richtige Brücke'	ADJ > ADJN *Adjektiv, attributiv, nachgestellt*	ADJD (s.u.)

Die Beispiele in der Tabelle[22] zeigen zwei im Neuhochdeutschen nurmehr ausnahmsweise vorkommende Fälle: Endungslosigkeit bei vorangestelltem attributivem Adjektiv und Nachstellung des attributiven Adjektivs. Ersteres wird nicht über die Wortartbestimmung erfasst, die sich zwischen HiTS und STTS auch nicht unterscheidet. Zu letzterem Schiller et al. (1999, S. 18): „Mit ADJD werden prädikativ und adverbial (auch wenn andere Adjektive modifiziert werden) gebrauchte, sowie nachgestellte, nicht flektierte Adjektive bezeichnet." Das hieße, die nachgestellten, endungslosen attributiven Adjektive nicht von prädikativen (und nach dem STTS auch nicht von adverbialen) zu unterscheiden, wohl aber von nachgestellten attributiven Adjektiven mit Endung (die allerdings nur sehr selten vorkamen).

5.3 Substantivierte Adjektive und Substantive

Beispiel	HiTS	STTS
die ***rehten*** *ſkinent* 'die gerechten/Gerechten strahlen'	ADJ > ADJS *Adjektiv, substituierend*	ADJA; NN *attr. Adj.; normales Nomen*
daz opfer deſ ***rehteſ*** 'das Opfer des Rechtes'	NA *Nomen, appellativ*	NN *normales Nomen*

Wie die Beispiele in der Tabelle[23] mit denen des vorigen Abschnittes zeigen, werden im HiTS substantivierte Adjektive sowohl von attributiven Adjektiven als auch von Substantiven unterschieden. Von ersteren dadurch, dass sie der Kopf ihrer Nominalphrase sind. Das wird auch dort angenommen, wo ein Substantiv als Bezugswort mehr oder minder leicht ergänzt werden könnte — „mehr oder minder leicht" deutet den Grund an:

2. Zu *recht* und anderen heißt es „diese Wortformen sind niemals ADJD, weil sie keine Kopulakonstruktion bilden können" (Schiller et al., 1999, S. 57). Diese Begründung gilt für *recht* aber nicht, vgl. *Das ist mir sehr recht*; hier wäre vielleicht das Kriterium eines Bedeutungsunterschiedes heranzuziehen, das aber bei den meisten anderen Beispielen für dieselbe Unterscheidung zwischen ADV und ADJA nicht wirkt (vgl. *ich habe ihn kürzlich*/ADV *gesehen* vs. *der Anlaß meines kürzlichen*/ADJA *Besuches*) und auch erst bei einer anderen Unterscheidung, der zwischen ADV und ADJD, angeführt wird.

[22] WNot 10rb, Rapp 17582
[23] WNot 21va und 6va

Die Unterscheidung wäre oft Spekulation, daher zeitaufwändig und für die Auswertung unergiebig. Ein kaum entscheidbarer Fall ist etwa der, dass in einer Aufzählung bald ein substantivisches Bezugswort beim Adjektiv steht, bald nur das Adjektiv:

(27) *Er ín hat noch nít vernv̄men waz blínden lāmen ſtv̄men Maladen dode lvde vnde waz vns daz bedvde*
'Ihr habt noch nicht vernommen, was Blinde, Lahme, Stumme, Aussätzige, tote Menschen (sind), und was uns das bedeute (was ihre übertragene Bedeutung ist)' (PrRei 18b,7)

Hier wäre man wohl eher geneigt, sich für leichte Ergänzbarkeit des Bezugswortes und damit gegen Substantiviertheit der Adjektive zu entscheiden. Das zeigt aber, wie unsicher solche Einschätzungen sind: Denn im westmitteldt. Dialekt dieses Textes sind die (gegen die Formregel gesetzten) schwachen Endungen der Adjektive *blínd**en**, lām**en**, ſtv̄m**en**, Malad**en*** ein starkes Zeichen dafür, dass es sich um Substantivierungen handelt, und entsprechend hat nur das offensichtlich nicht substantivierte Adjektiv in *dod**e** lvde* die starke Endung.

Was die Unterscheidung zwischen Substantiven und substantivierten Adjektiven angeht, gibt es einige nicht ganz selten vorkommende Stämme wie *reht*, *guot*, *vil*, die als Adjektiv und als Substantiv vorkommen können. Sie sind in vielen Kasus durch die Endung unterschieden, wie das zweite Beispiel der Tabelle für den Gen.Sg.Neutr zeigt, in anderen Kasus aber nicht, vgl. *daz iſt **recht*** 'das ist recht / (das) Recht' (Rupr 96,13). Dann bleiben unsicherere Anhaltspunkte: die Syntax (kongruierendes Attribut oder Bezugswort mit Genitivattribut) und die Semantik ('gut, Gutes' oder 'Gutes; Besitz, Güte, gute Absicht').

Im STTS sind substantivierte Adjektive entweder als ADJA oder, und zwar bei Großschreibung, als NN zu bestimmen (vgl. Schiller et al., 1999, S. 18f). Die Bindung an die Schreibung macht Spekulation unnötig, ist aber zumal für ältere Sprachstufen nicht sinnvoll, vielleicht auch für jüngere zu sehr von Außersprachlichem (wie der gerade gültigen Rechtschreibung) abhängig. — Ein Sonderfall: Wie üblich wird nach dem STTS in einem Fall wie *die **Schweizer** Schokolade* ein ADJ angesetzt (vgl. Schiller et al., 1999, S. 19); auf älteren Sprachstufen hat man hier ein Substantiv im Genitiv Plural anzusetzen, worauf Endung und Großschreibung noch im Nhd. hindeuten.

6 Lateinische Passagen

Die frühe deutsche Schriftlichkeit entwickelte sich in direkter Abhängigkeit von den lateinischen christlichen und theologisch-philosophischen Texten, d.h. ein Großteil der altdeutschen Literatur ist mehr oder weniger direkt vom Lateinischen beeinflusst. Diese Beeinflussung wirkt sich über das Inhaltliche hinaus z.B. auf syntaktische Konstruktionen oder auf graphische Repräsentationsformen aus. Ein Teil der lateinabhängigen ahd. Schriften ist in textueller Verquickung mit der lateinischen Vorlage überliefert, es handelt sich dabei u.a. um Interlinearglossierungen (z.B. B), Interlinearübersetzungen (z.B. T) oder um lat.-deutsche Mischtexte (z.B. DH).

(28) zeigt ein Beispiel für eine zeilengemäße lat.-ahd. Interlinearübersetzung. Das Beispiel illustriert, dass die ahd. Interlinearübersetzung immer wieder vom lat. Original abweicht, so z.B. in der zweiten Zeile, wo im Lateinischen *circumcideretur* 'beschnitten wurde' vor seinem Subjekt *puer* 'Kind' steht, während in der ahd. Übersetzung die umgekehrte Reihenfolge gewählt wurde. Die Zeilenumbrüche im Beispiel entsprechen dem Original der Handschrift, die Zuordnung der entsprechenden lat. und ahd. Konstituenten erfolgt nur der besseren Lesbarkeit wegen.

(28)

Lat. Z1	*et*	*postquam*				*consummati*	*sunt*	
Lat.-POS	KON	KOUS				VVPP	VAFIN	
Ahd. Z1		*after*	*thiu*	*tho*		*argangana*	*uuarun*	
Ahd.-POS		KOUS	DDA	AVD		VVPP	VAFIN	
Lat. Z2	*dies*	*octo*	*ut*	*circumcideretur*		*puer*		
Lat.-POS	NA	CARDN	KOUS	VVFIN		NA		
Ahd. Z2	*ahtu*	*taga*	*thaz*			*thaz*	*kind*	
Ahd.-POS	CARDA	NA	KOUS			DDA	NA	
Lat. Z3			*vocatum*	*est*	*nomen*	*eius*	*Ihesus*	
Lat.-POS			VVPP	VAFIN	NA	PPER	NE	
Ahd. Z3	*bisnitan vvurdi*		*uuard*	*imo*	*genemnit*	*namo*	*Heilant*	
Ahd.-POS	VVPP VAFIN		VAFIN	PPER	VVPP	NA	NA	

'Nachdem acht Tage vergangen waren, dass das Kind beschnitten wurde, wurde er Jesus/Heiland genannt' (ahd, T 7,1)

Die ahd. Paralleltexte werden gemeinsam mit ihrer lateinischen Vorlage in das Referenzkorpus Altdeutsch aufgenommen. Die lateinischen Texte werden dabei ebenfalls voll annotiert, so dass ohne weiteres lexikalische und grammatikalische Übersetzungsstrategien abrufbar sind. Die Übernahme des POS-Tagsets erfolgt dabei weitestgehend problemlos, lediglich die Hinzunahme eines Tags für das Gerundium ('VVINFG') bzw. Gerundivum ('VVPG') war notwendig.

(29) a.

Lat.	*Et*	*cum*	*stabitis*	*ad*	***orandum***
	und	wenn	stehen	zu	beten
Lat.-POS	KON	KOUS	VVFIN	APPR	**VVINFG**

'Und wenn ihr steht und betet' (ahd., T 121,4)

b.

Lat.	*filius*	*hominis*	***tradendus***	*est*	*in*
	Sohn	Mensch	ausliefern	sein	in
Lat.-POS	NA	NA	**VVPG**	VAFIN	APPR
Lat.	*manus*	*hominum*			
	Hand	Mensch			
Lat.-POS	NA	NA			

'der Menschensohn wird in die Hände der Menschen ausgeliefert' (ahd., T 93,1)

7 Schluss

In diesem Artikel wurde HiTS vorgestellt, ein Tagset für die Annotation historischer Sprachstufen des Deutschen.

Der Artikel legte den Schwerpunkt auf wichtige Unterschiede zum STTS. Dies beginnt bei der Tokenisierung der Daten, für die HiTS eigene Richtlinien und Tags bereitstellt, da die Verwendung von Spatien in historischen Sprachdaten ungleich inkonsistenter ist als in modernen Sprachdaten. Oft interagiert die Bestimmung der Wortgrenzen mit der entsprechenden grammatischen Analyse, so z.B. bei der Analyse komplexer Nomensequenzen als Genitivkonstruktion vs. Komposita.

Ein weiterer auffallender Unterschied ist die konsequente Unterscheidung zwischen der Kategorie des Lemmas und des konkreten Belegs. Diese Art der Doppelauszeichnung eignet sich insbesondere für die Untersuchung von Sprachwandelprozessen, die mit einem Wortartwechsel einhergehen, wie z.B. im Fall von Verbpartikeln, die sich aus Adverbien herausbilden.

Verglichen mit modernen Daten ergeben sich überdies in den historischen Daten weitaus häufiger Ambiguitäten, die auch nicht im Kontext auflösbar sind. Der Hauptgrund dafür ist, dass wir heutigen Leser und Annotatoren keine muttersprachlichen Intuitionen haben und deshalb z.B. sprachliche Umformungstests nicht zur Disambiguierung genutzt werden können. In solchen Fällen wird die historisch ältere Version angenommen (so sinnvoll bestimmbar). Beispielsweise wird *der sunnen schîn* 'der Sonne Schein' als Genitivkonstruktion analysiert und nicht als Kompositum. Damit wird erreicht, dass Sprachwandelprozesse nicht zu früh angesetzt werden können, sondern erst ab dem Zeitpunkt, zu dem eindeutige Evidenz vorliegt.

Man könnte argumentieren, dass es zu bevorzugen wäre, ambige Fälle als solche zu markieren. Dann könnte z.B. im Falle der Komposita anhand der Korpora untersucht werden, ob ab einem gewissen Zeitpunkt ambige Fälle gehäuft auftreten, bis dann die ersten eindeutigen Kompositakonstruktionen in Erscheinung treten.

Zwei Punkte sprechen gegen ein solches Vorgehen: Zum einen würde dadurch die Annotationsaufgabe deutlich komplexer und die resultierenden Annotationen (häufig) unzuverlässiger, so dass die Korpusnutzer sich letztlich nicht darauf verlassen könnten, dass tatsächlich *alle* ambigen Fälle als solche annotiert sind. Zum anderen erfordert eine ambige Annotation eine deutlich aufwändigere Art der Repräsentation, da sich Ambiguitäten über mehrere Annotationsebenen erstrecken können. Im schon genannten Beispiel würde in der Kompositalesart die Sequenz *sunnen schîn* beispielsweise als ein Token analysiert, so dass folglich nur ein POS-Tag und nur eine morphologische Annotation vergeben würde (30b).[24] In der Genitivanalyse hingegen handelt es sich um zwei Token mit ihren entsprechenden Annotationen (30a). Außerdem wird der Artikel *der* einmal als feminin und einmal als maskulin analysiert. Es ist nicht offensichtlich, wie man diese Informationen innerhalb einer Annotation quer über die Ebenen als zusammengehörig auszeichnen könnte.

[24] Die Analyse in (30b) entspricht nicht den HiTS-Richtlinien. (30a) zeigt die HiTS-konforme Annotation.

(30) a.

MHD	*der*	*sunnen*	*schîn*	(nach HiTS)
POS	DD > DDART	NA > NA	NA > NA	
MORPH	Fem.Gen.Sg	Fem.Gen.Sg	Masc.Nom.Sg	

b.

MHD	*der*	*sunnen schîn*
TOK	*der*	*sunnenschîn*
TOKTAG		US
POS	DD > DDART	NA > NA
MORPH	Masc.Nom.Sg	Masc.Nom.Sg

Will man die Entwicklung von Komposita im Deutschen untersuchen, so bietet ein nach HiTS annotiertes Korpus dennoch nützliche Information, vorausgesetzt, es enthält auch morphologische Information (die Referenzkorpora des Altdeutschen und Mittelhochdeutschen sind morphologisch annotiert):

Eindeutige Kompositakonstruktionen wie in (31), 'die Herzenskönigin', sind z.B. auffindbar durch eine Suche nach Wortformen mit der POS-Annotation „NA > NA" und dem Token-Tag „US"; (31) wäre hierfür ein Treffer. Potenziell ambige Konstruktionen können durch eine Suche nach einem Artikel und (Adjektiv und) Nomen im Genitiv, direkt gefolgt von einem weiteren Nomen abgefragt werden; hier wäre (30a) ein Treffer. Die Treffer müssen dann manuell daraufhin gesichtet werden, ob der Artikel sich alternativ auch auf das zweite Nomen beziehen könnte. D.h. die Annotation nach HiTS kombiniert mit morphologischer Annotation liefert immerhin genug Information, um sehr gezielt nach ambigen Kompositakandidaten zu suchen.

(31)

MHD	*die*	*herzen küniginne*	(nach HiTS)
TOK	*die*	*herzenküniginne*	
TOKTAG		US	
POS	DD > DDART	NA > NA	
MORPH	Fem.Nom.Sg	Fem.Nom.Sg	

Wünschenswert wäre natürlich, dass die Korpora von solchen Untersuchungen profitieren könnten, also dass beispielsweise derjenige, der die Kompositakandidaten einzeln überprüft, das Resultat seiner Überprüfung („ja, ambig" oder „nein, eindeutig Genitivkonstruktion") zum Korpus hinzufügen könnte. Die aktuellen Korpusarchitekturen unterstützen solche Wiki-ähnlichen Beitragsmöglichkeiten allerdings noch nicht.

Quellen

Altdeutsch

B	*Benediktinerregel*. — Edition: E. Steinmeyer (Hg.). Die kleineren althochdeutschen Denkmäler. Dublin/Zürich 1971 (Nachdruck von 1916). 190–289.
BGB3	*Benediktbeurer Glauben und Beichte 3*. — Edition: E. Steinmeyer (Hg.). Die kleineren althochdeutschen Denkmäler. Dublin/Zürich 1971 (Nachdruck von 1916). 357–362.
DH	*De Heinrico*. — Edition: E. Steinmeyer (Hg.). Die kleineren althochdeutschen Denkmäler. Dublin/Zürich 1971 (Nachdruck von 1916). 110–114.
FP2	*Freisinger Paternoster 2*. — Edition: E. Steinmeyer (Hg.). Die kleineren althochdeutschen Denkmäler. Dublin/Zürich 1971 (Nachdruck von 1916). 43–48.
Hel	*Heliand*. — Edition: B. Taeger (Hg.). Heliand. Tübingen 1984.
I	*Isidor*. — Edition: H. Eggers (Hg.). Isidor. Tübingen 1964.
M	*Muspilli*. — Edition: E. Steinmeyer (Hg.). Die kleineren althochdeutschen Denkmäler. Dublin/Zürich 1971 (Nachdruck von 1916). 66–81.
O	*Otfrid: 'Evangelienbuch'*. — Edition: O. Erdmann (Hg.). Otfrids Evangelienbuch. Tübingen ^{6}1973.
T	*Tatian*. — Edition: E. Sievers (Hg.). Tatian. Paderborn 1966 (Nachdruck der zweiten Ausgabe von 1892).
WK	*Weissenburger Katechismus*. — Edition: E. Steinmeyer (Hg.). Die kleineren althochdeutschen Denkmäler. Dublin/Zürich 1971 (Nachdruck von 1916). 29–38.

Mittelhochdeutsch

AlexiusA	*Alexius A*. Handschrift: Graz, Universitätsbibl., Ms. 1501, Bl. 70–134. — Edition: H. F. Maßmann (Hg.). Sanct Alexius Leben in acht gereimten mittelhochdeutschen Behandlungen. Quedlinburg/Leipzig 1843. 45–67.
ArnoltSieb	*Priester Arnold: 'Von der Siebenzahl'*. Handschrift: Vorau, Stiftsbibl., Cod. 276, Bl. 129vb–133vb. — Edition: F. Maurer (Hg.). Die religiösen Dichtungen des 11. und 12. Jahrhunderts. Bd. III. Tübingen 1970. 53–85.
BaGB	*Bamberger Glaube und Beichte*. Handschrift: München, BSB, Cgm 4460. — Edition: E. von Steinmeyer (Hg.). Die kleineren althochdeutschen Sprachdenkmäler. Berlin 1916. Nr. 28B.
Erlös	*Die Erlösung*. Handschrift: Krakau, Bibl. Jagiellońska, Berol. mgq 1412; Laubach, Graf zu Solms–Laubach'sche Bibl., Fragm. T. — Edition: F. Maurer (Hg.). Die Erlösung. Eine geistliche Dichtung des 14. Jahrhunderts. Leipzig 1934. 302–308.
GRud	*Graf Rudolf*. Handschrift: Braunschweig, Stadtbibl., Fragm. 36; Göttingen, Staats- und Universitätsbibl., 4° Cod. Ms. philol. 184:VII. — Edition: C. von Kraus (Hg.). Mittelhochdeutsches Übungsbuch, Heidelberg 1912. 54–71.

HTri	*Heinrich von Freiberg: 'Tristan'*. Handschrift: Florenz, Nationalbibl., Cod. B.R. 226, Bl. 103ra–139vb. — Edition: D. Buschinger (Hg.). Heinrich von Freiberg, Tristan. Göppingen 1982.
Iw	*Hartmann von Aue: 'Iwein'*. Handschrift: Gießen, Universitätsbibl., Hs. 97. — Edition: G. F. Benecke u. K. Lachmann (Hg.). Iwein. Eine Erzählung von Hartmann von Aue. Neu bearb. von L. Wolff. 7. Ausg. Bd. 1: Text. Berlin 1968.
Kchr	*Kaiserchronik A (V)*. Handschrift: Vorau, Stiftsbibl., Cod. 276, Bl. 1ra–73vb. — Edition: E. Schröder (Hg.). Kaiserchronik eines Regensburger Geistlichen (MGH Deutsche Chroniken I,1). Berlin 1895.
Nib	*Nibelungenlied*. Handschrift: Karlsruhe, LB, Cod. Donaueschingen 63, fol. 1r–89r (= Nibelungenlied C). — Edition: U. Hennig (Hg.). Das Nibelungenlied nach der Handschrift C (Altdeutsche Textbibliothek 83). Tübingen 1977.
Phys	*Wiener (Jüngerer) Physiologus*. Handschrift: Wien, Österr. Nationalbibl., Cod. 2721, Bl. 130r–158v. — Edition: F. Wilhelm (Hg.). Denkmäler deutscher Prosa des 11. und 12. Jahrhunderts, Abteilung A: Text. München 1914/16. 5–28.
PrF	*Frankfurter Predigtfragmente*. Handschrift: Frankfurt, SUB, Fragm. germ. I 1. — Edition: L. Diefenbach. Mitteldeutsche Predigtbruchstücke. In: Germania 19 (1874). 305–314.
PrRei	*Hessische Reimpredigten*. Handschrift: Hamburg, SUB, Cod. 99 in scrin. 12–312. — Edition: B. Lenz-Kemper (Hg.). Die Hessischen Reimpredigten. Bd. 2: Text. Berlin 2009.
Rapp	*Rappoltsteiner Parzifal*. Handschrift: Karlsruhe, LB, Codex Donaueschingen 97 (Parzival Gd, Hs. D der frz. Perceval-Forschung). — Edition: K. Schorbach (Hg.). Parzifal von Claus Wisse und Philipp Colin (1331–1336). Eine Ergänzung der Dichtung Wolframs von Eschenbach. Straßburg/London 1888.
RhMl	*Rheinisches Marienlob*. Handschrift: Hannover, Landesbibl., Ms. I 81, Bl. 1r–93v. — Edition: A. Bach (Hg.). Das Rheinische Marienlob. Eine deutsche Dichtung des 13. Jahrhunderts. Leipzig 1934.
Roth	*König Rother (H)*. Handschrift: Heidelberg, Universitätsbibl., Cpg 390. — Edition: Th. Frings u. J. Kuhnt (Hg.). König Rother. Bonn/Leipzig 1922.
Rupr	*Ruprecht von Freising: 'Rechtsbuch'*. Handschrift: München, StadtA, Zimelie 1. — Edition: H.-K. Claußen (Hg.). Freisinger Rechtsbuch. Weimar 1941.
TrHL	*St. Trudperter Hohes Lied (A)*. Handschrift: Wien, Österr. Nationalbibl., Cod. 2719. — Edition: H. Menhardt (Hg.). Das St. Trudperter Hohe Lied. Halle a.d. Saale 1934. 1–42.
VAlex	*Lambrechts Alexander (Vorauer Alexander)*. Vorau, Stiftsbibl., Cod. 276, Bl. 109ra–115va. — Edition: K. Kinzel (Hg.). Lamprechts Alexander. Nach den drei Texten mit dem Fragment des Alberic von Besançon und den lateinischen Quellen. Halle 1884.
Will	*Williram von Ebersberg: 'Hoheliedkommentar'*. Breslau / Wrocław, Stadtbibl., Cod. R 347 [Kriegsverlust]. — Edition: E. H. Bartelmez (Hg.). The „Expositio in Cantica Canticorum" of Williram, Abbot of Ebersberg 1048–1085. A Critical Edition. Philadelphia 1967.

WMEv *Wien–Münchner Evangelienübersetzung*. Handschrift: München, Staatsbibl., Cgm 5250/1; Oxford, Bodleian Libr., MS Germ. b. 3, f. 15; Wien, Österr. Nationalbibl., Cod. Ser. nova 249. — Edition: H. Kriedte (Hg.). Deutsche Bibelfragmente in Prosa des XII. Jahrhunderts. Halle a.d. Saale 1930. 11–14, 64–122.

WNot *Wiener Notker*. Handschrift: Wien, ÖNB, Codex 2681 („Wiener Notker“ = Notker Y). — Edition: R. Heinzel u. W. Scherer (Hg.). Notkers Psalmen nach der Wiener Handschrift. Straßburg/London 1876.

Literatur

Dudenredaktion (Hg.) (2009). *Die Grammatik*. Duden 4. Dudenverlag, Mannheim/Wien/Zürich. 8., überarb. Aufl.

Höchli, S. (1981). *Zur Geschichte der Interpunktion im Deutschen*. de Gruyter, Berlin/New York.

Lexer, M. (1872–1878). *Mittelhochdeutsches Handwörterbuch*. Hirzel, Leipzig.

Schiller, A., Teufel, S., Stöckert, C. und Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technischer Bericht, Universitäten Stuttgart und Tübingen, `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf`.

Schottelius, J. G. (1641). *Teutsche Sprach Kunst*. Braunschweig. (2. Aufl. 1651).

Splett, J. (1993). *Althochdeutsches Wörterbuch*. de Gruyter, Berlin/New York. 2 Bände.

Appendizes

Appendix I: Token-Tags

Appendix II: Wortart-Tags

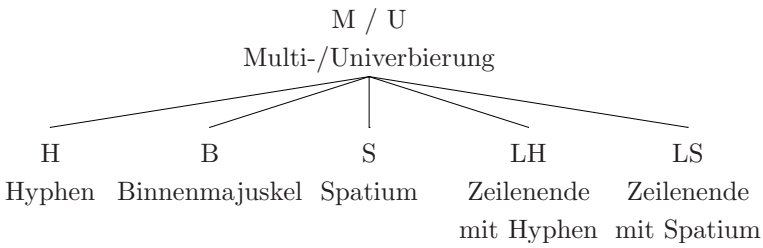
1	ADJ	Adjektive
2	AP	Adpositionen
3	AV	Adverbien
4	CARD	Kardinalzahlen
5	D, P	Determinativa und Pronomen
6	KO	Konjunktionen
7	N	Nomen
8	PAV	Pronominaladverbien
9	PTK	Partikeln
10	V	Verben
11	ITJ, FM	Verschiedenes
12	$	Interpunktion

Appendix III: Alphabetische Auflistung aller Beleg-Tags

Hinweise zu den Tabellen I und II in den Appendizes

- Die folgenden Tabellen geben eine komplette Auflistung aller in HiTS vorgesehenen Tags. Jeder Tabelle ist eine graphische Darstellung der verwendeten Abkürzungen in den Tagnamen vorangestellt.
- Die Tabellen zu den Wortart-Tags beginnen jeweils mit den „Normalfällen“, d.h. typischerweise solchen Fällen, in denen die Hauptwortart des Lemmas und des Belegs übereinstimmen (ein Beispiel aus der Adjektiv-Tabelle: ADJ > ADJA). Als nächstes werden (exemplarisch) Fälle gelistet, in denen die Lemmawortart abweicht (VVPP > ADJA), gefolgt von Fällen, in denen die Belegwortart abweicht (ADJ > AVD).
- Gelegentlich stehen unter den Tabellen noch Anmerkungen, die auf Unterschiede zum STTS hinweisen (z.B. wird darauf hingewiesen, dass und warum es in HiTS kein Tag APPRART gibt).

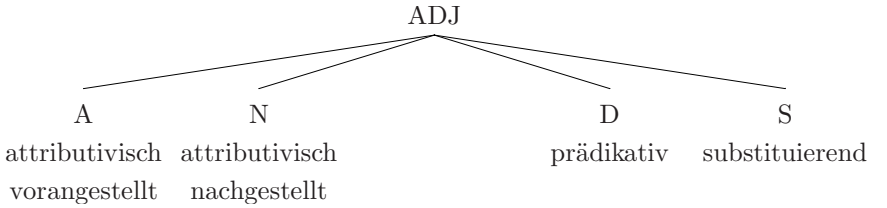
Appendix I: Überblickstabelle Token-Tags



Beispiel		**HiTS**
original	**modernisiert**	
indaz	*in daz*	**MS** (*Multiverbierung mit Spatium*)
LandGraff	*Landgraff*	**UB** (*Univerbierung mit Binnenmajuskel*)
Liebes=Ohnmachten	*Liebesohnmachten*	**UH** (*Univerbierung mit Hyphen*)
Schutz Gott	*Schutzgott*	**US** (*Univerbierung mit Spatium*)
in–↩ dem lande	*in dem lande*	**MLH** (*Multiverbierung am Zeilenende mit Hyphen*)
ou=↩ gen	*ougen*	**ULH** (*Univerbierung am Zeilenende mit Hyphen*)
ou↩ gen	*ougen*	**ULS** (*Univerbierung am Zeilenende mit Spatium*)

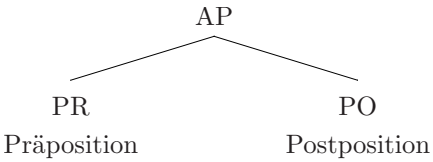
Appendix II: Überblickstabellen Wortart-Tags

1 Adjektive



Beispiel	HiTS	STTS
der [***reht***] *rihtari* 'der gerechte Richter'	**ADJ > ADJA** *Adjektiv, attributiv, vorangestellt*	ADJA *attr. Adjektiv*
die brugge [***reht***] 'die richtige Brücke'	**ADJ > ADJN** *Adjektiv, attributiv, nachgestellt*	(ADJD; selten)
daz almuoſen iſt [***reht***] 'das Almosen ist gerecht'	**ADJ > ADJD** *Adjektiv, prädikativ*	ADJD *präd. oder adv. Adj.*
die [***rehten***] *ſkinent* 'die gerechten/Gerechten strahlen'	**ADJ > ADJS** *Adjektiv, substituierend*	ADJA; NN *attr. Adj.; normales Nomen*
ůf ainem [***getouweten***] *chle* 'auf einem betauten Klee'	**VVPP > ADJA** *Partizip Präteritum, adjektivisch*	ADJA *attr. Adj.*
daz [***brinnent***] *ole* 'das brennende Öl'	**VVPS > ADJA** *Partizip Präsens, adjektivisch*	ADJA *attr. Adj.*
Außerdem:		
lebet [***rehto***] 'lebt gerecht!'	**ADJ > AVD** *Adjektiv, adverbial*	ADJD; ADV *präd. oder adv. Adj.; Adverb*

2 Adpositionen



Beispiel	HiTS	STTS
[***mit***] *der hant* 'mit der Hand'; *da* [***z***] *en* (= *zen*) *Hunin* 'dort bei den Hunnen/im Hunnenland'	**AP > APPR** *Präposition*	APPR; APPRART *Präposition (mit Artikel)*
inan [***úbari***] 'über ihn'	**AP > APPO** *Postposition*	APPO *Postposition*
Außerdem:		
dâ pâgant siu [***umpi***] 'darum streiten sie'	**AP > PAVAP** *Pron.adverb, präpositionaler Teil*	—

Anmerkungen:

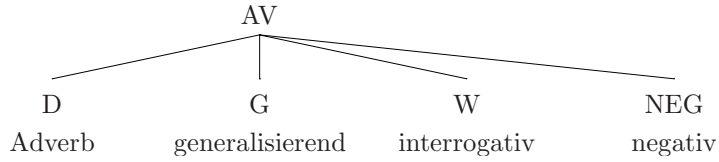
- Kontrahierte Präpositionen (im STTS: APPRART) existieren auch in den früheren Sprachstufen, werden jedoch in HiTS immer als zwei Token (Artikel + Präposition) analysiert:

  zen | *z*/APPR *en*/DDART (< *ze den*) | MS 'zu den'

- Der nachgestellte zweite Teil moderner Zirkumpositionen (im STTS: APZR) entspricht in früheren Sprachstufen einem Nomen:

  durch der liebe uillen/NA 'um der Liebe willen'

3 Adverbien



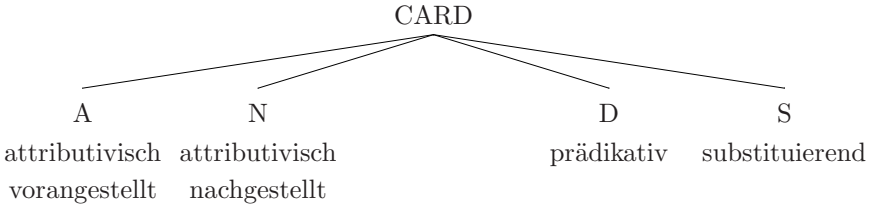
Beispiel	HiTS	STTS
ein gotishus [***uil***] *mechtic* 'eine sehr mächtige Kirche'	**AVD > AVD** *Adverb*	ADV *Adverb*
[***do***] *begagenda imo min trohtin* 'da begegnete ihm mein Herr'	**AVD-KO > AVD** *Konjunktionaladverb, adverbial*	ADV *Adverb*
[***swie***] *sie wellen* 'wie immer sie wollen'	**AVG > AVG** *Relativadverb, generalisierend*	PWAV *adv. Int.- o. Rel.pron.*
[***hwār***] 'wo'; [***wie***] *fro 'wie froh'*	**AVW > AVW** *Adverb, interrogativ*	PWAV *adv. Int.- o. Rel.pron.*
der daz tages lieht [***nie***] *negesah* 'der das Tageslicht nie (nicht) sah'	**AVD > AVNEG** *Adverb, negativ*	AVD *Adverb*
lebet [***rehto***] 'lebt gerecht!'; [***groze***] *willechomen* 'sehr willkommen'	**ADJ > AVD** *Adjektiv, adverbial*	ADJD *präd. oder adv. Adj.*
Außerdem:		
[***dâ***] *pâgant siu umpi* 'darum streiten sie'	**AVD > PAVD** *Pron.adverb, pronominaler Teil*	—
[***uf***] *huob er die hende* 'hoch hob er die Hände'	**AVD > PTKVZ** *Verbzusatz*	PTKVZ *abgetr. Verbzusatz*

Anmerkung:

- Relativische Verwendungen werden ebenfalls mit AVD getagged:

  nv was div ivnchfrovwe genomen her vz, da/AVD *si gevangen lac* 'nun war das Mädchen (dort) herausgeholt worden, wo sie gefangen gelegen hatte'

4 Kardinalzahlen



Beispiel	HiTS	STTS
[***hunderet***] *unde* [***uifzech***] *tage* '150 Tage'	**CARD > CARDA** *Kardinalzahl, attributiv, vorangestellt*	CARD *Kardinalzahl*
daz waren ceichen [***sibeniu***] 'das waren sieben Zeichen'	**CARD > CARDN** *Kardinalzahl, attributiv, nachgestellt*	—
(ohne eindeutigen Korpusbeleg)	**CARD > CARDD** *Kardinalzahl, prädikativ*	CARD *Kardinalzahl*
in [***driv***] *getailet 'in drei geteilt'*; *siben* [***hundert***] *siner manane* '700 seiner Leute; *diu* [***eine***] *'die eine'*	**CARD > CARDS** *Kardinalzahl, substituierend*	CARD *Kardinalzahl*

5 Determinativa und Pronomen

Anmerkungen zu den Determinativa und Pronomen:

- Tags, deren Name mit „D“ beginnt (≈ Determinativa), haben zwei Unterkategorien, für den Typ und die Position, z.B.:

  ein/DIS (Determinativ, indefinit, substituierend)

- Tags, deren Name mit „P“ beginnt (≈ Pronomen), gelten stets als substituierend und haben deswegen nur eine Unterkategorie für den Typ, z.B.:

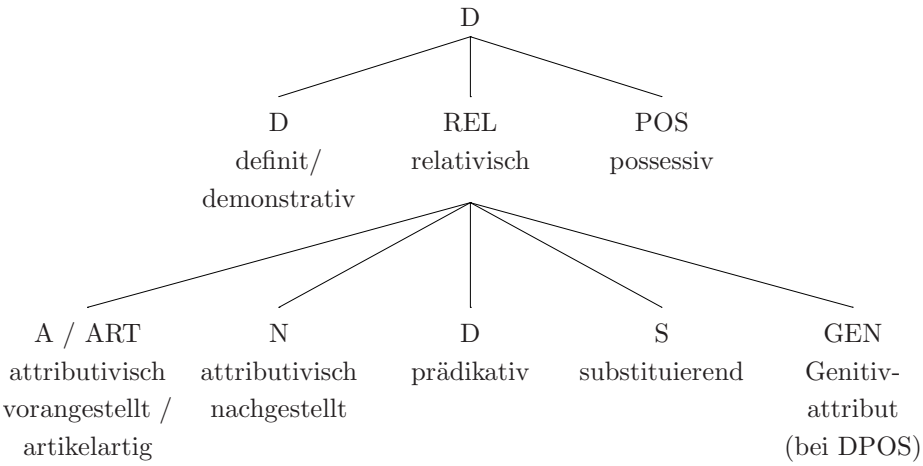
  man/PI (Pronomen, indefinit)

- Attributive Verwendungen im Genitiv, die im STTS mit „AT“ markiert werden (PRELAT, PWAT), werden in HiTS als substituierend annotiert (DRELS, PW), z.B.:

  Sanctus Johannes, des/DRELS *tac wir hivte begen* ‘Sankt Johannes, dessen Tag wir heute begehen’

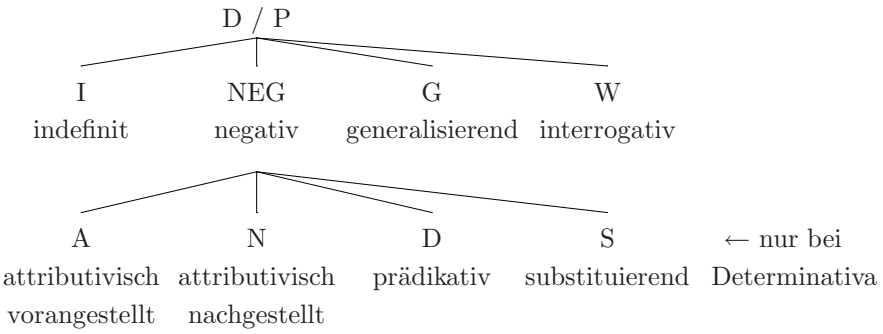
  wes/PW *sun* ‘wessen Sohn’

- Es stehen nicht zu allen theoretisch denkbaren Subklassen Beispiele in den Tabellen. Beispielsweise fehlen bei vielen prädikativen Verwendungen die Einträge (z.B. DDD, DID etc.). Sie sind aber prinzipiell nicht ausgeschlossen.



Beispiel	HiTS	STTS
	Definita/Demonstrativa: DD	
[***der***] *liehte tac* 'der helle Tag'	**DD > DDA** (ad.), **DDART** (mhd.) *Determinativ, definit, vorangestellt bzw. artikelartig*	ART *bestimmter Artikel*
[***dise***] *rede* 'diese Rede'; *ober alle* [***die***] *lant* 'über alle diese Länder'; *vf* [***ienem***] *sal witen* 'in jenem weiten Saal'	**DD > DDA** *Determinativ, definit/demonstrativ, vorangestellt*	PDAT *attr. Dem.pron*
der got [***selbe***] *'Gott selbst'*	**DD > DDN** *Determinativ, definit/demonstrativ, nachgestellt*	—
[***dizze***] *ist ein anphanclich zít* 'dieses ist eine angenehme Zeit'; *in* [***des***] *gewalt* 'in dessen Gewalt'	**DD > DDS** *Determinativ, definit/demonstrativ, substituierend*	PDS *subst. Dem.pron*

Beispiel	HiTS	STTS
	Relativpronomen: DREL	
di haiden, [***dî***] *dort waren* 'die Heiden, die dort waren'; *Sanctus Johannes,* [***des***] *tac wir hivte begen* 'Sankt Johannes, dessen Tag wir heute begehen'	**DD > DRELS** *Determinativ, relativisch, substituierend*	PRELS; PRELAT *subst. oder attr. Rel.pron*
	Possessiva: DPOS	
in [***mînem/irem***] *hûs* 'in meinem/ihrem Haus'	**DPOS > DPOSA** *Determinativ, possessiv, vorangestellt*	PPOSAT *attr. Poss.pron.*
der name [***sîn***] 'sein Name'	**DPOS > DPOSN** *Determinativ, possessiv, nachgestellt*	—
thaz dar thin [***thîn***] *ist* 'das da dein ist'	**DPOS > DPOSD** *Determinativ, possessiv, prädikativ*	PPOSS *subst. Poss.pron.*
alle die [***sîne***] 'all die Seinen'	**DPOS > DPOSS** *Determinativ, possessiv, substituierend*	NN *normales Nomen*
mid [***iro***] *handon* 'mit ihren Händen (= mit den Händen ihrer/von ihnen)'	**PPER > DPOSGEN** *Determinativ, personal-possessiv, Genitiv (aus Pronomen)*	—



Beispiel	HiTS	STTS
	Indefinita: DI/PI	
[***ein***] *tîer* 'ein Tier'	**DI > DIA** (ad.), **DIART** (mhd.) *Determinativ, indefinit, vorangestellt bzw. artikelartig*	ART *unbestimmter Artikel*
[***alliv***] *iar* 'alle Jahre'; *thúruh* [***thehéinan***] *wóroltruam* 'für irgendeinen irdischen Ruhm'	**DI > DIA** *Determinativ, indefinit, vorangestellt*	PIAT *attr. Indef.pron.*
under disen chunigen [***allen***] 'unter all diesen Königen'	**DI > DIN** *Determinativ, indefinit, nachgestellt*	—
der redet [***vil***] 'der redet viel'; [***thehein***] *thero fórasagono* '(irgend-) einer der Propheten'	**DI > DIS** *Determinativ, indefinit, substituierend*	PIS *subst. Indef.pron.*
[***íaman***] *hiar in lánte* 'jemand hier im Land'	**PI > PI** *Pronomen, indefinit*	PIS *subst. Indef.pron.*
[***huer***] *'(irgend)wer'*	**PW > PI** *Pronomen, indefinit*	PIS *subst. Indef.pron.*
[***man***] *'man'*	**NA > PI** *Pronomen, indefinit (aus Substantiv)*	PIS *subst. Indef.pron.*
Außerdem:		
di en salt du [***nit***] *schouwen* 'die sollst du nicht anschauen'	**PI > PTKNEG** *Negationspartikel (aus Pronominalsubstantiv)*	PTKNEG *Negationspartikel*

Beispiel	HiTS	STTS
	Negative: DNEG/PNEG	
ther heilant ni gab imo [***nohhein***] *antuurti* 'der Heiland (nicht) gab ihm keine Antwort'	**DI > DNEGA** *Determinativ, negativ, vorangestellt*	PIAT *attr. Indef.pron.*
ni si mán [***nihein***] *so véigi* 'kein Mann soll so gottlos sein'	**DI > DNEGN** *Determinativ, negativ, nachgestellt*	PIAT *attr. Indef.pron.*
[***nihéinan***] *ni gifíangun* 'sie nahmen keinen/niemanden gefangen'	**DI > DNEGS** *Determinativ, negativ, substituierend*	PIS *subst. Indef.pron.*
[***niowiht***] 'nichts'; [***nioman***] *giuuisso in taugle uuaz tuot* 'gewiß tut niemand etwas im Verborgenen'	**PI > PNEG** *Pronomen, indefinit, negativ*	PIS *subst. Indef.pron.*

Anmerkungen:

- Ad. *nohhein* und mhd. *nehein* treten nur in negativen Kontexten (d.h. mit Negation) auf und werden mit DI > DNEGA etc. annotiert:

  ther heilant ni gab imo nohhein/DI > DNEGA *antuurti* 'Der Heiland (nicht) gab ihm keine Antwort' (ad.)

  daz ne saget uns nehein/DI > DNEGA *puch daz deheiner so riche ware* 'davon (nicht) berichtet uns kein Buch, dass jemand so mächtig gewesen wäre' (mhd.)

- Ad. *thehein* kommt nur in nicht-affirmativen Kontexten (z.B. in Fragesätzen, hypothetischen Sätzen, mit Negation; wie Englisch *any(one)*) vor und wird mit DI > DIA/DIN etc. annotiert.

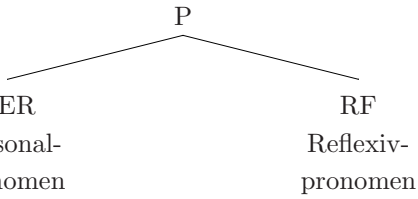
  nist ist thehein/DI > DIA *tuala* 'es gibt keinen Zweifel (= nicht ist irgendein Zweifel' (ad.)

  war imo súlih man thihéin/DI > DIN *so quami wísheiti héim* 'woher käme ihm ein solcher Mann (= solch Mann (irgend)ein) der Weisheit heim (ad.)

- Mhd. *dehein* (der Vorläufer von nhd. *kein*) wird teilweise noch wie ad. *thehein* in nicht-affirmativen Kontexten (wie Englisch *any(one)*) verwendet, teilweise drückt es aber auch schon allein die Negation aus (wie Englisch *no(one)*). Da die Unterscheidung teilweise nur schwer zu treffen ist, werden beide Fälle mit DI > DNEGA/DNEGS etc. annotiert.

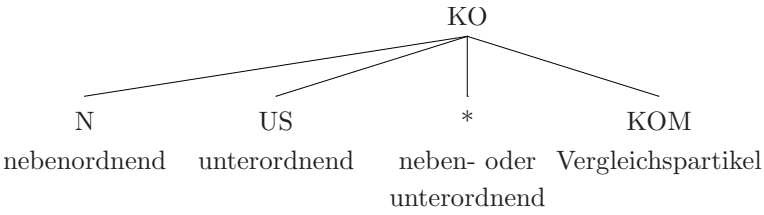
  daz ne saget uns nehein puch daz deheiner/DI > DNEGS *so riche ware* 'davon (nicht) berichtet uns kein Buch, dass jemand so mächtig wäre' (mhd.)

Beispiel	HiTS	STTS
	Generalisierende: DG/PG	
uernemet . . . , [***swelh***] *rat iv baz geualle* 'vernehmt . . . , welcher Rat euch besser gefällt'	**DG > DGA** *Determinativ, generalisierend, vorangestellt*	PWAT *attr. Interr.pron.*
[***swelh***] *baz entwíchen mag, der sol ouch entwichen* 'welcher [sc. Wagen] besser ausweichen kann, der soll auch ausweichen'	**DG > DGS** *Determinativ, generalisierend, substituierend*	PWS *subst. Interr.pron.*
[***swer***] *mir nîene tvot, der sol ôvch mich zefrivnde han* 'wer mir nichts tut, der soll auch mich zum Freund haben'; *ín* [***swez***] *wiltbanne* 'in wessen Jagdrevier'	**PG > PG** *Pronomen, generalisierend*	PWS; PWAT *subst. oder attr. Interr.pron.*
	Interrogativa: DW/PW	
owi, [***welh***] *mort da gefrumet wart!* 'oh weh, welch ein Mord wurde da verübt!'	**DW > DWA** *Determinativ, interrogativ, vorangestellt*	PWAT *attr. Interr.pron.*
[***wélih***]*?* 'welcher?'	**DW > DWS** *Determinativ, interrogativ, substituierend*	PWS *subst. Interr.pron.*
[***wer***]*?* 'wer?'; [***wes***] *beduncheт whc uon criste,* [***wes***] *sun er si?* 'Was meint ihr von Christus, wessen Sohn er sei?'	**PW > PW** *Pronomen, interrogativ*	PWS; PWAT *subst. oder attr. Interr.pron.*



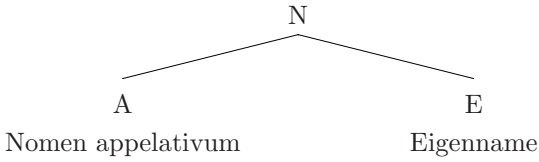
Beispiel	HiTS	STTS
	Sonstige Pronomen	
[***er***] 'er'	**PPER > PPER** *Pronomen, personal, irreflexiv*	PPER *irrefl. Pers.pron.*
im(o) 'sich'; [***einander***] 'einander'	**PPER > PRF** *Pronomen, personal, reflexiv*	PRF *refl. Pers.pron.*
[***sich***] 'sich'	**PRF > PRF** *Pronomen, personal, reflexiv*	PRF *refl. Pers.pron.*
Außerdem:		
mid [***iro***] *handon* 'mit ihren Händen (= mit den Händen ihrer/von ihnen)'	**PPER > DPOSGEN** *Pronomen, personal-possessiv, Genitiv*	—

6 Konjunktionen



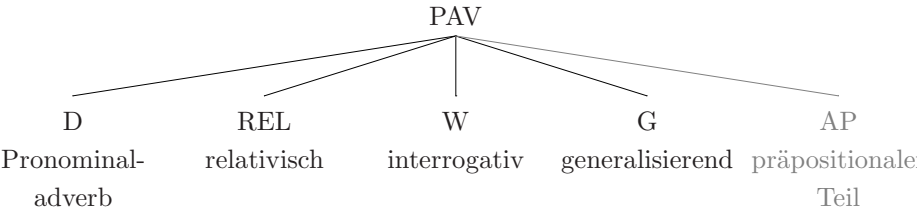
Beispiel	HiTS	STTS
ih [***unde***] *iohan* 'ich und Johann'	**KO > KON** *Konjunktion, nebenordnend*	KON *nebenordn. Konj.*
[***sundir***] *du solt in mich verwandelot werden* 'sondern du sollst/wirst in mich verwandelt werden'	**AVD-KO > KON** *Konjunktionaladverb, nebenordnend*	KON *nebenordn. Konj.*
[***daz***] *si giengen heim* 'dass sie heimgingen'	**KO > KOUS** *Konjunktion, unterordnend*	KOUS *unterordn. Konj.*
do sint si inne, [***want***] *her betit* 'darin sind sie, denn/da er betet'	**KO > KO*** *Konjunktion, neben- oder unterordnend*	KON; KOUS *neben- o. unterordn. Konj.*
mer [***dannî***] *ein iar* 'mehr als ein Jahr'	**KO > KOKOM** *Vergleichspartikel*	KOKOM *Vergleichskonj.*
Außerdem:		
[***do***] *begagenda imo min trohtin* 'da begegnete ihm mein Herr'	**AVD-KO > AVD** *Konjunktionaladverb, adverbial*	ADV *Adverb*

7 Nomen



Beispiel	HiTS	STTS
der [***kuninc***] 'der König'	**NA > NA** *Nomen appelativum*	NN *normales Nomen*
ze [***rome***] 'nach Rom'	**NE > NE** *Eigenname*	NE *Eigennamen*
mit [***suften***] *und mit* [***weinen***] 'mit Seufzen und mit Weinen'; *ez ist wol ze* [***lobenne***] 'es ist sehr zu loben'	**VVINF > NA** *Infinitiv, substantiviert*	NN *normales Nomen*

8 Pronominaladverbien



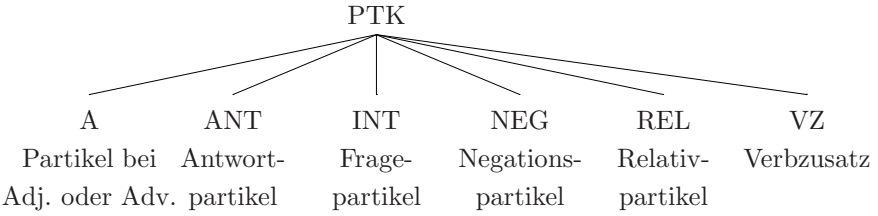
Beispiel	HiTS (zweiteilig) pronominal	präpositional	STTS
	Pronominaladverbien		
[***dâ***] *pâgant siu* [***umpî***] 'darum streiten sie'; [***dā***] *was vil volkes* [***inne***] 'darin waren viele Leute'	**AVD > PAVD** *Pron.adverb*	**+ AP > PAVAP**	PAV *Pron.adverb*
thaz fúndament . . ., [***thar***] *thiu érda ligit* [***úfe***] 'das Fundament, worauf die Erde ruht'; *der palas,* [***dā***] *. . .* [***inne***] *was* 'der Palast, in dem'	**AVD > PAVREL** *Pron.adverb, relativisch*	**+ AP > PAVAP**	PWAV *adv. Int.- o. Rel.pron.*
[***wā . . . umbe***] 'warum'	**AVW > PAVW** *Pron.adverb, interrogativ*	**+ AP > PAVAP**	PWAV *adv. Int.- o. Rel.pron.*
[***swā . . . umbe***] 'warum (auch immer)'	**AVG > PAVG** *Pron.adverb, generalisierend*	**+ AP > PAVAP**	PWAV *adv. Int.- o. Rel.pron.*

Anmerkung:

- Anstelle der Präpositionen können auch bestimmte Adverbien in Pronominaladverbien vorkommen. Diese erhalten als Lemmawortart ebenfalls AP zugewiesen:

  ez chnite fuor im dar/PAV > PAVD *nider*/AP > PAVAP 'es kniete vor ihm nieder'

9 Partikeln



Beispiel	HiTS	STTS
[***ze***] *lanc* 'zu lang'; [***ze***] *uil* 'zu viel'	**PTK > PTKA** *Partikel bei Adjektiv oder Adverb*	PTKA *Part. bei Adj. o. Adv.*
[***nêin***]*, sprach der herre Gawêin* "'nein", sagte Herr Gawein'	**PTK > PTKANT** *Antwortpartikel*	PTKANT *Antwortpartikel*
[***eno***] *ni lâsut ir in giscribun* 'Last ihr etwa nicht in der Heiligen Schrift?'	**PTK > PTKINT** (ad.) *Fragepartikel*	—
di [***en***] *salt du nit schouwen* 'die sollst du nicht anschauen'	**PTK > PTKNEG** *Negationspartikel*	PTKNEG *Negationspartikel*
than is im sô them salte, [***the***] *man bi sêes staðe uuidô teuuirpit* 'dann geht es ihm wie dem Salz, das man am Meeresufer weithin zerstreut'	**PTK > PTKREL** *Relativpartikel*	—
[***uf***] *huob er die hende* 'hoch hob er die Hände'	**AVD > PTKVZ** *Verbzusatz*	PTKVZ *abgetr. Verbzusatz*
di en salt du [***nit***] *schouwen* 'die sollst du nicht anschauen'	**PI > PTKNEG** *Negationspartikel (aus Pronominalsubstantiv)*	PTKNEG *Negationspartikel*

Anmerkungen:

- Anders als im STTS werden in HiTS Verbpartikeln und Basisverben immer getrennt annotiert, selbst bei Kontaktstellung:

  (*als er den brief*) *anesach* | *ane*/PTKVZ *sach*/VVFIN 'als er den Brief ansah'

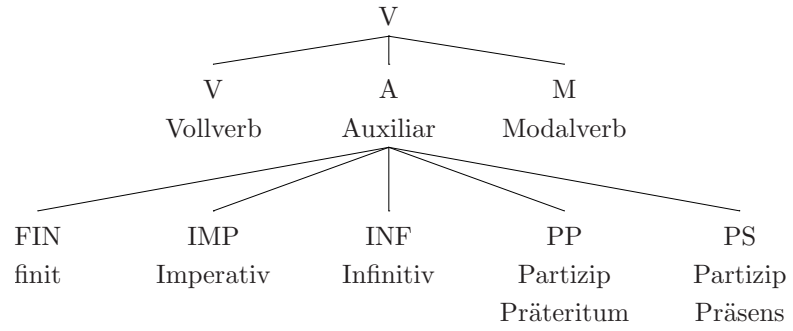
- Zum STTS-Tag PTKZU für „*zu* vor Infinitiv" gibt es in HiTS keine Entsprechung. In den älteren Sprachstufen gibt es eine ähnlich aussehende Vorgängerkonstruktion,

in der der Infinitiv allerdings flektiert ist (Dativ) und *zu* daher als Präposition zu analysieren ist:

ez ist wol ze/AP > APPR *lobenne*/VVINF > NA 'es ist sehr zu loben'

daz si ime ze/AP > APPR *helfen*/VVINF > NA *chomen* 'dass sie ihm helfen kommen/zu Hilfe kommen'

10 Verben



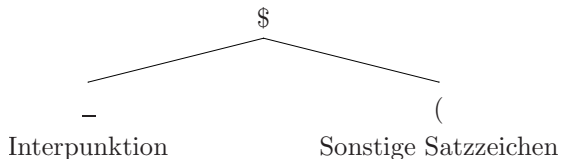
Beispiel	HiTS	STTS
	Vollverben: VV	
er [***sprach***] 'er sprach'	**VV > VVFIN** *Vollverb, finit*	VVFIN *finites Verb, voll*
[***gloubet***] *mir* 'glaubt mir'	**VV > VVIMP** *Vollverb, Imperativ*	VVIMP *Imperativ, voll*
si begunden in [***fragen***] 'sie begannen ihn zu fragen'	**VV > VVINF** *Vollverb, Infinitiv*	VVINF *Infinitiv, voll*
der winter was [***vergan***] 'der Winter war vergangen'	**VV > VVPP** *Vollverb, Partizip Präteritum, im Verbalkomplex*	VVPP *Partizip Perfekt, voll*
dannen wirt der lîp [***sennende***] 'deshalb beginnt der Leib, sich zu sehnen'; *. . .unt wir da sin* [***mendenti***] 'und wir uns dort freuen mögen (= uns freuend seien)'	**VV > VVPS** *Vollverb, Partizip Präsens, im Verbalkomplex*	ADJD *adv. oder präd. Adj.*
Außerdem:		
mit [***suften***] *und mit* [***weinen***] 'mit Seufzen und mit Weinen'	**VVINF > NA** *Infinitiv, substantiviert*	NN *normales Nomen*
ûf ainem [***getouweten***] *chle* 'auf einem betauten Klee'	**VVPP > ADJA** etc. *Partizip Präteritum, adjektivisch*	ADJA etc.
daz [***brinnent***] *ole* 'das brennende Öl'	**VVPS > ADJA** etc. *Partizip Präsens, adjektivisch*	ADJA etc.

Beispiel	HiTS	STTS
	Auxiliare und Modalverben: VA/VM	
der winter [***was***] *vergan* 'der Winter war vergangen'	**VA** > **VAFIN** *Auxiliar, finit*	VAFIN *finites Verb, aux*
...	**VA** > **VAFIN** etc.	VAFIN etc.
er [***scol***] *unser helfare wesen* 'er soll unser Helfer sein'	**VM** > **VMFIN** *Modalverb, finit*	VMFIN *finites Verb, modal*
...	**VM** > **VMINF** etc.	VMINF etc.

11 Verschiedenes

Beispiel	HiTS	STTS
[***o we***] 'o weh'	**ITJ > ITJ** *Interjektion*	ITJ *Interjektion*
Díu sûoze dînero [***gratię***] *ist bézzera dánne díu scárfe déro* [***legis***] 'die Süße deiner Gnade (gratiae) ist besser als die Schärfe des Gesetzes (legis)'	**FM > FM** *Fremdsprachliches Material*	FM *Fremdspr. Material*

12 Interpunktion



Beispiel	HiTS	STTS
[. : , / ? !]	**$_** > **$_** *originale Interpunktion*	$, $. *Komma; satzbeendende Interpunktion*
[( ) ’ “ ”]	**$(** > **$(** *sonstige Satzzeichen*	$(*sonstige Satzzeichen*

Appendix III: Alphabetische Auflistung aller Beleg-Tags

Tag	Beschreibung
ADJA	*Adjektiv, attributiv, vorangestellt*
ADJD	*Adjektiv, prädikativ*
ADJN	*Adjektiv, attributiv, nachgestellt*
ADJS	*Adjektiv, substituierend*
APPO	*Postposition*
APPR	*Präposition*
AVD	*Adverb*
AVG	*Relativadverb, generalisierend*
AVNEG	*Adverb, negativ*
AVW	*Adverb, interrogativ*
CARDA	*Kardinalzahl, attributiv, vorangestellt*
CARDD	*Kardinalzahl, prädikativ*
CARDN	*Kardinalzahl, attributiv, nachgestellt*
CARDS	*Kardinalzahl, substituierend*
DDA	*Determinativ, definit, attributiv, vorangestellt* (ad.)
DDART	*Determinativ, definit, artikelartig* (mhd.)
DDD	*Determinativ, definit/demonstrativ, prädikativ*
DDN	*Determinativ, definit/demonstrativ, attributiv, nachgestellt*
DDS	*Determinativ, definit/demonstrativ, substituierend*
DGA	*Determinativ, generalisierend, attributiv, vorangestellt*
DGD	*Determinativ, generalisierend, prädikativ*
DGN	*Determinativ, generalisierend, attributiv, nachgestellt*
DGS	*Determinativ, generalisierend, substituierend*
DIA	*Determinativ, indefinit, attributiv, vorangestellt* (ad.)
DIART	*Determinativ, indefinit, artikelartig* (mhd.)
DID	*Determinativ, indefinit, prädikativ*
DIN	*Determinativ, indefinit, attributiv, nachgestellt*
DIS	*Determinativ, indefinit, substituierend*
DNEGA	*Determinativ, negativ, attributiv, vorangestellt*
DNEGD	*Determinativ, negativ, prädikativ*
DNEGN	*Determinativ, negativ, attributiv, nachgestellt*
DNEGS	*Determinativ, negativ, substituierend*
DPOSA	*Determinativ, possessiv, attributiv, vorangestellt*
DPOSD	*Determinativ, possessiv, prädikativ*
DPOSGEN	*Determinativ, personal-possessiv, Genitiv*
DPOSN	*Determinativ, possessiv, attributiv, nachgestellt*
DPOSS	*Determinativ, possessiv, substituierend*
DRELS	*Determinativ, relativisch, substituierend*
DWA	*Determinativ, interrogativ, attributiv, vorangestellt*
DWD	*Determinativ, interrogativ, prädikativ*
DWN	*Determinativ, interrogativ, attributiv, nachgestellt*
DWS	*Determinativ, interrogativ, substituierend*

FM	*Fremdsprachliches Material*
ITJ	*Interjektion*
KO*	*Konjunktion, neben- oder unterordnend*
KOKOM	*Vergleichspartikel*
KON	*Konjunktion, nebenordnend*
KOUS	*Konjunktion, unterordnend*
NA	*Nomen appelativum*
NE	*Eigenname*
PAVAP	*Pronominaladverb, präpositionaler Teil*
PAVD	*Pronominaladverb, pronominaler Teil*
PAVG	*Pronominaladverb, pronominaler Teil, generalisierend*
PAVREL	*Pronominaladverb, pronominaler Teil, relativisch*
PAVW	*Pronominaladverb, pronominaler Teil, interrogativ*
PG	*Pronomen, generalisierend*
PI	*Pronomen, indefinit*
PNEG	*Pronomen, indefinit, negativ*
PPER	*Pronomen, personal, irreflexiv*
PRF	*Pronomen, personal, reflexiv*
PTKA	*Partikel bei Adjektiv oder Adverb*
PTKANT	*Antwortpartikel*
PTKINT	*Fragepartikel* (ad.)
PTKNEG	*Negationspartikel*
PTKREL	*Relativpartikel*
PTKVZ	*Verbzusatz*
PW	*Pronomen, interrogativ*
VAFIN	*Auxiliar, finit*
VAIMP	*Auxiliar, Imperativ*
VAINF	*Auxiliar, Infinitiv*
VAPP	*Auxiliar, Partizip Präteritum, im Verbalkomplex*
VAPS	*Auxiliar, Partizip Präsens, im Verbalkomplex*
VMFIN	*Modalverb, finit*
VMIMP	*Modalverb, Imperativ*
VMINF	*Modalverb, Infinitiv*
VMPP	*Modalverb, Partizip Präteritum, im Verbalkomplex*
VMPS	*Modalverb, Partizip Präsens, im Verbalkomplex*
VVFIN	*Vollverb, finit*
VVIMP	*Vollverb, Imperativ*
VVINF	*Vollverb, Infinitiv*
VVPP	*Vollverb, Partizip Präteritum, im Verbalkomplex*
VVPS	*Vollverb, Partizip Präsens, im Verbalkomplex*
$_	*originale Interpunktion*
$(	*sonstige Satzzeichen*

Swantje Westpfahl, Thomas Schmidt

POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch

1 Einleitung

Im Rahmen des FOLK-Projekts (Forschungs- und Lehrkorpus Gesprochenes Deutsch), das am *Institut für Deutsche Sprache* (IDS) ein großes wissenschaftsöffentliches Gesprächskorpus aufbaut, soll mit Hilfe des *TreeTaggers* (SCHMID 1995) und des *Stuttgart-Tübingen-Tagsets* (STTS), (SCHILLER ET AL. 1999) ein automatisiertes Part-of-Speech-Tagging (POS-Tagging) für Spontansprache ermöglicht werden. Zuerst nur auf FOLK angewendet, soll dieser Tagger später auch für weitere Korpora spontansprachlicher Daten in der *Datenbank für Gesprochenes Deutsch* (DGD), (INSTITUT FÜR DEUTSCHE SPRACHE) genutzt werden. Da das *Forschungs- und Lehrkorpus* kontinuierlich ausgebaut wird, muss das POS-Tagging aus Effizienzgründen mittelfristig vollautomatisch erfolgen. Dabei wird eine Fehlerquote von unter 5 Prozent angestrebt.

Weil sowohl das Tagset als auch der Tagger für geschriebene Sprache konzipiert bzw. trainiert wurden und beim automatisierten Taggen der Transkripte die Fehlerquote bei fast 20 Prozent lag, muss eine Anpassung sowohl des Tagging-Verfahrens als auch des Tagsets an Spontansprache vorgenommen werden. Aus diesem Grund wurden die Fehler, die bei einem ersten Versuch des automatisierten Taggings dreier Transkripte des Korpus mit dem TreeTagger und dem STTS auftraten, auf ihre Ursachen hin analysiert. Daraufhin konnten Vorschläge zur Verbesserung des POS-Taggings in Hinblick auf eine Anpassung des Tagsets sowie des Tagging-Verfahrens gemacht werden.

2 Methodik

2.1 Auswahl der Transkripte

Für einen ersten Versuch des automatisierten POS-Taggings von spontansprachlichen Daten wurden Transkripte aus möglichst unterschiedlichen Bereichen der Alltagskommunikation ausgewählt:

1. Eine Berufsschulinteraktion[1],
2. ein Alltagsgespräch von Studenten in der Mensa[2],
3. und eine Kind-Kind-Vorleseinteraktion[3].

Die Transkripte sowie deren Metadaten sind auf der DGD-Webseite abrufbar.[4] Durch die Auswahl dieser unterschiedlichen Kommunikationssituationen sollte vermieden werden, dass Probleme beim Taggen, die einer bestimmten Art der Kommunikation geschuldet sind, einerseits zu sehr in den Vordergrund gelangen, andererseits eventuell unberücksichtigt blieben. Die Berufsschulinteraktion ist stark regionalsprachlich und, durch die Frage-Antwort-Struktur des Unterrichts und durch die geregelte Rederechtsverteilung, stark insti-

tutionell geprägt. Im Gegensatz dazu ist das studentische Alltagsgespräch kaum dialektal geprägt und durch das ungezwungene Beisammensein der Studenten in der Mensa eher persönlich. Zuletzt lässt das Kind-Kind-Vorlesen, bedingt durch das Alter der Kinder, Fälle von nicht standardsprachlicher Wortstellung sowie auch, durch das Vorlesen von Textpassagen, konzeptionell schriftliches Sprechen im Sinne des Nähe/Distanz-Modells (KOCH & OESTERREICHER 1985) erwarten. Mit einer Anzahl von insgesamt 11.029 Tokens kann die Auswahl als ausreichend große und differenzierte Grundlage für die Analyse eines ersten Tagging-Versuchs gelten.

2.2 Erstellung und Beschaffenheit der Transkripte

Transkripte des FOLK-Korpus werden wie folgt erstellt: Die Audiodaten werden mit dem Transkriptionseditor FOLKER[5] konform nach cGAT[6] als Minimaltranskripte[7] in literarischer Umschrift[8] transkribiert und mit dem Originalton aligniert. (SCHMIDT 2012) Auf Interpunktion und Annotation von Intonationsverläufen oder von pragmatischen Einheiten wird dabei gemäß cGAT verzichtet, um die Transkribenten von zeitaufwändigen und oft stark interpretativen Entscheidungen zu entlasten. Des Weiteren werden Pausen, die länger als 0,2 Sekunden sind, keinem Sprecher zugeordnet, sodass durch Pausen unterbrochene Äußerungseinheiten (seien sie syntaktischer, prosodischer oder pragmatischer Natur) in mehrere Segmente zerteilt werden (SCHMIDT & SCHÜTTE 2011).

Nach mindestens zweifacher Korrektur eines Transkriptes wird es ‚normalisiert', das heißt mit Hilfe des Programms *OrthoNormal*[9] semi-manuell[10] auf einer weiteren Ebene mit den standardorthographischen Entsprechungen der literarisch transkribierten Tokens annotiert, was die Suche nach Wörtern und Wortverbindungen in der Datenbank für gesprochenes Deutsch erleichtern soll (SCHMIDT 2012). Wie Abbildung 1 verdeutlicht, werden dabei u.a. elliptische, umgangssprachliche oder dialektale Formen auf ihre standardorthographischen Entsprechungen, sowie kleingeschriebene Substantive auf großgeschriebene Formen abgebildet.

Transkription	da	gehst	de	jetz	einfach	über	dem	bild
Normalisierung	da	gehst	**du**	**jetzt**	einfach	über	dem	**Bild**

Abbildung 1: Transkriptausschnitt FOLK_E_00086_SE_01_T_01 aus dem FOLK-Korpus mit normalisierten Formen

Wie ein initialer Test gezeigt hat, ist die orthographische Normalisierung auch eine unabdingbare Voraussetzung für ein erfolgreiches Part-of-Speech-Tagging. Während ein Tagging mit dem Default-TreeTagger-Parameterfile auf Transkripten mit literarischer Umschrift Fehlerquoten zwischen 30 und 35 Prozent ergibt, verbessern sich diese bereits auf etwa 20 Prozent, wenn statt mit der literarischen Umschrift mit den normalisierten Formen gearbeitet wird (s.u.).

2.3 Tagging und manuelle Korrektur des ersten Taggingversuchs

Für das initiale POS-Tagging wurde der TreeTagger (über den Java-Wrapper TT4J, ECKART 2013) mit dem Default-Parameterfile verwendet. Dabei stellt sich aufgrund der nicht vorhandenen Interpunktion in den Transkripten (s.o.) die Frage, welche Einheiten idealerweise an den Tagger zu übergeben sind. Bei einer Übergabe eines Transkripts als Gesamttext würden Sprecherwechsel, die in aller Regel auch syntaktische Grenzen darstellen, ignoriert. Die Transkripte wurden daher nicht als Gesamttexte, sondern (sprecher-)beitragsweise an den Tagger übergeben.

Für die manuelle Korrektur der Tags der automatisiert getaggten Transkripte wurde ebenfalls das Programm *OrthoNormal*[11] verwendet. Zwei Screenshots sollen die Arbeitsschritte verdeutlichen. Der erste Screenshot (Abbildung 2) zeigt, wie man in der Beitragsansicht die einzelnen Tags der Wörter und die Wahrscheinlichkeit der Tags aufrufen kann. In diesem Falle wurde das Tag *Partizip Perfekt*, *Vollverb* mit hundertprozentiger Wahrscheinlichkeit, vergeben.

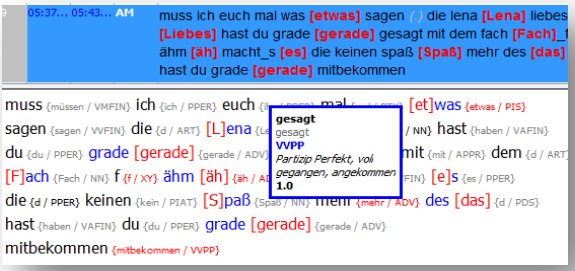


Abbildung 2: Anzeige der Wahrscheinlichkeiten und des Tags in OrthoNormal (Version 0.6)

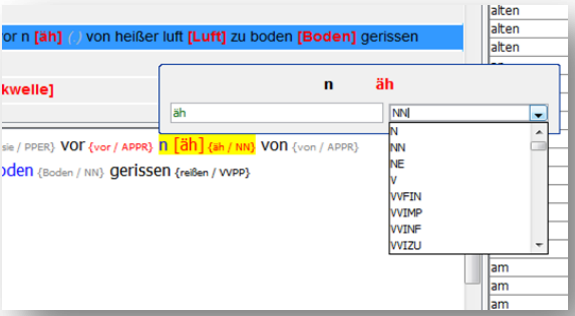


Abbildung 3: Korrektur der Tags in OrthoNormal (Version 0.6)

Der zweite Screenshot (Abbildung 3) verdeutlicht die Vorgehensweise der Korrektur: Das Auswählen des zu korrigierenden Items und die Neuzuordnung des Tags aus der Liste der verfügbaren Tags im Tagset. Da ‚*äh*' kein Nomen ist, muss ein anderes Tag ausgewählt werden.

Die Korrektur orientierte sich bei der Zuordnung der Wortarten an der Duden-Grammatik, da diese nach den Normalisierungskonventionen von ARNULF DEPPERMANN, WILFRIED SCHÜTTE und JENNY WINTERSCHEID (2012) als Grundlage für die Normalisierung im Workflow des FOLK-Projekts dient. Hierbei ergaben sich einige Probleme bei der Bestimmung der Wortarten, da, im Gegensatz zu geschriebener Sprache, in Transkripten von Spontansprache der zur Wortartenbestimmung notwendige syntaktische Kontext nicht immer vorhanden ist.

Beispielsweise stellten sich die Fragen, ob „*gut*" und „*richtig*" im Unterrichtsgespräch Antwortpartikeln des Lehrers oder eher elliptische Adjektive sind und ob „*hoch*" in „*Hand hoch*" eher als ein Anakoluth von „*hochheben*", also als Verbzusatzpartikel oder als Adverb getaggt werden sollte. Zudem ergaben sich Probleme, die dem Aufbau des Tagsets geschuldet sind: So erwies sich die Unterscheidung zwischen attribuierend oder substituierend, die in der Spontansprache durch häufig auftretende Anakoluthe zur Interpretationssache des Hörers wird, als weitere Herausforderung bei der Korrektur des Taggings. Bei der Äußerung „*irgendwie ist das alles so ein bisschen…*" stellt sich beispielsweise die Frage, ob „*bisschen*" hier attribuierend oder substituierend ist.

3 Ergebnisse

3.1 Auswertung der manuellen Korrektur

Tabelle 1 zeigt, dass in Transkript 1 „Berufsschulinteraktion" bei einer Gesamtmenge von 3.976 Tokens 3.229 korrekt getaggt wurden, also die Fehlerquote bei 18,79 Prozent lag. In Transkript 2 „Studentisches Alltagsgespräch" wurden bei einer Gesamtmenge von 5.033 Tokens 4.096 korrekt getaggt, die Fehlerquote liegt bei diesem bei 18,62 Prozent. In Trans-

	Transkript 1	Transkript 2	Transkript 3	Gesamt
Tokens gesamt	3976	5033	2020	11029
Richtig	3229	4096	1626	8951
Richtig in %	81,21	81,38	80,5	81,16
Korrigiert	747	937	394	2078
Korrigiert in %	18,79	18,62	19,5	18,84
Richtig (Superkategorie)	3381	4295	1702	9378
Richtig in % (Superkategorie)	85,04	85,34	84,26	85,03
Korrigiert (Superkategorie)	595	738	318	1651
Korrigiert in % (Superkategorie)	14,96	14,66	15,74	14,97

Tabelle 1: Auswertung Gesamt

kript 3, der „Kind-Kind-Vorleseinteraktion“, wurden bei 2.020 Tokens 1.626 korrekt getaggt, sodass hier die höchste Fehlerquote von 19,5 Prozent vorliegt. Insgesamt beträgt die durchschnittliche Fehlerquote aller drei Transkripte 18,84 Prozent.

Bei der Fehleranalyse ist es von Interesse herauszufinden, ob die Fehler durch Zuweisung eines komplett falschen Tags oder durch eine falsche Subklassifikation entstanden sind, also das Tag zwar der richtigen Wortart zugewiesen wurde, die genauere Differenzierung derselben jedoch nicht erfasst wurde. Nimmt man die Superkategorie als Ausgangspunkt, liegt die Fehlerquote bei 14,97 Prozent. Das bedeutet, dass nur 3,87 Prozent der Tags, circa ein Fünftel der Fehler (20,55 Prozent), aufgrund der Subkategorisierung falsch getaggt wurden. Umgekehrt bedeutet dies, dass 79,45 Prozent der Fehler durch eine falsche Kategorisierung entstanden sind.

	Transkript 1	Transkript 2	Transkript 3	Gesamt
	Anzahl der Korrigierten			
V gesamt	126	115	45	286
	korrigiert zu V			
V gesamt	79	71	26	176
	korrigiert zu V in %			
V gesamt	62,70	61,74	57,78	61,54

Tabelle 2: Fehlerhafte Subkategorisierung bei Verben

Um eine fehlerhafte Subkategorisierung exemplarisch zu verdeutlichen, kann man die Auswertung der Verwechslung von Infinitiven und finiten Verben heranziehen – ein Problem von besonderer Häufigkeit, auf das auch schon SCHMID (1995) hinweist. Wie Tabelle 2 zu entnehmen ist, treten 62,7 Prozent der Fehlzuweisungen bei Verben innerhalb der Klasse der Verben auf.

Auffällig ist, dass die Werte der Fehlerquoten bei den drei Transkripten trotz deren Unterschiedlichkeit sehr nahe beieinander liegen. Eine erste Analyse zeigt, dass besonders häufig die Zieltags für Partikeln und Interjektionen nicht als solche getaggt wurden, wie aus Tabelle 3 hervorgeht. Sie verursachen mit insgesamt 51,59 Prozent die meisten Fehler beim Taggen. In Transkript 1 sind 55,56 Prozent der Fehler den Partikeln und Interjektionen geschuldet, in Transkript 2 57,84 Prozent und in Transkript 3 29,19 Prozent.

Des Weiteren geht aus den Daten hervor, dass 13,43 Prozent der Fehler durch fehlerhaftes Taggen von Pronomina, vor allem von substituierenden Demonstrativpronomina und Personalpronomina, entstanden. Ebenfalls haben Verben und Wörter, die in die Kategorie XY (Nichtwort) fallen, mit insgesamt 9,14 Prozent beziehungsweise 8,18 Prozent eine auffällig hohe Fehlerquote.

Fehlerquoten unter 5 Prozent erreichte der Tagger bei Eigennamen (4,33 Prozent), Konjunktionen (3,8 Prozent), Adverbien (3,27 Prozent), Adjektiven (3,13 Prozent), Präpositionen (1,2 Prozent), Kardinalzahlen (0,72 Prozent), Artikeln (0,63 Prozent), fremdsprachlichem Material (0,38 Prozent) und Pronominaladverbien (0,19 Prozent).

	Transkript 1	Transkript 2	Transkript 3	Gesamt
	Korrekturen in %			
Partikeln/Interjektionen	55,56	57,84	29,19	51,59
Pronomen	10,17	15,90	13,71	13,43
Verben	11,24	7,90	8,12	9,14
XY Nichtwörter	6,56	2,88	23,86	8,18
Nomen/Eigennamen	5,76	2,88	5,08	4,33
Konjunktionen	2,54	3,74	6,35	3,80
Adverbien	1,07	3,20	7,61	3,27
Adjektive	4,15	2,56	2,54	3,13
Präpositionen	1,47	0,85	1,52	1,20
Kardinalzahlen	0,13	1,17	0,76	0,72
Artikel	0,80	0,21	1,27	0,63
Fremdsprachliches Material	0,27	0,64	0,00	0,38
Pronominaladverbien	0,27	0,21	0,00	0,19

Tabelle 3: Korrekturen zu Zieltags

3.2 Analyse der Fehler in Hinsicht auf Probleme der Verwendung des TreeTaggers und des STTS mit spontansprachlichen Daten

Die Liste der Wortformen geschlossener Wortarten

Zunächst ist zu bemerken, dass das STTS und somit auch die *Liste der Wortformen geschlossener Wortarten*[12] 1995 konzipiert und seither nicht aktualisiert wurden. Da die *Liste der Wortformen geschlossener Wortarten* Teil des Lexikons ist, anhand dessen der vorliegende Tagger Informationen über die Wahrscheinlichkeiten der Zugehörigkeit bestimmter Wörter zu Wortarten bezieht, birgt sie für das Tagging von spontansprachlichen Daten verschiedene Problemquellen:

Erstens orientiert sie sich an der alten Rechtschreibung, weshalb Wörter wie „*bisschen*“ oder „*dass*“ nicht durch eine Suche im Lexikon ermittelt werden können, da sie nicht mit „*bißchen*“ oder „*daß*“ übereinstimmen.

Zweitens wurde die Liste anhand der Daten konzipiert, die bei der Analyse des Zeitungskorpus auftraten, an dem der Tagger trainiert wurde. Sie beinhaltet somit teilweise nicht die vollständigen Wortreihen der Wortklassen oder teilweise sogar fehlerhafte Einordnungen, wie ein Abgleich mit dem Duden schnell deutlich macht. Beispielsweise ist die Liste der Pronominaladverbien fehlerhaft und unvollständig. Ersteres, da „*trotzdem*“ und „*deshalb*“ nicht zu den Pronominaladverbien gehören und dennoch als solche in der Liste aufgeführt sind, letzteres, da die Liste wesentlich weniger Items enthält als die Liste der Pronominaladverbien im Duden (Duden op. 2006). Ebenso ist auffällig, dass es eine Kategorie und eine Liste für Pronominaladverbien, aber keine für Konjunktionaladverbien gibt (Duden op. 2006). Weitere Unvollständigkeiten sind:

- „*sondern*“, „*trotzdem*“, „*wo*“ und „*außer*“ fehlen in der Liste der Konjunktionen,
- „*selber*“ fehlt bei substituierenden Demonstrativpronomina (PDS),
- in Hinsicht auf spontansprachliche Daten (dialektaler Gebrauch) fehlt „*wo*“ in der Liste der substituierenden Relativpronomina (PRELS) beispielsweise „*die wo…*“ und
- „*irgendetwas*“ und „*irgendwas*“ fehlen in der Kategorie der attribuierenden Indefinitpronomina (PIAT).

Dies sind Probleme der Kategorisierung, die mit der *Liste der Wortformen der geschlossenen Wortarten* zusammenhängen. Andere hängen mit der prinzipiellen Systematisierung der Wortarten zusammen.

Pronomina

Wie bereits erwähnt, entstanden über 13 Prozent der Fehler durch das fehlerhafte Zuweisen von Tags zu Pronomina. Hier schien das Erkennen vor allem von substituierenden Relativpronomina, Personalpronomina und substituierenden Demonstrativpronomina besonders problematisch. Gründe dafür lassen sich unter anderem von der Kategorisierung des Tagsets herleiten. Es unterscheidet zwischen reflexiven und irreflexiven Personalpronomina, wobei schon in den Guidelines darauf hingewiesen wird, dass es „Überschneidungen bei *mir, dir, dich, mich, euch, uns*[13], die sowohl reflexiv als auch irreflexiv sein können“ (SCHILLER ET AL. 1999), gibt.

Auch der Duden weist auf diese Nicht-Unterscheidbarkeit hin (Duden op. 2006), weshalb eine Unterscheidung zwischen diesen Wortarten fragwürdig erscheint, zumal weder in den Guidelines, noch im Duden geklärt wird, unter welchen Bedingungen oben genannte Wörter den reflexiven oder irreflexiven Pronomina zuzuordnen sind. In Hinblick auf den Nutzen ist fraglich, ob die Unterscheidung in einer Suchanfrage an das Korpus in Bezug auf Personalpronomina hilfreich ist.

Verben

Den drittgrößten Anteil der Fehlerquote machten mit insgesamt 9 Prozent die falschen Zuweisungen der Tags für Verben aus. Ein möglicher Grund dafür zeigt sich, wenn man die im Tagset für Verben vorgesehenen Kategorien betrachtet. Das Tagset beinhaltet Kategorien für Vollverben, Auxiliarverben und Modalverben. Bei allen drei Subklassen wird weiter unterschieden zwischen finitem Verb, Infinitiv und Partizip Perfekt. Bei den Vollverben und Auxiliaren gibt es jeweils noch eine eigene Kategorie für den Imperativ, bei den Modalverben jedoch nicht. Zugegeben tritt eine solche Form im Sinne von „*du hast zu wollen*“ also „*wolle!*“ sicherlich sehr selten auf, ist jedoch denkbar und so ist es fraglich, warum keine Kategorie für VMIMP (Modalverb Imperativ) existiert.

Wesentlich problematischer erscheint jedoch, dass es eine Kategorie für eine markierte Modusform gibt, nämlich für den Imperativ, nicht aber für den Konjunktiv: Verbformen im Konjunktiv jeglicher Art werden nur als finite Verbform getaggt. Eine Einführung der Kategorie ‚Konjunktiv‘ würde in einem Korpus für gesprochenes Deutsch Zugriff auf weitere Informationen ermöglichen.

Kategorie XY

Ein weiteres Kategorisierungsproblem birgt die Kategorie XY, die Nichtwörter. Hier stellt sich zunächst die Frage, was als Nichtwort zu definieren ist. SCHILLER ET AL. (1995) geben in ihren Guidelines keine Definition von Nichtwörtern, vielmehr beschreiben sie, dass das Tag „vor allem bei größeren Symbolgruppen, Nichtwörtern sowie Kombinationen aus Ziffern und Zeichen, die sich nicht als CARD (Kardinalzahlen) oder ADJA (attributives Adjektiv) einordnen lassen“ (ebd.) vergeben wird, mit besonderem Vermerk, dass „Nichtalphabetische Zeichen (§, ©, $ etc.), römische Zahlzeichen etc. [...] so zu taggen [seien], wie das ausgeschriebene Wort getaggt würde, in Analogie zu Abkürzungen“ (ebd.).

Dies birgt gleich zwei Probleme. Erstens werden nach den Normalisierungskonventionen im FOLK-Projekt „Abgebrochene Wörter, die nicht zweifellos rekonstruierbar sind“, „idiolektale Wörter“ und „nicht lexikalisierte Laute“ (DEPPERMANN ET AL. 2012) durch die Sonderzeichen %, § bzw. # markiert (ebd.). Wie aber bereits zitiert, werden solche ‚nichtalphabetischen Zeichen‘ behandelt, als seien sie ausformulierte Wörter, das heißt sie werden als die Wörter ‚Prozent‘, ‚Paragraph‘ bzw. ‚Raute‘ interpretiert. Der Tagger wird also zwangsläufig allen so markierten Wörtern das Tag NN, Nomen, zuweisen. In mehr als einem Viertel aller Fälle, in denen fälschlicherweise das Tag NN zugewiesen wurde, wurde es manuell zu XY korrigiert.

Zweitens birgt die oben zitierte Ausführung noch immer keine Definition dessen, was der Kategorie Nichtwort zuzuordnen ist. In Grammatiken und linguistischen Aufsätzen ist keine Definition zu finden. In der psycholinguistischen und neurolinguistischen Forschung wird Nichtwort, oder Pseudowort, als eine Graphem- oder Phonem-Abfolge bezeichnet, die kein bekanntes Lexem einer bestimmten Sprache formt (HARLEY 2008). Nach dieser Definition würden typisch gesprochensprachliche Phänomene wie Abbrüche sowie Phänomene, die Besonderheiten des Transkriptionsprozesses geschuldet sind, wie beispielsweise für den Transkribenten unverständliche Äußerungen oder die Übertragung von Buchstabiertem oder silbischem Lachen in die Transkription, als Nichtwörter bezeichnet werden müssen. Würde man alle diese Phänomene jedoch unter die Kategorie XY fassen, wie es zunächst geschehen ist, wäre dies in zweierlei Hinsicht problematisch. Einerseits kommt es zu einer hohen Fehlerquote, da für den Tagger Nichtwörter nur sehr schwer zu erfassen sind – mehr als 8 Prozent aller Fehler sind durch fehlerhafte Tagzuweisung der Kategorie XY entstanden. In dem Transkript der Kind-Kind-Vorleseinteraktion machte sie fast ein Viertel der Fehler aus. Dies liegt daran, dass der Tagger durch Mangel eines Abgleichs im Lexikon nach dem ‚default entry‘ das Tag allein aufgrund der Wahrscheinlichkeitsberechnung durch die zwei vorhergehenden Tags berechnet. In den Wahrscheinlichkeitsberechnungen dürften Nichtwörter jedoch äußerst geringe Wahrscheinlichkeiten haben, da sie in dem Zeitungskorpus nur sehr selten vorkamen.

Das zweite Problem, das aus einer solchen Kategorisierung entsteht, ist, dass viele Informationen über das mit XY getaggte Wort nicht erfasst werden können. Gerade weil Abbrüche und Korrekturen ein so prominentes Phänomen der gesprochenen Sprache sind, sollten sie in einem Korpus des gesprochenen Deutsch als solche markiert werden und nicht der Restkategorie XY zugewiesen werden. Auch gehen Informationen über das verloren, was in

den meisten Fällen als Buchstabiertes erscheint, wie beispielsweise dass „*i ce e*“ für ICE (Inter City Express) im Prinzip ein Nomen ist (mit Ausnahme von ‚echt‘ buchstabierten Äußerungen wie „*wer nämlich mit* ha *schreibt ist dämlich*“). Weiterhin ist es problematisch, unverstandene Äußerungen als Nichtwörter zu bezeichnen, da sie höchstwahrscheinlich doch geäußerte Wörter waren, die der Transkribent nur nicht verstanden hat.

Zusammenfassend kann man sagen, dass auch im Falle der Nichtwörter die Kategorisierung im STTS in Hinblick auf Phänomene gesprochener Sprache unzureichend ist.

Partikeln und Interjektionen

Wie bereits erwähnt, entstanden über 50 Prozent der Taggingfehler durch Nicht-Erkennung von Partikeln. Die zehn prominentesten waren „*ja, äh, halt, mal, hm, aber, so, doch, also, gut*“, und „*einfach*“. Wendet man sich der Wortart Partikel zu, so trifft man schnell auf verschiedene Taxonomien nach unterschiedlichen Kriterien: nach grammatischen Kriterien, funktionalen Unterscheidungen oder nach ihrer Sequenzstruktur. Da das Wortfeld sehr groß ist, sind die Taxonomien meist nicht für alle auftretenden Formen geeignet, bei anderen werden bestimmte Wörter doppelt kategorisiert. Beispielsweise verfolgt die Duden-Grammatik einen funktionalen Ansatz (Duden op. 2006), JOHANNES SCHWITALLA hingegen einen Ansatz der Kategorisierung nach ihrer Sequenzstruktur im Gespräch (SCHWITALLA 2002) und in der „Grammatik der deutschen Sprache“ nimmt LUDGER HOFFMANN eine Einteilung nach funktionalen und vor allem distributionellen Kriterien vor (HOFFMANN 1997).

Im *Stuttgart-Tübingen-Tagset* (STTS) sind folgende Kategorisierungen der Partikeln vorgenommen worden (STTS-Tagtable 1995/1999):

PTKZU: ‘zu’ vor Infinitiv, beispielsweise *zu [gehen]*
PTKNEG: Negationspartikel, beispielsweise *nicht*
PTKVZ: abgetrennter Verbzusatz, beispielsweise *[er kommt] an, [er fährt] rad*
PTKANT: Antwortpartikel, beispielsweise *ja, nein, danke, bitte*
PTKA: Partikel bei Adjektiv oder Adverb, beispielsweise *am [schönsten], zu [schnell]*

Eine weitere Kategorie existiert für Interjektionen:

ITJ: Interjektion, beispielsweise *mhm, ach, tja*

Schon auf den ersten Blick wird deutlich, dass die Kategorisierung, wie sie hier vorgenommen wurde, keiner der oben genannten Taxonomien entspricht. Abgesehen von der Antwortpartikel und der Negationspartikel schließt die Kategorisierung jegliche Gesprächspartikeln aus, beispielsweise Lautmalerei, Hesitationspartikeln (gefüllte Pausen), Backchannelpartikeln und Interjektionen. Zu letzteren gibt es, im Gegensatz zu den anderen, zwar eine Kategorie im STTS und damit verbundene Einträge im Lexikon, diese sind jedoch auf solche begrenzt, die in den manuell getaggten Zeitungsartikeln vorkamen. Sie beschränken sich auf einige wenige, die JOHANNES SCHWITALLA als „primäre“ Interjektionen bezeichnet (SCHWITALLA 2006). „Sekundäre“ Interjektionen, die sich aus Lexemen ableiten wie z. B. „*mist*“ oder „*gott*“ können nicht als solche erkannt werden, da keine Wahrscheinlichkeitswerte dafür vorliegen und sie so nur als Nomen getaggt werden können.

Ebenso verhält es sich mit den anderen Interjektionen und Partikeln. Eintragungen in der Liste der *Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset* in Bezug auf Partikeln beschränken sich auf „*allzu*“, „*am*“ und „*zu*“ für die Kategorie PTKA (Partikel bei Adjektiv oder Adverb), „*bitte*“, „*danke*“, „*doch*“, „*ja*“, „*nein*“ für die Kategorie PTKANT (Antwortpartikel), „*nicht*“ als PTKNEG (Negationspartikel). Zudem sind dort Partikeln, die als abgetrennte Verbzusätze fungieren und „*zu*“ für die Kategorie PTKZU (‚zu' vor Infinitiv) aufgelistet. Dass Modalpartikeln bzw. Abtönungspartikeln bewusst nicht in die Kategorisierung aufgenommen wurden, zeigen Beispiele der „Vorläufigen Guidelines für das Tagging deutscher Textcorpora mit STTS“. (SCHILLER ET AL. 1999) Hier werden Abtönungspartikeln explizit der Wortklasse der (echten) Adverbien zugeordnet. Über Modalpartikeln wird keine Aussage gemacht.

Da die Kategorie ADV (Adverb) offensichtlich als ‚Restkategorie' genutzt wurde[14], ist es nicht verwunderlich, dass 35 Prozent aller korrigierten Wörter als Adverb getaggt waren. In circa 86 Prozent der Fälle, in denen fälschlicherweise das Tag ADV vergeben wurde, wurde es zu der Wortart Partikel korrigiert. Die größte Problemquelle des Taggens spontansprachlicher Daten mit dem STTS ist also dahingehend zu identifizieren, dass eine unzulängliche Kategorisierung der Partikeln vorhanden ist und solche folglich auch gar nicht korrekt getaggt werden können; ein Tag für die Oberkategorie Partikel gibt es im STTS nicht. Bei der manuellen Korrektur der automatisiert getaggten Daten wurde jedoch jedem Wort, das nach der Definition des Dudens eine Partikel jedweder Art ist, das Tag PTK zugewiesen.

Wie ist mit dieser problematischen Kategorie umgehen? Partikeln als Adverbien zu taggen scheint nicht plausibel. In keiner der beschriebenen Kategorisierungen von Partikeln werden diese als Adverbien bezeichnet. Um den Erscheinungsformen gesprochener Sprache gerecht zu werden, ist es also notwendig, das Tagset um einige Kategorien zu erweitern. Hierbei sollte vor allem auf die Umsetzbarkeit, das heißt die Erkennbarkeit der Wortart für den Tagger, geachtet werden – eine zu differenzierte Spezifizierung könnte einerseits zu weiteren Fehlern führen, andererseits müsste man sich dann zwangsläufig einer Richtung der Spezifizierung anschließen. Mit Letzterem würde man jedoch Interpretation an das Korpus herantragen. In Anbetracht der Tatsache, dass man nicht weiß, nach welchen Gesichtspunkten Forscher das Korpus durchsuchen wollen, wäre dies von Nachteil. Prämisse ist also, Partikeln zwar als solche zu taggen, dabei jedoch die Präzision der Ergebnisse hochzuhalten und das Tagging so interpretationsfrei wie möglich zu gestalten.

Im Rahmen des *KiezDeutsch-Korpus* (KiDKo) haben INES REHBEIN, SÖREN SCHALOWSKI und HEIKE WIESE schon Vorschläge für eine Anpassung des Tagsets an gesprochene Sprache formuliert. Sie unterscheiden zwischen Rückversicherungs-, Backchannel-, Hesitations- und Antwortpartikeln sowie Onomatopoetika und unspezifischen Partikeln (REHBEIN ET AL. 2012). Eine solche differenzierte Unterscheidung, vor allem in Bezug auf die ersten zwei Kategorien, ist durch ein automatisiertes Tagging nicht möglich. Die bereits erwähnte Segmentierungsproblematik und fehlende Interpunktion[15] machen es für den Tagger unmöglich, Informationen darüber abzuleiten, ob das Wort am Anfang, in der Mitte oder am Ende einer Äußerung steht. Nimmt man das Wort „*ja*“, so kann es beispielsweise sowohl als Rückversicherungspartikel, als Responsiv, als Abtönungspartikel oder als Antwortpartikel fungieren, je nachdem, wo und wie es in der Äußerung verwendet wird.

Für ein automatisiertes Tagging ist eine Differenzierung also nur an solchen Stellen sinnvoll, wo es eindeutige Formen gibt. Dies ist zum Beispiel bei Hesitationspartikeln der Fall. Nach den Normalisierungskonventionen im FOLK-Projekt werden alle Formen der gefüllten Pause zu „*äh*“ normalisiert. Es ist also möglich, einen Lexikon-Eintrag zu erstellen, der die Information beinhaltet, dass „*äh*“ mit hundertprozentiger Wahrscheinlichkeit das Tag PTKFILL zugewiesen wird. Für alle anderen Formen von Partikeln, die noch nicht in den Tagset-Kategorien enthalten sind, ließe sich das Tag PTK, also unspezifizierte Partikel, verwenden. Die Kategorie müsste dafür in die Liste der *Wortformen der geschlossenen Wortarten* aufgenommen werden. Gleichermaßen könnte mit Interjektionen verfahren werden.

Unabhängig von der Aufnahme der Kategorie in die Liste der *Wortformen geschlossener Wortarten* stellt sich die Frage, ob Interjektionen zukünftig als Unterkategorie von Partikeln behandelt werden und sie demnach das Tag PTKITJ erhalten sollten. Es wird vorgeschlagen, auch hier der Duden-Grammatik zu folgen und diese Änderung vorzunehmen. Allerdings ist eine solche Entscheidung in Hinblick auf die bereits geschilderte Definitionsproblematik diskutierbar.

Zusammenfassend kann man sagen, dass die Ursache für das größte Problem des Taggens von normalisierten spontansprachlichen Daten in der unzureichenden Kategorisierung des STTS in Hinsicht auf Partikeln liegt. Nicht nur, weil diese in konzeptionell gesprochener Sprache häufiger vorkommen als in konzeptionell geschriebener, sondern auch, weil die Tag-Zuweisung von Partikeln vom TreeTagger mit dem STTS nur bei einigen wenigen Spezialformen funktioniert. Findet man für dieses Problem eine Lösung, könnte man die Fehlerquote um die Hälfte reduzieren.

Fazit

Das automatisierte POS-Tagging von spontansprachlichen Daten mit dem TreeTagger und dem STTS – ohne manuelle Nachkorrektur – erreicht im Durchschnitt eine Präzision von 81,16 Prozent. Die Ursachen für die hohe Fehlerquote liegen darin, dass das Tagset und der Tagger für konzeptionell geschriebene Sprache entwickelt bzw. trainiert worden sind und das Tagging-Ergebnis somit aufgrund der Unterschiede von gesprochener und geschriebener Sprache an Präzision stark einbüßt. Viele Tagging-Fehler sind vor allem der Problematik der Segmentierbarkeit gesprochener Sprache in Hinblick auf fehlende Annotation syntaktischer Einheiten, Kontextabhängigkeit sowie den Unterschieden zwischen gesprochener und geschriebener Sprache geschuldet. In Bezug auf Letzteres sind vor allem das häufige Vorkommen von Gesprächspartikeln, die andersartige Verwendung von Pronomina und Verbformen sowie Phänomene, die der Kategorie XY zugewiesen werden müssen, die größten Problemquellen.

Für das Taggen des FOLK müssen also der Tagger und das Tagset für spontansprachliche Daten optimiert werden, damit das Ziel erreicht werden kann, die Präzision des automatisierten POS-Taggings auf mindestens 95 Prozent zu steigern. Um eine Verbesserung herbeizuführen werden folgende Vorschläge gemacht:

1. Die Einführung neuer Kategorien im Tagset, die den Eigenheiten und der Transkription von gesprochener Sprache gerecht werden,

2. Änderungen der Normalisierungskonventionen für den FOLK-Normalisierungsprozess, damit eine präzise Zuordnung der Tags zu Phänomenen gesprochener Sprache möglich wird,
3. das Einführen eines Post-Processings, das die Zuweisung bestimmter Tags zu den in der Normalisierung markierten Phänomenen möglich macht,
4. die Überarbeitung der Liste der Wortformen geschlossener Wortarten in Bezug auf ihre Aktualität, Vollständigkeit und Richtigkeit
5. und ein Neutraining des TreeTaggers an normalisierten spontansprachlichen Daten.

In Bezug auf den ersten Punkt, die Überarbeitung der Kategorisierung im STTS, wird Folgendes vorgeschlagen:

- Die Einführung eines Tags für unverständliche Äußerungen mit der Bezeichnung UI (uninterpretierbar),
- die Neukategorisierung der Tags für Partikeln mit der Einführung des Tags PTK für Partikeln jeglicher Art sowie PTKFILL für Hesitationspartikeln und PTKITJ für Interjektionen,
- eine Überarbeitung der Kategorisierung der Tags für Verben in Hinblick auf die Berücksichtigung des Modus, im Besonderen auf die Einführung eines Tags für imperativisch gebrauchte Modalverben und generell die Einführung eines Tags für konjunktivisch gebrauchte Verben,
- die Einführung eines Tags für Abbrüche, das bei nicht-rekonstruierbaren Abbrüchen allein steht und bei rekonstruierbaren Abbrüchen dem Wortarten-Tag hinzugefügt wird, beispielsweise mit der Bezeichnung AB,
- die Einführung eines Tags für Buchstabiertes, um ‚echt' Buchstabiertes von Akronymen zu unterscheiden
- und schließlich die Einführung eines Tags für Konjunktionaladverbien, beispielsweise mit der Bezeichnung KAV.

Zuerst sollte also das Tagset in dieser Weise überarbeitet sowie die Liste der Wortformen geschlossener Wortarten in Hinblick auf ihre Aktualität, Vollständigkeit und Richtigkeit unter Einbezug der neu vorgeschlagenen Kategorien aktualisiert werden. Vor allem ist eine Anpassung an moderne Rechtschreibung und die Ergänzung und Korrektur von den jeweiligen Listen der Wörter geschlossener Wortarten durch die des Dudens notwendig. Dies betrifft vor allem die Listen der Pronominaladverbien, Demonstrativpronomina, Indefinitpronomina und Konjunktionen. Bei Letzteren sollten vor allem „*weil*", „*obwohl*" und „*wobei*" auch in die Liste der nebenordnenden Konjunktionen aufgenommen werden, da sogar im Duden im Kapitel der Grammatik der gesprochenen Sprache belegt ist, dass sie auch nebenordnend sein können (Duden op. 2006). Außerdem sind die Abkürzungen „*d. h.*", „*bzw.*" und „*z. B.*" aus der Liste der Konjunktionen zu entfernen. Weiterhin müsste, um den Eigenheiten der gesprochenen Sprache gerecht zu werden, die Liste der Personalpronomina um „*der*", „*die*" und „*das*" erweitert werden, da sie in der Umgangssprache häufig als solche verwendet werden (BARBOUR & STEVENSON 1998).

Darauf aufbauend können durch ein Post-Processing bestimmten Tokens die korrekten Tags nachträglich zugewiesen werden. Dies ist immer dann sinnvoll, wenn Wortarten eindeutige Formen zugeordnet werden können. Dies greift in gewisser Hinsicht auch für die

Items in der Liste der *Wortformen geschlossener Wortarten*. Es ist wahrscheinlich, dass man die Fehlerquote fast um die Hälfte reduzieren kann, wenn man eine Liste der häufigsten Gesprächspartikeln und Modalpartikeln erstellt und ihnen im Post-Processing das Tag PTK zuweist. Analog dazu kann mit vielen Items der Liste der *Wortformen geschlossener Wortarten* verfahren werden. Um Buchstabiertes und Abbrüche als solche erkennbar zu machen, können sie in der Normalisierung mit einem Sonderzeichen versehen werden. Dieses kann im Post-Processing dann zu einer Zuweisung der genannten Tags dienen.

Ein weiterer Schritt ist das Neutraining des Taggers. Dies soll in naher Zukunft, unter Einbezug der in dieser Arbeit vorgeschlagenen neuen Tags und Kategorisierungen, an noch weiteren manuell korrigierten Transkripten vorgenommen werden. Es ist anzunehmen, dass auf diese Weise verschiedene Eigenheiten gesprochener Sprache statistisch repräsentiert werden und ein weiteres automatisiertes Taggen spontansprachlicher Daten daraufhin deutlich bessere Ergebnisse erzielt.

---

1 Das gesamte Transkript sowie die Metadaten dazu sind abrufbar auf der Webseite der DGD http://dgd.ids-mannheim.de unter dem Transkriptcode: FOLK_E_00001_SE_01_T_01_DF_01.

2 Transkript und Metadaten auf o.g. Webseite unter dem Transkriptcode: FOLK_E_00046_SE_01_T_01_DF_01.

3 Transkript und Metadaten auf o.g. Webseite unter dem Transkriptcode: FOLK_E_00076_SE_01_T_01_DF_01.

4 Zugang zum FOLK Korpus, den Transkripten sowie den Metadaten ist nach einer Registrierung auf o.g. Webseite möglich.

5 FOLKER ist ein Programm zur computergestützen Transkription spontansprachlicher Daten, das von Thomas Schmidt speziell für die Arbeit und den Workflow des FOLK-Projekts entwickelt wurde. (Schmidt 2012).

6 GAT – Gesprächsanalytisches Transkriptionssystem, ist ein System, das Konventionen für das Erstellen gesprächsanalytischer Transkriptionen aufstellt. cGAT ist größtenteils konform zu GAT 2, der aktuellsten Form der Konventionen, jedoch mit einigen Abweichungen für die Umsetzung der Transkription mit Hilfe von Computer gestützten Verfahren.

7 GAT 2 sieht Konventionen für drei Detailliertheitsstufen der Transkripte vor: das Minimal-, Basis- und Feintranskript. Siehe Selting et al. 2009.

8 Literarische Umschrift bedeutet, dass zwar im lateinischen Alphabet transkribiert wird, jedoch dabei die Aussprache des Sprechers möglichst ‚lautungsgetreu' dargestellt werden soll, beispielsweise wenn ein süddeutscher Sprecher „weisch" sagt anstelle von „weißt du" wird es wie erstere Form auch transkribiert.

9 OrthoNormal ist ein Programm, das, wie FOLKER, ebenfalls für die Arbeit und den Workflow des FOLK-Projekts von Thomas Schmidt entwickelt wurde. Es nimmt eine in FOLKER erstellte Transkription als Grundlage und ermöglicht Wort für Wort die Annotation der orthographisch korrekten Form. (Schmidt 2012).

10 Das Programm speichert in einer Datenbank eingegebene Korrekturen, sodass Formen, die häufig zu anderen Formen korrigiert wurden, automatisiert normalisiert werden. Eine manuelle Korrektur ist jedoch in jedem Fall notwendig.
11 Für die manuelle Korrektur wurde die Version 0.6 verwendet.
12 Die Liste der Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset dient dem Tagger für den Lexikon-Abgleich. Er erhält dort Informationen über Wahrscheinlichkeiten für die Zugehörigkeit von Items zu geschlossene Wortarten.
13 Hervorhebung durch den Autor.
14 Als Adverbien werden laut Guidelines auch Ordinalzahlen, Präpositionen+einander, Abkürzungen wie „z. B." oder „bzw." und Multiplikativzahlen getaggt. (Schiller et al. 1999).
15 Im KiDKo-Projekt wurde nach den Konventionen des Systems der halbinterpretativen Arbeitstranskription (HIAT) transkribiert. HIAT ist ebenso wie GAT 2 ein System von Konventionen zur Verschriftlichung gesprochener Sprache, unterscheidet sich jedoch in einigen Punkten von GAT, beispielsweise durch interpretative Annotation syntaktischer Abschlusspunkte. Daher kann erstens aus der Stellung der Partikel im Satz Information über ihre Funktion abgeleitet werden, zweitens kann bei manueller Annotation der Erfahrungshorizont des Annotators ausreichende Information bieten. Im FOLK-Projekt wird, wie bereits erwähnt, nach cGAT, einer Modifikation des Gesprächsanalytischen Transkriptionssystems 2 transkribiert, in dem es keine Annotation von Interpunktion gibt (Selting et al. 2009) und (Schmidt & Schütte 2011). Da automatisiert getaggt wird, kann ein Informationsbezug über die Stellung im Satz nicht hergestellt werden.

Literatur

BARBOUR, S.; STEVENSON, P. (1998). Variation im Deutschen. Soziolinguistische Perspektiven. Berlin [u.a.]: De Gruyter.

Duden. Die Grammatik. Unentbehrlich für richtiges Deutsch (op. 2006). 7. Aufl. Mannheim: Dudenverlag.

DEPPERMANN, A.; WINTERSCHEID, J.; SCHÜTTE, W. (2012). Regeln für die orthografische Transkription mit OrthoNormal.

ECKART, R. (2013). TreeTagger for Java (TT4J). Online verfügbar unter https://code.google.com/p/tt4j/, zuletzt geprüft am 21.11.2013.

HARLEY, T. A. (2008). The Psychology of Language. From Data to Theory. 3. Aufl. Hove, East Sussex, UK: Psychology Press. Online verfügbar unter http://media.routledgeweb.com/pp/common/sample-chapters/9781841693828.pdf, zuletzt geprüft am 29.12.2012.

HOFFMANN, L. (1997). "Interjektionen und Responsive." In: Zifonun, G. et al. (Eds.) (1997). Grammatik der deutschen Sprache. Berlin [etc.]: De Gruyter, 360–408.

INSTITUT FÜR DEUTSCHE SPRACHE (Hg.). DGD. Datenbank für Gesprochenes Deutsch. Online verfügbar unter http://dgd.ids-mannheim.de, zuletzt geprüft am 15.12.2013.

INSTITUT FÜR DEUTSCHE SPRACHE (Hg.) (2012). FOLK. Forschungs- und Lehrkorpus Gesprochenes Deutsch. Online verfügbar unter http://agd.ids-mannheim.de/folk.shtml, zuletzt aktualisiert am 29.10.2013, zuletzt geprüft am 15.12.2013.

REHBEIN, I. ET AL. (2012). Annotating spoken language. POTSDAM UNIVERSITY. LREC 2012 'Workshop Best Practices for Speech Corpora in Linguistic Research', Hamburg. Online verfügbar unter http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf, zuletzt aktualisiert am 21.05.2012, zuletzt geprüft am 15.12.2013.

SCHILLER, A. ET AL. (1999). Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset). INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG (STUTTGART); UNIVERSITÄT TÜBINGEN SEMINAR FÜR SPRACHWISSENSCHAFT (TÜBINGEN). Online verfügbar unter ftp://ftp.ims.uni-stuttgart.de/pub/corpora/stts_guide.pdf, zuletzt geprüft am 15.12.2013.

SCHMID, H. (1995). Improvements In Part-of-Speech Tagging With An Application To German. INSTITUT FÜR MASCHINELLE SPRACHVERARBEITUNG (STUTTGART). Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. Online verfügbar unter ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf, zuletzt geprüft am 15.12.2013.

SCHMIDT, T. (2012). EXMARaLDA and the FOLK tools. In: Proceedings of the Language Resource and Evalutation Conference (LREC). Istanbul, Paris: ELRA.

SCHMIDT, T. & SCHÜTTE, W. (2011). FOLKER Transkriptionseditor für das "Forschungs- und Lehrkorpus gesprochenes Deutsch" (FOLK). Transkriptionshandbuch. INSTITUT FÜR DEUTSCHE SPRACHE; ARCHIV FÜR GESPROCHENES DEUTSCH. Online verfügbar unter http://agd.ids-mannheim.de/download/FOLKER-Transkriptionshandbuch.pdf, zuletzt aktualisiert am 02.09.2011, zuletzt geprüft am 15.12.2013.

SCHWITALLA, J. (2002). "Kleine Wörter. Partikeln im Gespräch." In: Dittmann, J. & Schmidt, C. (Eds.) (2002). Über Wörter: Grundkurs Linguistik. Freiburg im Breisgau: Rombach, 259–281.

SCHWITALLA, J. (2006). Gesprochenes Deutsch. Eine Einführung. 3. Aufl. Berlin: Erich Schmidt.

SELTING, M. ET AL. (2009). "Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)." In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* (10), 353–402. Online verfügbar unter http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf, zuletzt geprüft am 15.12.2013.

STTS-Tagtable (1995/1999). Online verfügbar unter http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html, zuletzt geprüft am 16.12.2013.

Thomas Bartz, Michael Beißwenger, Angelika Storrer

Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge

1 Einleitung

Die Erschließung, Dokumentation und Aufbereitung von Sprachdaten aus Genres internetbasierter Kommunikation für die Zwecke der korpusgestützten empirischen Sprachanalyse stellt gegenwärtig eine große Herausforderung für den Bereich der Korpuslinguistik und Sprachbeschreibung dar (Beißwenger & Storrer 2008): Die Besonderheiten der schriftlichen Sprachverwendung in E-Mails, Chats, Online-Foren, Twitter, Wikipedia-Diskussionsseiten, Weblog-Kommentaren, sozialen Netzwerken, Instant-Messaging-Anwendungen oder Online-Computerspielen (MMORPGs) lassen sich, wie die neuere linguistische Forschung in diesem Bereich gezeigt hat, weder mit Kategorien und Modellen für die Analyse geschriebener Texte noch mit Kategorien und Modellen für die Analyse mündlicher Gespräche zufriedenstellend beschreiben. Auch computerlinguistische Verfahren für die automatische Analyse und Annotation geschriebener Sprache lassen sich zum gegenwärtigen Stand nur sehr bedingt für die Aufbereitung von Sprachdaten aus Genres internetbasierter Kommunikation heranziehen. Darüber hinaus existieren bis dato keinerlei Standards für die Strukturannotation und Repräsentation solcher Daten in Korpora. Korpora internetbasierter Kommunikation stellen neben den Text- und Gesprächskorpora einen „Korpustyp dritter Art“ (Storrer i.Dr.) dar, für deren Aufbau geeignete Standards, Verfahren und Gütekriterien erst noch entwickelt werden müssen.

Weshalb die Sprachverwendung in der internetbasierten Kommunikation neuer, dem Untersuchungsgegenstand angemessener Beschreibungs- und Analyseansätze bedarf, wurde in der linguistischen Forschung zum Thema bereits ausführlich theoretisch begründet und empirisch gezeigt (vgl. z.B. Herring 1996, 1999, 2010/11; Beißwenger 2000, 2007; Crystal 2001, Storrer 2001, Storrer i.Dr.; Bittner 2003; Schönfeldt & Golato 2003). Die entsprechenden Befunde sind aber bislang kaum in korpus- und computerlinguistische Ansätze zur Behandlung von Sprachdaten aus diesem Kommunikationsbereich eingeflossen. Erste Schritte, die linguistische Theoriebildung zu diesem Bereich mit korpus- und computerlinguistischen Beschreibungsansätzen zusammenbringen, werden gegenwärtig im europäischen Netzwerk „Building and Annotating CMC Corpora“[1] sowie in der Special Interest Group „Computer-Mediated Communication“ im Rahmen der *Text Encoding Initiative* (*TEI*)[2] unternommen; ein erster Entwurf zu einem Annotationsschema für Genres internetbasierter Kommunikation, das linguistische Ergebnisse zum Gegenstand mit einem existierenden Repräsentationsstandard im Bereich der E-Humanities zusammenführt, ist in Beißwenger et al. (2012) beschrieben.

Werkzeuge und Verfahren für die automatische Wortartenannotation (im Folgenden auch: *Part-of-speech-Tagging*, *POS-Tagging*) stellen neben Repräsentationsstandards ein weiteres

grundlegendes Desiderat beim Aufbau linguistisch aufbereiteter Korpora zur Sprachverwendung in der internetbasierten Kommunikation dar (Beißwenger & Storrer 2008: 302f., King 2009: 312f.): Existierende Verfahren, die in der Regel auf Korpora mit redigierten Texten (Zeitungstexten u.Ä.) trainiert wurden, lassen sich zum gegenwärtigen Stand der Kunst nicht ohne deutliche Einbußen bei der Genauigkeit der Klassifikation auf schriftliche Sprachdaten aus E-Mails, Chats, Online-Foren, Twitter, Wikipedia-Diskussionsseiten, Weblog-Kommentaren, sozialen Netzwerken oder Instant-Messaging-Anwendungen anwenden.

Giesbrecht & Evert (2009) zeigen in einem systematischen Vergleich der Performanz verschiedener Wortartentagger, dass sich die Zuverlässigkeit der Wortartenzuordnung drastisch verschlechtert, wenn anstelle von Texten mit redigierter Schriftlichkeit Webkorpora verarbeitet werden. Korpusausschnitte mit geringerer Konformität zum geschriebenen Standard (z.B. Postings aus Online-Foren, die Giesbrecht & Evert entsprechend als „hard genres" bezeichnen) verursachen dabei größere Probleme als Dokumente mit höherer Standardkonformität („easy genres"). Bick (2010) kommt zu ähnlichen Befunden; als Beispiele für Phänomene, die in der E-Mail- und in noch stärkerem Maße in der Chat-Kommunikation Verarbeitungsprobleme verursachen, benennt er u.a. kontraktierte Formen (*dont*, *gotta*), Fälle der graphischen Nachbildung von Phänomenen der gesprochenen Sprache („phonetic writings": *Ravvvvvvvvvvvvveeee*), netztypische Akronyme sowie „subject-less sentences" wie *dances [around] wild and naked*. Weitere Herausforderungen für die automatische Analyse ergeben sich durch Schnellschreibphänomene (Tippfehler, Wortauslassungen, Verzicht auf normkonforme Großschreibung) sowie durch Einheiten, die keine direkten Pendants in Genres redigierter Schriftlichkeit haben; zu nennen wären hier u.a. Emoticons, Adressierungen (*@bienchen*, *@alle*) sowie Aktionswörter wie *grins*, *lach*, *lol*, *stirnrunzel* oder *diestirninfaltenleg*.

Mit existierenden, für linguistische Anwender direkt nutzbaren Verarbeitungswerkzeugen können Phänomene dieser Art bislang nur unzureichend analysiert und disambiguiert werden. Entsprechend sind korpusgestützte Analysen zu den sprachlichen Besonderheiten und zur sprachlichen Variation in Genres internetbasierter Kommunikation bislang noch mit großem Aufwand verbunden: Entweder kann sich die Analyse nur auf Rohdaten oder auf einige wenige in die Daten eingebrachte Annotationen stützen oder die verwendeten Korpora müssen für Analysezwecke vorab zeit- und kostenintensiv handannotiert werden.

Inzwischen gibt es zumindest für das Englische und das Niederländische allerdings erste Arbeiten, in denen Verfahren zur Wortartenannotation auf einzelne Genres internetbasierter Kommunikation angepasst wurden. So wurden in Ritter et al. (2011), Gimpel et al. (2011) und Owoputi et al. (2012) Wortarten-Tagsets und darauf bezogene Verarbeitungswerkzeuge für die Annotation englischer Twitter-Daten optimiert; die zugrunde gelegten Tagsets enthalten Kategorien für Hashtags, Adressierungen („@-mentions"), Emoticons, URLs und Retweet-Marker (*RT*), bei Gimpel et al. (2011) auch für konktraktierte Formen wie *someone's*, *I'm*, *let's*, *book'll*, *Mark'll*. Avontuur et al. (2012) nutzen die in Gimpel et al. (2011) beschriebenen Twitter-spezifischen Tagset-Erweiterungen in Kombination mit einem existierenden Tagset für das Niederländische für die Annotation niederländischer Tweets. Die Ergebnisse der auf der Grundlage dieser Tagsets entwickelten automatischen Verfahren

sind ermutigend; die Entwicklung vergleichbarer Verfahren für das Deutsche und unter Einbeziehung auch anderer Genres als nur Twitter sowie weiterer sprachlicher Phänomene ist aber noch zu leisten. Speziell mit Blick auf die korpusgestützte linguistische Analyse sprachlichen Wandels durch internetbasierte Kommunikation ist es zudem wünschenswert, netztypische sprachliche Innovationen nicht einfach als neue, „online-spezifische" (Gimpel et al. 2011: 2) Kategorien in die verwendeten Tagsets einzufügen, sondern sie in einen wortartentheoretischen Beschreibungsrahmen zu integrieren und – wo begründbar – als online-spezifische Erweiterungen vorhandener Kategorien darzustellen. In Beißwenger et al. (2012: 3.5.1) wird die Möglichkeit einer solchen Integration für Emoticons, Aktionswörter und Adressierungen aufgezeigt.

Welche Potenziale sich der Variations- und Medienlinguistik durch die Analyse großer annotierter Korpora zur internetbasierten Kommunikation bieten, zeigen exemplarisch die Untersuchungen von Bick (2010) zu Mündlichkeitsphänomenen und grammatischen Merkmalen in einem E-Mail- und Chat-Korpus mit insgesamt 117 Millionen Tokens. An welche Kapazitätsgrenzen man stößt, wenn man die Analyse eines großen Korpus zur internetbasierten Schriftlichkeit ohne die Möglichkeit zur Suche über linguistischen Kategorien (Annotationen) durchführen muss, belegt die quantitative Untersuchung von Storrer (2013) zu Frequenzunterschieden bei der Verwendung von Emoticons auf den Artikel- und Diskussionsseiten der deutschen Wikipedia: Obwohl anhand formaler Merkmale recht gut per Volltextsuche identifizierbar, gibt es in der Wikipedia zu Emoticons homonyme Zeichenkombinationen, die selbst keine Emoticons sind. Je fünfstellige Trefferzahlen für die Emoticon-Formen *;-)*, *:-)* und *:)* lassen sich aber intellektuell nicht mehr mit vertretbarem Aufwand disambiguieren; entsprechend ist in solchen Fällen eine Differenzierung nach falsch positiven Treffern (Pseudotreffern) und echten Belegen ohne die Möglichkeit zur Suche über Kategorien nicht mehr möglich. Emoticons stellen hierbei noch einen vergleichsweise „einfachen" Fall dar; für die empirische Analyse komplexerer sprachlicher Phänomene dürften, wenn lediglich die Möglichkeit einer Volltextsuche zur Verfügung steht, die methodischen Probleme noch ungleich größer sein.

Linguistisch annotierte Korpora mit Sprachdaten aus Genres internetbasierter Kommunikation stellen dabei nicht nur ein Desiderat desjenigen Bereichs variations- und medienlinguistischer Forschung dar, der sich speziell mit den Besonderheiten der Sprachverwendung und Kommunikation im Internet beschäftigt; auch empirische Untersuchungen zur Beschreibung und Analyse von aktuellen Tendenzen in der Entwicklung der deutschen Gegenwartssprache kommen im Zeitalter digitaler Kommunikationstechnologien nicht umhin, die internetbasierte Kommunikation als einen wichtigen und aktiven Bereich sprachlicher Variation und sprachlichen Wandels einzubeziehen und gestützt auf authentische Daten zu untersuchen. Nicht zuletzt werden Verfahren für die automatische linguistische Analyse von Sprachdaten aus Genres internetbasierter Kommunikation auch im Bereich der Sprachtechnologie benötigt, wenn es darum geht, große, aus dem World Wide Web erhobene „Webkorpora" mit automatischen Methoden für die Nutzung in sprachtechnologischen Anwendungen aufzubereiten und zu analysieren.

Um die Wortartenannotation für Sprachdaten aus Genres internetbasierter Kommunikation zu verbessern, benötigt man Kategorien, Ressourcen und Verfahren auf unterschiedlichen Ebenen:

(i) Kategorien und linguistische Beschreibungen zu charakteristischen Phänomenen internetbasierter Schriftlichkeit, die bei der Verarbeitung mit auf redigierten Texten trainierten Verarbeitungswerkzeugen typischerweise Probleme verursachen;

(ii) eine Typologie der Probleme, die sich bei der Behandlung dieser Phänomene auf unterschiedlichen Ebenen des Verarbeitungsprozesses (Tokenisierung, Taggingverfahren, Tagset) ergeben;

(iii) ein Tagset, das für die Aufgabe des Wortartentaggings für Genres internetbasierter Kommunikation erweitert wurde;

(iv) Trainingsdatensets mit manuell annotierten Daten, denen die Kategorien des erweiterten Tagsets zugrunde liegen und die die unter Punkt (ii) formulierten Probleme in einer für linguistische Analysezwecke sinnvollen Art und Weise behandeln (Goldstandard);

(v) Verarbeitungswerkzeuge, die auf diesen Trainingsdaten auf die Anwendung des erweiterten Tagsets und auf den Umgang mit den unter (ii) beschriebenen Problemtypen trainiert wurden.

Der vorliegende Beitrag präsentiert Beschreibungen und Kategorien zu den Punkten (i) und (ii) und leitet daraus Vorschläge für die Erweiterung des STTS ab (Punkt iii). Er stellt damit die wesentlichen konzeptuellen Grundlagen bereit, um Trainingsdatensets (iv) aufzubauen und auf deren Basis Verarbeitungswerkzeuge auf den Umgang mit den sprachlichen Besonderheiten internetbasierter Kommunikation zu trainieren (v).

Der Beitrag ist wie folgt aufgebaut: Nach einer Vorbemerkung zum Status sprachlicher „Besonderheiten“ und zu deren Verteilung in Kontexten internetbasierter Kommunikation (Abschnitt 2) skizzieren wir in Abschnitt 3 eine Typologie charakteristischer Phänomene bei der schriftlichen Sprachverwendung in Genres internetbasierter Kommunikation. Die in der Typologie erfassten Phänomentypen werden anhand von Datenbeispielen aus Chats und aus Wikipedia-Diskussionsseiten veranschaulicht. Der Fokus der Typologie liegt auf sprachlichen Einheiten und auf Phänomenen der schriftlichen Realisierung. Einheiten, in denen sich sprachliche Ausdrücke bzw. Schriftlichkeitsphänomene mit Phänomenen hypermedialer Vernetzung überlagern (z.B. Hashtags in Tweets), bedürfen weiterer empirischer und konzeptueller Klärung und werden daher zwar erwähnt, aber nicht systematisch mitbehandelt (vgl. Abschnitt 3, Phänomentyp VII).

In Abschnitt 4 geben wir einen datengestützten Überblick über Probleme bei der Verarbeitung einiger der in Abschnitt 3 vorgestellten Phänomene mit existierenden Werkzeugen für die automatische Tokenisierung und Wortartenannotation deutscher Sprachdaten, die von linguistischen Anwendern „off the shelf“ über die Oberfläche der webbasierten Analyseplattform *WebLicht* (https://weblicht.sfs.uni-tuebingen.de/) genutzt werden können und die das „Stuttgart-Tübingen-Tagset“ (STTS) als Ressource für die Zuordnung von Wortartentags zu Wort-Tokens nutzen. Der Problemaufriss zeigt, dass die

Probleme bei der automatischen Zuordnung von Wortartenkategorien auf unterschiedlichen Ebenen des Verarbeitungsprozesses zu verorten sind. Betroffen ist nicht nur die Ebene des Wortartentaggings selbst, sondern auch die Ebene der Tokenisierung sowie die Ebene der Festlegung eines geeigneten Kategorien- und Tagsets. Entsprechend müssen Ansätze zur Optimierung des Wortartentaggings nicht nur Verarbeitungsverfahren auf verschiedenen Analyseebenen an die Besonderheiten der Domäne „Sprachverwendung in der internetbasierten Kommunikation" anpassen; vielmehr müssen auch die Tagsets, die diesen Verfahren als Ressource zugrunde liegen, an die sprachlichen Besonderheiten der Domäne angepasst werden. Abschnitt 4 zeigt, welche der in Abschnitt 3 vorgestellten Phänomene einer Anpassung der Verfahren und welche einer Erweiterung des POS-Tagsets bedürfen. Für Phänomene der letzteren Art formulieren wir in Abschnitt 5 einen Vorschlag, wie sie in einer Erweiterung des STTS berücksichtigt werden könnten.

2 Sprachliche Besonderheiten und sprachliche Variation in der internetbasierten Kommunikation

Die schriftliche Sprachverwendung in Genres internetbasierter Kommunikation weist eine Reihe von Phänomenen auf, hinsichtlich derer sie sich von der Sprachlichkeit in redigierten Texten unterscheidet. Als „Besonderheiten" erscheinen diese Phänomene unter zweierlei Perspektiven:

a) Im Vergleich mit redigierten Texten (z.B. journalistischen Genres), die eine hohe Konformität mit den Normen der geschriebenen Standardsprache aufweisen, erscheinen sie als Abweichungen von der Norm bzw. als sprachliche Mittel, die in standard-nahen Textsorten nicht oder nur in besonderen Fällen auftreten oder erwartbar sind;

b) aus der Perspektive der Automatischen Sprachverarbeitung erscheinen sie als Phänomene, die mit gängigen, an redigierten Texten trainierten Verarbeitungswerkzeugen nicht oder nur unzureichend analysiert werden können und die deshalb bei der linguistischen Annotation von Sprachdaten einer besonderen Behandlung bedürfen.

Die Etikettierung der schriftlichen Sprachverwendung in der internetbasierten Kommunikation als ‚abweichend von den schrift(sprach)lichen Normen' bzw. ‚nichtstandardisiert' ist dabei lediglich im Sinne einer ‚Nicht-*Bezogenheit* auf den schriftlichen Standard' aufzufassen und nicht als Abweichung im Sinne einer nur unzureichend umgesetzten oder defizitären Erfüllung der Anforderungen an standardschriftlich realisierte Texte. Im Gegensatz zu prototypischen Textäußerungen, die – etwa i.S.v. Ehlich (1983, 1984) – situationsunabhängig rezipiert und verstanden werden sollen, unterliegt die sprachliche Gestaltung von Beiträgen im Rahmen getippter Dialoge in Foren, Chats, Weblog-Kommentaren, Diskussionen in Wikis oder auf den Profilseiten sozialer Netzwerke eigenen Normen, die den Normen für die Gestaltung mündlicher Gesprächsbeiträge nicht unähnlich sind: Leitlinie bei der Gestaltung der Kommunikationsbeiträge ist weniger die Sicherung der situationsunabhängigen Verständlichkeit eines sprachlichen *Produkts* als vielmehr der kommunikative Erfolg der damit realisierten sprachlichen Handlung(en) im Kontext der laufenden Interaktion. Die

sprachliche Form wird dabei optimiert für Adressaten, die die im Kommunikationsgeschehen vorausgegangenen kommunikativen Schritte kennen und über den aktuellen Stand des Geschehens auf dem Laufenden sind (vgl. Storrer 2012, 2013).

Dies gilt nicht nur für synchrone Genres wie Chat und Instant Messaging, in denen alle Beteiligten zeitgleich auf das Kommunikationsgeschehen orientiert sind und die Vorbeiträge noch unmittelbar mental präsent haben, sondern auch für asynchrone dialogische Genres wie Diskussions-Threads in Online-Foren, Wikis und sozialen Netzwerken, bei denen die schriftlichen Beiträge entsprechend ihrer zeitlichen Abfolge und häufig auch thematisch strukturiert auf einer Bildschirmseite vorgehalten werden. Auch wenn die Beteiligten nicht zeitgleich auf die Fortentwicklung des Dialoggeschehens orientiert sind, ist hier die jeweilige Vorkommunikation jederzeit im Wortlaut nachlesbar und kann, da neue Beiträge am Bildschirm direkt an den dokumentierten Verlauf der Vorkommunikation angefügt werden, bei der Formulierung neuer Beiträge auch bei den Adressaten eine Orientiertheit über den aktuellen Stand des Kommunikationsgeschehens vorausgesetzt werden.

Die in Abschnitt 3 vorgestellte Typologie beschreibt sprachliche Besonderheiten, die *charakteristischerweise* in der internetbasierten Kommunikation auftreten – was nicht bedeutet, dass sie in jedem Genre, in jedem Nutzungskontext und in jedem konkreten Kommunikationsereignis in gleicher Frequenz und Verteilung vorzufinden sind. Vielmehr hat die neuere linguistische Forschung zur internetbasierten Kommunikation wiederholt nachgewiesen, dass die Nutzer internetbasierter Kommunikationstechnologien ihre Sprachverwendung in ähnlicher Weise an soziale, institutionelle, situative und individuelle Rahmenbedingungen anpassen sowie in Abhängigkeit zu Themen und zu den jeweils instanziierten kommunikativen Gattungen variieren, wie dies in mündlichen Gesprächen oder in schriftlichen Texten der Fall ist. So zeigen z.B. Androutsopoulos/Ziegler (2003), dass die Verwendung von Regionalismen in stadtspezifischen Chat-Kanälen nicht unabhängig von der Dialekt-Standard-Relation der zugehörigen „realweltlichen" Region zu denken ist. Jarbou/al-Share (2012) beschreiben in einer Korpusuntersuchung zu Schreibvarianten in jordanischen Chats die Abhängigkeit graphematischer Variation von Gender und Dialekt. Storrer (2013) weist in einer quantitativen Auswertung eines Korpus zur deutschen Wikipedia nach, dass das Auftreten „typischer" Stilelemente systematisch in Abhängigkeit zum Genre (Wikipedia-Artikelseiten vs. Wikipedia-Diskussionsseiten) variiert. Luckhardt (2009) arbeitet in ihrer Dissertation auf Basis der Analyse eines Chat-Korpus heraus, dass die Verwendung „typischer" Stilmerkmale der internetbasierten Kommunikation zudem auch stark von individuellen Präferenzen der Nutzer beeinflusst wird. Auch Arbeiten, die sich mit der Nutzung der Chat-Technologie für unterschiedliche professionelle Nutzungskontexte (Beratung, Bildung, Medien) beschäftigen, haben differenziert die technischen und sozialen Faktoren beschrieben, die auf die Struktur und sprachliche Gestaltung des kommunikativen Austauschs Einfluss nehmen und die für eine erfolgreiche Nutzbarmachung der Technologie in professionellen Kontexten zu kontrollieren sind (vgl. z.B. die Beiträge in Beißwenger/Storrer (Hrsg.) 2005 sowie das „Chat-Szenario"-Modell in Beißwenger/Storrer 2005a).

Internetbasierte Kommunikationstechnologien konstituieren Kommunikations*formen* und keine *kommunikativen Gattungen*. Die *Formen* (z.B. Chat-Kommunikation, Forenkommu-

nikation, Twitter-Kommunikation usw.) sind durch die technischen Rahmenbedingungen der Technologie determiniert; einzelne *Gattungen* werden hingegen erst in der Nutzung einer Form für konkrete kommunikative Zwecke instanziiert (vgl. Beißwenger 2003, 2007: 107–112 in Anlehnung u.a. an vergleichbare Differenzierungen für den Bereich der Textlinguistik in Brinker 2001; Dürscheid 2005a). Entsprechend gibt es nicht „*die* Sprache des Internet“ oder „*die* Sprache des Chat / der E-Mail / der sozialen Netzwerke / der Blogs“, die sich nur mit Bezug auf die technologischen Rahmenbedingungen und in Absehung von sozialen und pragmatischen Faktoren bestimmen ließe (Storrer 2000, Storrer 2009; Dürscheid 2004).

Die Zusammenstellung verschiedener existierender oder derzeit in Aufbau befindlicher Korpora zur internetbasierten Kommunikation trägt dieser Tatsache Rechnung und berücksichtigt systematisch verschiedene Kommunikationsformen und deren Nutzungskontexte – vgl. z.B. Reynaert et al. (2010) zur Zusammensetzung des niederländischen Referenzkorpus *SoNaR*, Beißwenger (2013) zum 2002-2008 aufgebauten, nach Handlungsbereichen differenzierten „Dortmunder Chat-Korpus“, Beißwenger et al. (2013) zum laufenden Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (*DeRiK*).

3 Für das Wortartentagging relevante sprachliche Besonderheiten: eine Typologie

In diesem Abschnitt skizzieren wir eine Typologie sprachlicher Besonderheiten in Genres internetbasierter Kommunikation. Die unterschiedliche Verteilung der in der Typologie beschriebenen Phänomene in unterschiedlichen kommunikativen Genres sowie unter unterschiedlichen sozialen, institutionellen, situativen und individuellen Rahmenbedingungen (vgl. Abschnitt 2) wird dabei ausgeblendet. Ziel der Typologie ist es nicht, sprachliche und stilistische Variation in der internetbasierten Kommunikation zu beschreiben, sondern eine differenzierte Übersicht über diejenigen Phänomene zu bieten, hinsichtlich derer sich die schriftliche Sprachverwendung in der internetbasierten Kommunikation von der Sprachverwendung in redigierten Texten unterscheidet, um auf dieser Grundlage Verfahren für die automatische Erkennung und Disambiguierung dieser Phänomene entwickeln zu können. Der Fokus liegt dabei auf Einheiten, die bei der automatischen Wortartenannotation (POS-Tagging) Probleme bereiten. Da POS-Annotationen die Basis für alle weiteren Ebenen der linguistischen Annotation von Korpora darstellen, ist diese Annotationsebene beim Aufbau von Korpora von zentraler Bedeutung. Erst die Behandlung der in der Typologie erfassten Phänomene sowie der zugehörigen Verarbeitungsprobleme mit Werkzeugen für die automatische Sprachanalyse (s. Abschnitt 4) wird es ermöglichen, große, reichhaltig linguistisch annotierte Korpora zur internetbasierten Kommunikation aufzubauen, mit denen die Verteilung IBK-typischer sprachlicher Phänomene relativ zu unterschiedlichen Kontexten auf breiter Datenbasis quantitativ und qualitativ untersucht werden kann.

Um die in der Typologie erfassten Phänomene und die zugehörigen Verarbeitungsprobleme in den Griff zu bekommen, bieten sich prinzipiell zwei Ansätze an:

– die *Normalisierung*, bei der Wortformen, die von den genutzten Verarbeitungswerkzeugen aufgrund nicht-standardkonformer Formmerkmale nicht

sinnvoll analysiert und klassifiziert werden können, in einem Vorverarbeitungsschritt auf einer separaten Annotationsebene in eine Form überführt werden, die diese Merkmale nicht mehr aufweist. Dabei werden z.B. Phänomene geschriebener Umgangssprache durch ihre standardkonformen Pendants ersetzt (*machste* ⇨ *machst du*, *wat isn* ⇨ *was ist denn* usw.) und Schnellschreibphänomene beseitigt (*idese aufgabe* ⇨ *diese Aufgabe*). Anschließend wird die normalisierte Version der Ausgangsdaten mit den Verarbeitungswerkzeugen analysiert und annotiert.

– die *Anpassung existierender oder die Entwicklung neuer, spezialisierter Verarbeitungsverfahren* für den Umgang mit den Probleme bereitenden Phänomenen.

Welcher Ansatz sich im Einzelfall als praktischer erweist, muss in Abhängigkeit von der Forschungsfrage sowie vom Status der Verarbeitungsprobleme bereitenden Phänomene im jeweiligen Projekt entschieden werden. Grundsätzlich kommt eine Normalisierung nur für solche Phänomene in Betracht, zu denen sich eine normalisierte Form angeben lässt (z.B. im Falle der kontraktierten Formen, die sich durch Aufhebung der Klise in zwei standardorthographisch reguläre Formen auflösen lassen: *machste* ⇨ *machst du*). Hingegen entziehen sich Einheiten wie Emoticons, Aktionswörter und Adressierungen, die als spezifisch für Genres internetbasierter Kommunikation gelten können, einer Normalisierung. Für ihre Behandlung bei der POS-Annotation ist eine Erweiterung des zugrunde liegenden Tagsets unumgänglich.

Die nachfolgende Typologie zielt darauf, Besonderheiten der schriftlichen internetbasierten Kommunikation möglichst feinkörnig zu erfassen und Phänomenbeschreibungen bereitzustellen, die als Grundlage für die Erweiterung von Tagsets, für die Erarbeitung von Normalisierungsverfahren und die Anpassung von Verarbeitungswerkzeugen dienen können.

Der Fokus der Typologie liegt auf *sprachlichen Besonderheiten* und auf *Phänomenen der schriftlichen Realisierung*. Einheiten, bei denen sich Sprachliches mit Phänomenen der hypermedialen Vernetzung überlagert, bedürfen weiterer empirischer und konzeptueller Klärung und sind nur am Rande berücksichtigt (Phänomentyp VII).

Die Experimente, die wir anschließend in Abschnitt 4 vorstellen, zeigen, welche der im Folgenden unterschiedenen Phänomene bei der automatischen Verarbeitung Probleme auf welchen Verarbeitungsebenen verursachen. Für diejenigen Phänomene, die nur durch Erweiterung der Tagsets in den Griff zu bekommen sind, formulieren wir in Abschnitt 5 einen Vorschlag, wie sie im STTS dargestellt werden könnten.

I. Schnellschreibphänomene

- I.1 Irreguläre Verwendung von Spatien
- I.2 Tippfehler
- I.3 Ökonomiebedingte Abweichung von den Normen für die Groß- und Kleinschreibung

II. Graphische Nachbildung suprasegmentaler Elemente der gesprochenen Sprache

- II.1 Vollgroßschreibung von Wortformen und ganzen Äußerungen
- II.2 Iteration von Graphemen
- II.3 Iteration von Interpunktemen

III. Geschriebene Umgangssprache

- III.1 Umgangssprachlich fundierte Wortschreibungen
- III.2 Umgangssprachliche kontraktierte Formen
- III.3 Umgangssprachliche Lexik

IV. Verfremdungsschreibungen

V. IBK-typische Akronyme

VI. IBK-spezifische interaktive Einheiten

- VI.1 Emoticons
- VI.2 Aktionswörter
- VI.3 Adressierungen

VII. Weitere Phänomene

Abb. 1: Typologie sprachlicher Besonderheiten in der internetbasierten Kommunikation.

Im Folgenden geben wir Kurzbeschreibungen zu den einzelnen Phänomentypen (und Subtypen) (Abb. 1) und erläutern diese anhand von Beispielen aus Wikipedia-Diskussionsseiten [WPD] und aus dem Dortmunder Chat-Korpus [CHAT]:

I. Schnellschreibphänomene

Da es den Kommunikationsbeteiligten mit ihren schriftlichen Beiträgen primär darum geht, die laufende Interaktion weiterzuentwickeln, und weniger darum, ein situationsunabhängig verständliches schriftliches Produkt herzustellen, geschieht die Planung und Versprachlichung von Beiträgen häufig schnell und spontan und auf eine Korrekturdurchsicht der Beiträge vor der Verschickung wird häufig verzichtet. Typische Begleiterscheinungen der schnellen Produktion sind Tippfehler, untypische Platzierungen von Spatien sowie ein Verzicht auf die Anwendung der Regeln für die Großschreibung am Wort- und Satzanfang:

I.1 Irreguläre Verwendung von Spatien

Beim schnellen Tippen werden Spatien entweder – bedingt durch den Anschlag der Space-Taste mit zu geringer Dynamik – ausgelassen (Beispiele 1 und 2) oder kommt es durch versehentliches Betätigen von Tasten in einer anderen als der anvisierten Abfolge zur Platzierung von Spatien an einer anderen als der anvisierten Stelle (Beispiel 3). Entsprechend ergeben sich im verschickten Beitrag Token-Grenzen, die nicht den Grenzen der vom Produzenten realisierten Wort-Tokens entsprechen:

(1) *nu is tombefreit* (anstelle von: *Tom befreit*) [CHAT]

(2) *sind die B-Städter jetzt virtuell on? odersollen wir warten?* [CHAT]

(3) **wünk* und wehe ihr verleumdet mich *grml* *brummel* imme rnoch nicht fassen kann* [CHAT]

Es ist anzunehmen, dass irregulär gesetzte Spatien in der Mehrzahl der Fälle unbeabsichtigt als Folge schneller Tippaktivität auftreten; in Einzelfällen können sie aber auch auf ein bewusstes Spiel mit der medialen Schriftlichkeit der Kommunikation zurückgeführt werden (vgl. das Beispiel in Beißwenger/Storrer 2012: 100).

I.2 Tippfehler

Tippfehler sind irreguläre orthographische Realisierungen von Wortformen, die typischerweise daraus resultieren, dass die Tasten der Tastatur – flüchtigkeitsbedingt – falsch, ungenau, zu stark, zu schwach oder in falscher Reihenfolge angeschlagen werden. Die in Genres internetbasierter Kommunikation auftretenden Tippfehler lassen sich in vier Typen einteilen (Beißwenger 2000: 73f.): „Vertipper" (Typ I), bei denen anstelle der Taste mit dem anvisierten Zeichen fälschlich eine auf der Tastatur nebengelegene Taste angeschlagen wird (Beispiel 4); „Vertipper" (Typ II), bei denen zusätzlich zur Taste mit dem anvisierten Zeichen eine weitere Taste angeschlagen wird, die entweder ein zusätzliches Zeichen realisiert oder – im Falle der Shift-Taste – eine Realisierung der in der Folge getippten Zeichen in Versalien verursacht (Beispiele 5 und 6); anschlagsdynamisch bedingte Auslassungen (Beispiel 7) oder Mehrfachrealisierungen eines anvisierten bzw. an der betreffenden Position in der Graphemstruktur vorgesehenen Zeichens (Beispiel 8); „Buchstabendreher", bei denen zwei oder mehrere in der Graphemstruktur aufeinander folgende Zeichen in falscher Reihenfolge realisiert werden (Beispiel 9):

(4) *Wer gib eigentlich seinen Namen hier preis? Ingeborbbachmann. Seltsam!* [CHAT]

(5) *31,5 werd ich auchg noch schaffen* [CHAT]

(6) *dass prinzipiell auchg alles, was sie "sAQGEN2; MITgeschnitten werden kann...* [CHAT]

(7) *Bin ja garnicht böe* [CHAT]

(8) *ich bin durcheinander weil ich nochmals von vorne anfangen muss du dagegenkannst anffangen zu planen* [CHAT]

(9) *da sollten wir wohl schleunigst den mantel des schweigens über idese aufgabe breiten / dat ebste findeste eigentlich wenn du gar nich suchst* [CHAT]

I.3 Ökonomiebedingte Abweichung von den Normen für die Groß- und Kleinschreibung

Als typisches Schnellschreibphänomen gilt weiterhin die liberale Anwendung oder radikale Nichtanwendung der Normen für die Großschreibung. Dabei wird, um Tippaufwand zu minimieren, auf die Betätigung der Umschalttaste verzichtet, die dazu benötigt wird, den Buchstaben, mit dem eine gleichzeitig angeschlagene Buchstabentaste auf der Tastatur belegt ist, in Versalien zu realisieren. Im Einzelfall kann die irreguläre Anwendung der Regeln für die Großschreibung im Deutschen natürlich auch kompetenzbedingt sein.

Das Abweichen von bzw. die Nichtanwendung der Großschreibung kann sich dabei sowohl auf die Großschreibung von Substantiven und Eigennamen wie auch auf die Großschreibung von Satzanfängen (innerhalb von Nutzerbeiträgen) und Textanfängen (bezogen auf die internetbasierte Kommunikation: die Großschreibung zu Beginn eines neuen Kommunikationsbeitrags) beziehen:

(10) Reguläre Großschreibung der Substantive, aber keine Großschreibung am Beginn des Beitrags: *ich habe mich heute bei Wikipedia angemeldet (nach Inspiration durch meinen Berufsschullehrer *zwinker*) und hoffe auf einen Ausbau des Portals* [WPD]

(11) Großschreibung zu Beginn des Beitrags, aber konsequente Kleinschreibung aller weiteren Substantive: *Dinge, die körperlich nicht passieren können oder kommunikation, die eigetnlich widersinnig wird, wir sofort akzeptiert und lediglich als philosophischer bruch empfunden.* [CHAT]

(12) Radikale Kleinschreibung (Substantive, Satzanfänge, Beginn des Beitrags): *immer wieder schön beim arbeiten gestört zu werden werden so kleinigkeiten wie ein externer link. entweder machst du die seite oder lässt mich einfach mal alles hier fertig machen und dann kannse von mir aus mal mal drüber guggn.* [WPD]

II. Graphische Nachbildung suprasegmentaler Elemente der gesprochenen Sprache

Ein weiterer Phänomentyp ist die Nachbildung suprasegmental-prosodischer Elemente der gesprochenen Sprache mit den Mitteln der Schrift. Typische Verfahren sind die Vollgroßschreibung von Wortformen und ganzen Sätzen oder Beiträgen sowie die Graphemiteration (vgl. Beißwenger 2000: 104f.). Dürscheid (2005: 48f.) klassifiziert Normabweichungen dieser Art als *Stilmittel*; im Gegensatz zur liberalen Anwendung der Normen für die Groß- und Kleinschreibung, die zwar – sofern es sich nicht um Kompetenzfehler handelt – ebenfalls beabsichtigt ist, aber nur der Ökonomisierung der Beitragsproduktion dient, handelt es sich bei den Phänomenen in dieser Gruppe um Schreibungen, mit denen zudem ein kommunikativer Zweck verfolgt wird:

II.1 Vollgroßschreibung von Wortformen und ganzen Äußerungen

Die Vollgroßschreibung von Wortformen, Sätzen oder ganzen Beiträgen nutzt Versalien als typographische Metapher für Intensität. Dabei werden Großbuchstaben verwendet, um lautes Sprechen oder Schreien darzustellen. Groß geschrieben werden entweder einzelne Wortformen oder ganze Sätze bzw. Beiträge. In ersterem Fall (Beispiele 13–16) dient die Vollgroßschreibung häufig der Fokussierung, in zweiterem Fall der Symbolisierung von Emphase bzw. von lautem Sprechen/Schreien (Beispiel 17):

(13) *Mehr Arbeitsplätze müssen her, die Konjunktur muß angekurbelt werden. Rot/Grün setzt die FALSCHEN Signale, trifft die FALSCHEN Entscheidungen .....armes Deutschland .. zahlen müssen leider jetzt auch die, die Rot/Grün nicht gewählt haben.. :-(((* [CHAT]

(14) *latinumsprüfung ist NICHT egal... aber ich nehme mal an, du meintest das auch nicht...* [CHAT]

(15) *Hab übrigens mitm Hersteller Kontakt aufgenommen, die wollen das Logo erst dann hochladen, wenn der Löschantrag NICHT durchgegangen ist* [WPD]

(16) (Autor 1:) *Einfach die Tabelle aus einer alten Version in die Zwischenablage kopieren, in die letzte Version einfügen, speichern, fertig :-)* – (Autor 2:) *:-) DAS habe ich ja die ganze Zeit versucht! *heul* Und ich habe es an zwei Computern mit drei verschiedenen Browsern probiert: Nullanzeige :-(* [WPD]

(17) *mathe mündlich? MATHE MÜNDLICH! BRUTAL! das war bei uns schriftlich und schon schlimm genug!* [CHAT]

II.2 Iteration von Graphemen

Die Iteration von Graphemen dient der Nachbildung von Dehnung zur Setzung von Emphaseakzenten in der gesprochenen Sprache und zum Ausdruck von Emotion:[3]

(18) *ein seeeehr heikles Thema auf jeden Fall, wer da einen fairen und treffenden Absatz zustande bringt, bekommt von mir einen Orden;-)* [WPD]

(19) *Tiggi ist ja sooooo erwachsen ;-)* [CHAT]

(20) *leider nicht, nö *schaaaaade* [CHAT]

Die Iteration tritt auch in akronymischen, silbisch interpretierten Formen auf (im folgenden Beispiel im Aktionswort *lol*):

(21) *FRagt mal den Koch in Hessen nach dem Spendenkonto*looooooool** [CHAT]

Bisweilen treten Vollgroßschreibung (II.1) und Graphemiteration auch in Kombination auf:

(22) *Burzel nimmt anlauf und stürzt sich im HUUUUUURRRRRAAAAAA auf SX zum SUPERDUPPERHYPERMEGAKKKKKNNNNUUUUUDDDDLLLLÄÄRRRRR* [CHAT]

II.3 Iteration von Interpunktemen

Die Iteration von Interpunktemen dient der Symbolisierung von Emphase und Emotion (Beispiele 23 und 24) bzw. der (performativen) Nachbildung von Planungspausen aus der gesprochenen Sprache im Medium der Schrift (Beispiel 25):

(23) *Wir wollen die Welt verbessern... OHNE Drogen!!!!!* [CHAT]

(24) *Jeder muss zugeben: dass kanns doch nicht sein!!!* [WPD]

(25) *Warum stehe ich nur mit Klarnamen in dieser Kategorie?.....mmh..... Werbung?.....Schwarzes Brett???.....* [WPD]

Während die Vervielfältigung von Vokal- und Konsonantengraphemen eine Iteration der graphischen Repräsentationen einzelner Lautsegmente *innerhalb* eines Tokens darstellt, wird bei der Iteration eines Interpunktionszeichens das komplette Token (das standardgemäß nur aus einem einzigen Zeichen besteht) vervielfältigt.

III. Geschriebene Umgangssprache

Mit der internetbasierten Kommunikation ist die Schrift als Realisierungsmedium für sprachliche Äußerungen im großen Stil auch für solche Handlungsbereiche nutzbar geworden, die zuvor eher der gesprochenen Sprache vorbehalten waren (vgl. Storrer 2001:439). Insbesondere in solchen Nutzungskontexten internetbasierter Kommunikationstechnologien, die neben der Dialogizität weitere Kommunikationsbedingungen der Nähe – z.B. einen vertrauten, informellen Umgang der Partner, eine freie Themenentwicklung, eine hohe Spontaneität bei der Planung und Realisierung von Kommunikationsbeiträgen – aufweisen, lässt sich eine Orientierung an einer „Sprache der Nähe“ (i.S.v. Koch/Oesterreicher 1994) feststellen, die sich auf Wortebene in einer Realisierung umgangssprachlicher (z.T. auch mundartlicher) Formen und Strukturen im Medium der Schrift niederschlägt („Geschriebene Umgangssprache“, vgl. z.B. Kilian 2001).

III.1 An der umgangssprachlichen Lautung orientierte Wortschreibungen

Bei der Wortschreibung zeigt sich die Orientierung an der Umgangssprache in der Anwendung der orthographischen Prinzipien des Deutschen auf die umgangssprachliche anstelle der standardsprachlichen Lautung von Wortformen. Dabei werden z.T. auch typisch sprechsprachliche Elisionen ins schriftliche Medium transponiert (wie z.B. *is* < *ist*, *ne* < *eine*, *wunder* < *wundere* in den nachfolgenden Beispielen):

(26) *Jut, ich find die Variante mit "Die" auch besser und "richtiger" in diesem Fall.* [WPD]

(27) *japp da habe ich es auch gelesen, aber sonst nirgens* [WPD]

(28) *Das sind zuviele Artikel drinne, die mit Recht im weiteren Sinne nichts mehr zu tun haben.* [WPD]

(29) *Nee, mit Vadder is hier Kim Il Sung gemeint, der Sohn demnach Kim Jong Il.* [WPD]

(30) *Isch ja gut, es hier noch anderes zu tun als solcher Kleckerleskram.* [WPD]

(31) *ick wunder mir über jarnischt mehr* [WPD]

(32) *Ick weeß, auch det is Jeschichte ... aber hässlich bleibt hässlich.* [WPD]

(33) *Noch ne kleene Nachfrage: Brauchen wir demnächst ooch noch ne Themenkat:NS-Opfer?* [WPD]

(34) *entweder machst du die seite oder lässt mich einfach mal alles hier fertig machen und dann kannse von mir aus mal mal drüber guggn.* [WPD]

(35) *Kandidaturdiskussion auf Artikeldiskussion kopieren und auf WP:KALP löschen. Dann noch einen Abbruchvermerk, Kandidaturbaustein aus Artikel entfernen - zack, feddich!* [WPD]

(36) *Moinsen, wirf mal einen Blick auf WP:WEB, das hatte ich dir schon heude mittach als IP geraten* [WPD]

(37) *dit is selbstverständlich* [CHAT]

(38) *ozelot sacht moin zu stoeps* [CHAT]

(39) *p.s.: knuffine ist ne olle petze ;) *lol** [CHAT]

(40) *gun tach* [CHAT]

III.2 Umgangssprachliche kontraktierte Formen

In der internetbasierten Kommunikation finden sich unterschiedlichste Arten von kontraktierten Formen, die aus der Verschmelzung zweier aufeinanderfolgender syntaktischer Wörter resultieren. Einerseits begegnen kontraktierte Formen wie *im* (< *in dem*), *am* (< *an dem*), *zur* (< *zu der*), *zum* (< *zu dem*), *ans* (< *an das*), *ins* (< *in das*), die auch standardsprachlich – d.h. in redigierten Texten – gebräuchlich sind, einen hohen Grammatikalisierungsgrad aufweisen und sich in vielen Fällen nicht mehr ohne Weiteres durch ihre Ausgangsformen ersetzen lassen (z.B. in *zum Erliegen/Stocken/Erlahmen kommen*, *ins Schleudern/Trudeln/Grübeln/ Schwärmen geraten*). Daneben finden sich aber auch kontraktierte Formen wie *haste*, *biste*, *isn* und *aufm*, bei denen zwei (oder mehrere) aufeinanderfolgende Wortformen koartikulationsbedingt miteinander

verschmolzen werden und die somit als typisch sprechsprachlich zu gelten haben. Mit ihrer Übernahme in geschriebene Äußerungen – bei deren Produktion Koartikulation keine Rolle spielen *kann* – wird ein Phänomen medialer Mündlichkeit in die mediale Schriftlichkeit transponiert (‚verschriftet' i.S.v. Koch/Oesterreicher).

Typische Bildungsmuster sind (z.B.):

- Präposition + Artikel: *innem, aufm, aus(s)n*
- Adverb + Artikel: *noch(e)n*
- Konjunktion + Personalpronomen: *fallste* (< *falls du*), *obse* (< *ob sie*)
- Auxiliarverb + Personalpronomen: *haste, biste*
- Vollverb + Personalpronomen: *machste, gehste, denkste, schreibste*
- Vollverb + zwei Personalpronomina: *machstes, gibstes*
- Kopulaverb + Personalpronomen: *warens*
- Modalverb + Personalpronomen: *kannste, willste, sollste, darfste*
- Auxiliarverb + Abtönungspartikel: *(was) isn (passiert)*

Beispiele für die Verwendung in Wikipedia-Diskussionen und in Chats sind:

(41) *Prinzipiell schon, wobei auch das mit überregional so ne Sache ist. Innem halbwegs großen Staat erreichen seriöse Regionalzeitungen deutlich mehr Leser als die gesamte Einwohnerschaft Kosovos.* [WPD]

(42) *wieso "war" aswad denn eine band? sind die nich dieses jahr aufm summerjam, d.h. noch oder wieder aktiv?* [WPD]

(43) *Ich hab's auch nicht verstanden, rinn' inne Kartoffeln, raus aussn Kartoffeln. Wenn's der Account so will...* [WPD]

(44) *Nochen Vorschlag: Die Episodenliste unter „Die Simpsons: Episodenliste" ablegen.* [WPD]

(45) *nicht so gut wie alt biste und wo kommste her* [CHAT]

(46) *Fallste das kannst, ist dir wohl der Nobelpreis sicher.* [WPD]

(47) *Kategorie:Gesellschaft ist schlicht falsch; obs Dir passt oder nicht.* [WPD]

(48) *...war das die ursprüngliche zielsetzung? oder wars nich einfach nur der wunsch, neulingen bei ihren erstschritten zu helfen - unabhängich davon, obse auch bleibn... ??* [WPD]

(49) *hm, magste mal genauer erläutern, warum ein exzellenter artikel über eine skisprungschanze „fatal" sei?* [WPD]

(50) *alles klar, ich schreibs nochmal neu* [WPD]

(51) *na da haste aber was verschlimmbessert, machstes selber rückgängig? :)* [WPD]

(52) *naja,bei mir warens einige semester nachrichtentechnik halt.....* [CHAT]

(53) *aber wie gesagt,meinerseits hättestes ok da du ja nun quasi angefragt hast...* [CHAT]

(54) *Aber meine Überleitung zu den einzelnen Strömungen des Renaissancehumanismus sollste nicht wegkürzen* [WPD]

(55) *warum wollteste jemanden trffen der so etwas macht* [CHAT]

(56) *Könnteste die Datei nochmal auf Commons hochladen und hier SLA stellen? Danke.* [WPD]

Im Einzelfall können sich in ein- und demselben Token unterschiedliche Phänomene überlagern. Die Beispiele 57 und 58 zeigen umgangssprachliche kontraktierte Formen, die zusätzlich einen Tippfehler (57) bzw. eine Orientierung an einer umgangssprachlichen Formvariante (58) aufweisen:

(57) *dann kansstes dur auch direkt ausdenken...* [CHAT] (kannstes < kannst du es)

(58) *gibbet denn da kein bild?* [CHAT] (gibbet < gibt et „gibt es“)

III.3 Umgangssprachliche Lexik

Geschriebene Umgangssprache findet sich nicht nur in der Graphie (III.1) und in der Verwendung kontraktierter Formen (III.2), sondern auch im Bereich der Lexik. Charakteristisch insbesondere für informelle Kontexte internetbasierter Kommunikation ist die Verwendung umgangssprachlicher, in der gesprochenen Sprache z.T. regional gebundener Lexik:

(59) *ozelot sacht moin zu stoeps* [CHAT]

(60) *Moinsen, wirf mal einen Blick auf WP:WEB, das hatte ich dir schon heude mittach als IP geraten* [WPD]

(61) *Und wer schreibt nun den Kurierbeitrag? Gruß vonne Maloche* [WPD]

(62) *p.s.: knuffine ist ne olle petze ;) *lol** [CHAT]

IV. Verfremdungsschreibungen

Nicht-standardkonforme Schreibungen wie in den Beispielen 63–65 lassen sich keinem der bislang unterschiedenen Phänomentypen zuordnen: Im Unterschied zu Schnellschreibphänomenen (I) lässt sich ihre Hervorbringung nicht aus der beschleunigten Textproduktion erklären, im Gegensatz zu Fällen des Typs II geht es bei ihnen auch nicht um die Nachbildung prosodischer Elemente. Gegenüber Fällen geschriebener Umgangssprache (III) wird gezielt vom phonographischen Prinzip (den Graphem-Phonem-Korrespondenzen für das Deutsche) abgewichen und für bestimmte Lautsegmente eine graphematische Repräsentation gewählt, die gegenüber der orthographischen Norm als stark markiert gelten kann. Funktion solcher Schreibungen

ist die gezielte und kreative graphematische Verfremdung – z.T. in Anlehnung an den Schreibgebrauch in bestimmten Szenen (z.B. graphematische Ersetzung <s> → <z> im Auslaut in Anlehnung an Schreibungen in der Hip-Hop-Szene, Beispiele 63 und 64). Auch die Nutzung von Zahlzeichen zur Repräsentation von Wörtern oder Wortbestandteilen, die phonologisch der Lautung des Zahlzeichens entsprechen (*n8* < *Nacht*, Beispiel 65), rechnen wir zu dieser Kategorie:

(63) *Congratulations Mädelz, das habt ihr gut gemacht. :)* [WPD]

(64) *Tach Wurm, geh mich doch fott mit die Plörre...Gibt's heute übrignz das nächste Waterloo, und ganz ohne Bierbecherweitwurf ???* [WPD]

(65) serIan: *ich hau mich ne ecke aufs ohr....*
stoeps: *n8 seri*
Tigerelse: *nacht, seri :)* [CHAT]

V. IBK-typische Akronyme

In Genres internetbasierter Kommunikation finden sich zum einen okkasionelle Abkürzungen, bei denen um der Zeitersparnis willen Ausdrücke für als bekannt vorausgesetzte Redegegenstände mit den Mitteln der deutschen Kurzwortbildung regulär gekürzt werden. Daneben haben sich für bestimmte, häufig genutzte Wendungen stabile Akronyme eingespielt, die einen z.T. hohen Idiomatisierungsgrad aufweisen. Viele der Abkürzungen und Akronyme stammen aus dem Englischen (Beispiele 68–76; deutschen Ursprungs hingegen 64–66 sowie *lg* in 67), einige sind spezifisch für bestimmte Communities (z.B. Wikipedia-Diskussionen: *POV* in Beispiel 76), andere sind Community-übergreifend gebräuchlich.

(64) *Einfach kann jeder, fragen kostet nix. Vllt. kann Dir Leonhardt ja schon 'ne Lageinfo liefern?* [WPD] (Vllt. < vielleicht)

(65) *Positionen zu Umweltpolitik. Jmd fleißig genug, die zu finden und einzuarbeiten?* [WPD] (Jmd < jemand)

(66) *kA was das kostet* [CHAT] (kA < keine Ahnung)

(67) *hi FIST, du was muss ich tun wenn ich eine neue Kategorie anlegen will bzw. weißt du wo die Seite ist wo man das nachlesen kann? Mir ist es dabei wichtig zu wissen welche Kriterien erfüllt werden müssen. thx und lg* [WPD] (thx < thanks, lg < liebe Grüße)

(68) *Dass die Veikkausliiga im Sommer spielt, führt hier IMO zu weit, das gehört in den entsprechenden Artikel* [WPD] (IMO < in my opinion)

(69) *Imho ist die Kritik an der Interpretation nicht richtig.* [WPD]
(Imho < in my humble opinion)

(70) *Btw: Diesen Edit solltest du genauer erklären.* [WPD] (Btw < by the way)

(71) *Sry, keine Ahnung warum der Artikel in der Form überhaupt existiert, für mich eindeutig enz. irrelevant, und SLA-fähig* [WPD] (sry < sorry)

(72) *Das Reisen mit Buchungsbestätigung statt klassischem Papierticket ist ja z.B. inzwischen afaik heutzutage eher die Regel als eine Ausnahme.* [WPD] (afaik < as far as I know)

(73) *Danke für den Hinweis - wird ASAP geändert.* [CHAT] (ASAP < as soon as possible)

(74) *cu biene bis samstag* [CHAT] (cu < see you)

(75) *bin mal eben für ca. 5 minuten "afk" - melde mich dann gleich wieder zurück.* [CHAT] (afk < away from keyboard)

(76) *Was ist daran so schwer zu verstehen, dass die Aussage „xy schmeckt gut" POV ist?* [WPD] (POV < point of view)

VI. IBK-spezifische interaktive Einheiten

Auf lexikalischer Ebene finden sich mit den *Emoticons*, den *Aktionswörtern* und den *Adressierungen* Elemente, die als spezifische Erweiterungen des Inventars sprachlicher Einheiten in der internetbasierten Kommunikation gelten können. Allen drei Elementen ist gemeinsam, dass sie sich syntaktisch-positional wie auch hinsichtlich ihrer Funktionen sehr ähnlich verhalten wie Einheiten, die in Grammatiken des Deutschen – in z.T. unterschiedlichem kategorialen Zuschnitt – als *Interjektionen* (GDS, Duden-4[7]), *Responsive* (GDS) oder *Gesprächspartikeln* (Duden-4[5]), in Grammatiken des Englischen als *interjections* (Greenbaum 1996, McArthur et al. 1998, Blake 2008), *inserts* (Biber et al. 1999; 2002) oder *discourse markers* (Schiffrin 1986) beschrieben werden: Sie sind in aller Regel nicht syntaktisch integriert, tragen also nicht zum kompositionalen Aufbau der Satzbedeutung bei, und können sowohl im linken oder rechten Außenfeld von Sätzen auftreten wie auch an nahezu beliebiger Position in Form von Parenthesen eingeschoben werden. Funktional sind sie spezialisiert auf Aufgaben im Bereich der Handlungskoordination im Dialog, der emotionalen Kommentierung und der Respondierung vorangegangener Partneräußerungen.

Die Grammatik der deutschen Sprache (*GDS*, Zifonun et al. 1997) führt Interjektionen (*ach*, *äh*, *mhm*, *ne*, *tja*) und Responsive (*ja*, *nein*, *okay*) aufgrund ihrer auf die spezifischen, auf die Organisation interaktiven Austauschs spezialisierten Funktionen in einer eigenen Kategorie ‚Interaktive Einheiten'. Will man Emoticons, Aktionswörter und Adressierungen in einen grammatischen Beschreibungsrahmen einordnen, so eignet sich diese Kategorie in besonderer Weise: Emoticons, Aktionswörter und Adressierungen lassen sich dann als spezifische Erweiterung des sprachlichen Inventars für die besonderen Bedürfnisse bei der Organisation *schriftlicher* dialogischer Interaktion beschreiben (vgl. Beißwenger et al. 2012: 3.5.1).

Als IBK-spezifische interaktive Einheiten sind Emoticons, Aktionswörter und Adressierungen auf unterschiedliche Aufgaben spezialisiert. Entsprechend bilden wir für ihre typologische Einordnung drei eigene Subtypen. Diesen sind die „klassischen" Typen von interaktiven Einheiten – Interjektionen und Responsive – nebengeordnet (Abb. 2).

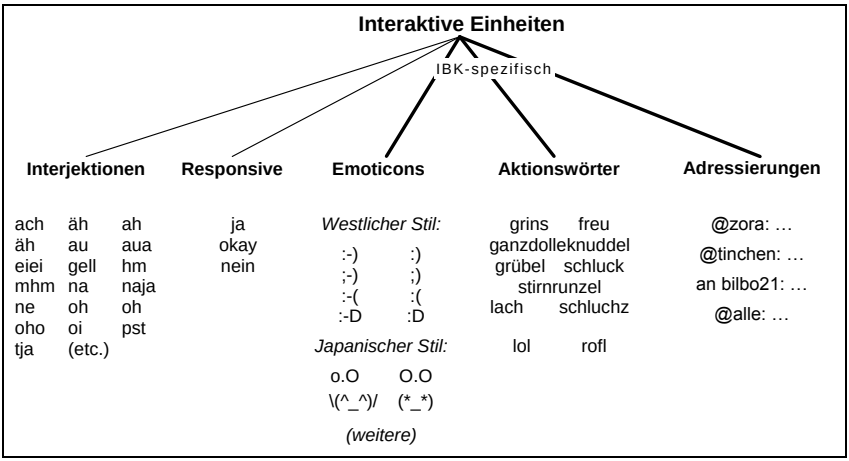


Abb. 2: Emoticons, Aktionswörter und Adressierungen als IBK-spezifische Erweiterungen der Kategorie der ‚interaktiven Einheiten' (GDS).

VI.1 Emoticons

Emoticons werden typischerweise durch die Kombination von Interpunktions- und Sonderzeichen gebildet; bisweilen können sie auch Buchstabenzeichen enthalten. Durch ihre ikonische Fundierung sind sie übereinzelsprachlich verwendbar. In unterschiedlichen Kulturkreisen haben sich unterschiedliche Stile herausgebildet (z.B. westlicher, japanischer, koreanischer Stil), deren Verwendung aber nicht auf die jeweiligen Ursprungskulturen beschränkt geblieben ist. So sind in vielen deutschsprachigen Online-Communities neben den „klassischen" Emoticons westlichen Stils inzwischen u.a. auch japanische Emoticons gebräuchlich.

Emoticons können am Ende eines Satzes bzw. einer satzwertigen kommunikativen Einheit, am Ende eines Beitrags oder – seltener – in Form von Parenthesen auftreten sowie alleine einen Kommunikationsbeitrag realisieren. Sie werden u.a. zur emotionalen Kommentierung, zur Respondierung von Vorgängeräußerungen oder als Illokutions- und Ironiemarker verwendet:

(77) *och,die fischbude am heumarkt is ok**:-)*** [CHAT]
(Emotionale Kommentierung der eigenen Äußerung)

(78) ***:(((*** *Mit mir will einfach keiner chatten!* ***:(((*** [CHAT]
(Emotionale Kommentierung der eigenen Äußerung)

(79) *Ach nee, jetze isses plötzlich wieder eine Stadt?* ***:-P*** (WPD)
(Ironiemarkierung)

(80) ***:-/*** *Nein, nicht wirklich. Na ja, aber was ist den der Sinn des ganzen?0* [WPD] (Evaluative Respondierung einer Vorgängeräußerung)

(81) Weswolf: *Weswolf trabt zu tränchen rüber und begrüßt sie lieb*
tränchen: *tränchen hallöchen an alle :) besonders an WESWOLF :))))))))*
Weswolf: ***:-)))***

[CHAT] (Evaluative Respondierung einer Vorgängeräußerung)

(82) Raebchen: *ich habe als kind mal auf einem delphin gestanden (wirklich)*
Matrose: *cool*
Raebchen: *naja, der war gestrandet :(*
Matrose: *raebchen: und du bist dann darauf rumgeklettert - toll* ***:-(***

[CHAT] (Illokutionsmarkierung: Indikator für indirekten Sprechakt)

In Beißwenger et al. (2012: 3.5.1.4) wird für die Behandlung von Graphiken, die ähnliche Funktionen wie Emoticons und Aktionswörter übernehmen, eine Kategorie ‚interaction template' vorgeschlagen. Da wir uns in der hier beschriebenen Typologie auf (sprachliche bzw. tastaturschriftliche) Einheiten konzentrieren, die für das Wortartentagging relevant sind, sind graphisch realisierte Einheiten hier nicht als ein eigener Typ erfasst. Selbstverständlich ist bei der automatischen linguistischen Analyse von IBK-Daten aber auch mit in die Nutzerbeiträge eingebetteten Medienobjekten (Graphiken, animierte Graphiken, Videodateien) umzugehen – worunter auch die sog. „Graphik-Smileys" fallen –, dann aber nicht auf der Ebene der Wortartenklassifikation, sondern im Rahmen von Vorverarbeitungsschritten (vgl. VII zu weiteren Phänomenen in IBK-Daten). Entsprechend umfasst die hier beschriebene Kategorie ‚Emoticon' tatsächlich nur tastaturschriftlich erzeugte Einheiten.

VI.2 Aktionswörter

Aktionswörter sind einzelsprachlich gebundene symbolische Einheiten. Strukturell basieren sie auf einem Wort – zumeist einem unflektierten Verbstamm (‚Inflektiv', Schlobinski 2001) –, das entweder alleine steht oder um weitere Einheiten erweitert sein kann – im Falle von Inflektiven um vom Verb geforderte Ergänzungen oder Angaben. Im Falle solcher Konstruktionen werden die einzelnen Wortformen sehr häufig zusammengeschrieben, sodass sie formal

als ein Token erscheinen; sehr verbreitet ist zudem die Markierung mit ein- und ausleitenden Asterisken.

Aktionswörter werden zur Beschreibung von Gesten, mentalen Zuständen oder Handlungen verwendet. Sie dienen als Emotions- oder Illokutionsmarker (Beiträge 865, 876, 880 in Beispiel 83), als Ironiemarker (Beiträge 875, 878, 879), zur spielerischen Nachbildung fiktiver Handlungen (Beitrag 864) oder dazu, sich selbst (oder dem eigenen virtuellen Charakter) Charaktermerkmale oder innere Zustände zuzuschreiben (Beitrag 881).

Einige sehr gebräuchliche Aktionswörter haben die Form von Akronymen – z.B. **lol** (< *lauging out loud*), **rofl** (< *rolling on the floor laughing*), **g** (< *grin(s)*), **s** (< *smile*).

(83) Ausschnitt aus einem Chat-Mitschnitt:

858	Turnschuh:	*OHNE DEUTSCHLAND FAHRN WIR ZUR EM!*
859	system:	*Ryo hat die Farbe gewechselt*
860	Gangrulez:	*jo schade*
861	system:	*Windy123 geht in einen anderen Raum: Forum*
862	juliana:	*alle leute müssen ihre fernseher bei media markt bezahlen*
863	juliana:	*haha*
864	Turnschuh:	*Es gab mal ein Rudi Völler.....es gab mal ein Rudi Völler.....♫**sing**♫*
865	Ryo:	***g***
866	Gangrulez:	*hehe..das wurd eh gerichtlich gestoppt juliana*
867	juliana:	*echt?*
868	oz:	*gang: echt ??*
869	Gangrulez:	*ja*
870	juliana:	*wieso?*
871	Gangrulez:	*wettbewerbsverzerrung*
872	Naturkonstantler:	*Fussball ist sooo unendlich unwichtig...*
873	juliana:	*versteh ich nicht. ich fand es war ein cooler trick*
874	Gangrulez:	*aber es war eine Art Glücksspiel*
875	Turnschuh:	*mag auch keinen Fussball.....nur wollte ich das letzte Deutschlandspiel sehen **fg***
876	Chris-Redfield:	***s*** *aber net erlaubt @ juli*
877	juliana:	*fußball ist nen dreck wichtig. es ist ein spiel. hauptsache, die jungen männer haben sich fitgehalten und ihrer gesundheit was getan :)*
878	Gangrulez:	*und das entspircht nicht dem Handel **g***
879	juliana:	*chris, du weißt doch, daß ich ein gesetzesbrecher bin **g***
880	Chris-Redfield:	*ja ich weiß **s***
881	juliana:	***wildsei***

[CHAT]

VI.3 Adressierungen

Adressierungen sind sprachliche Ausdrücke, mit denen ein Kommunikationsbeitrag an einen bestimmten anderen Kommunikationsbeteiligten oder eine Gruppe von Kommunikationsbeteiligten adressiert wird. Ihr zentraler Bestandteil ist ein Ausdruck, mit dem der Adressat benannt (Angabe des Adressatennamens oder einer Variante davon) oder charakterisiert wird (Paraphrase). In vielen Fällen ist der Angabe des Adressaten ein Adressierungsmarker – häufig das auch aus E-Mail-Adressen bekannte <@>-Zeichen – vorangestellt. Sie stehen in aller Regel initial oder final zu der sprachlichen Äußerung, deren Adressiertheit kenntlich gemacht werden soll. Im Falle initialer Verwendung ist die Adressierung zumeist durch einen Doppelpunkt von den folgenden Einheiten abgesetzt.

Adressierungen dienen als ökonomische Mittel zur Anknüpfung an Beiträge oder Themen aus der Vorkommunikation oder zur Adressierung von Personen bzw. Personengruppen. In letzterer Funktion weisen sie Parallelen zu Anredeformen in Gesprächen auf; in ersterer Funktion dienen sie der Explizitmachung sequenzieller oder thematischer Kontinuität quer zur am Bildschirm angezeigten Beitragsabfolge. Da insbesondere in synchronen Formen internetbasierter Kommunikation (Chat, Instant Messaging) für die Beitragsproduzenten aufgrund der technischen Rahmenbedingungen nicht exakt planbar ist, ob handlungssequenziell oder thematisch aufeinander bezogene Beiträge am Bildschirm auch unmittelbar adjazent angezeigt werden, werden Adressierungen dazu eingesetzt, eine Rekonstruierbarkeit dieser Bezüge durch die Adressaten zu ermöglichen.

(84) Ausschnitt aus einem Chat-Mitschnitt (gekürzt):

4 Bates: *wir müssen aus baden raus, jasmin*
5 Jasmin: *weiß net..aber so weit ist stuttgart ja nicht weg*
[…]
8 Jasmin: *ja und dann auch noch nach schwaben..das ist eigentlich ne große überwindung für einen urbadner *g**
[…]
18 Bates: *ob die uns leben lassen?*
[…]
24 baloo: *loool@bates ich als franke trau mich auch nüber und hab es überlebt*
[…]
28 Biene: *baloo: ihr franken seid jaeh so ein völkchen *grins* nicht bös gemein, mein freund ist ja auch einer*
29 Jasmin: *@biene *gg**

Die Beiträge in Beispiel 84, das einem Dokument aus dem Dortmunder Chat-Korpus entnommen ist, enthalten drei Belege für Adressierungen. In Beitrag 24 ist die Adressierung an das Aktionswort *loool* angehängt; dadurch wird explizit markiert, dass dieser Teil des Beitrags der evaluativen Respondierung eines Vorbeitrags der Chatterin *bates* (Beitrag 18) dient; der folgende Teil des Beitrags (*ich als franke trau mich auch nüber und hab es überlebt*) knüpft daran dann thematisch an und entwickelt das aktuell verhandelte Thema weiter. In den Beiträgen 28 und 29 finden sich Adressierungen, die dem gesamten Beitrag vorangestellt sind; in 29 unter Verwendung eines Adressierungsmarkers, in 28 ohne Marker, dafür mit anschließendem Doppelpunkt, um den Ausdruck als von der syntaktischen Struktur der folgenden Einheiten abgesetzt zu kennzeichnen.

VII. Weitere Phänomene

Die Typologie erfasst unter den Subtypen I–VI ausdrücklich rein *sprachliche* Phänomene, d.h. solche Einheiten, die von den Kommunikationsbeteiligten per Tastatureingabe erzeugt werden. Nicht berücksichtigt sind „Graphik-Smileys“, die von den Nutzern per Auswahl aus einem Menü in ihre Beiträge eingefügt werden oder die vom System auf Basis bestimmter Tastatureingaben automatisch erzeugt werden. Solche – nicht-sprachlichen – Einheiten wären eher im Rahmen eines Vorverarbeitungsschrittes als auf der Ebene der Wortartenannotation zu behandeln; entsprechend sind sie in der hier beschriebenen Typologie nicht berücksichtigt (vgl. unsere Ausführungen zu Phänomentyp VI.1).

In der Typologie ebenfalls nicht berücksichtigt sind Einheiten, bei denen sprachliche Einheiten – häufig von den Autoren gezielt eingesetzt – mit Einheiten der Hypertextstruktur konvergieren. Beispiele für solche Einheiten sind Hashtags sowie Adressierungen in Tweets und in Facebook-Pinnwandeinträgen, die serverseitig automatisch zu Hyperlinks umgewandelt und von den Nutzern gezielt für die thematische Vernetzung ihrer Tweets mit Beiträgen anderer Nutzer (Hashtags in Twitter) bzw. für die Sichtbarmachung ihrer Beiträge auf den individuellen Startseiten der spezifizierten Adressaten (Adressierungen in Twitter und Facebook) genutzt werden. Die Beschreibung solcher Einheiten spielt für die Analyse von Kommunikationsbeiträgen in Genres internetbasierter Kommunikation zweifelsohne eine wichtige Rolle; um die Besonderheiten hypertextuell vernetzten Kommunizierens in Korpora zu beschreiben, können diese Einheiten aber nicht auf lediglich eine ihrer verschiedenen Funktionen – z.B. die sprachliche, die technische, die hypertextuelle – reduziert werden. Vielmehr gewinnen sie ihre spezifische Funktion im Rahmen der Kommunikation gerade durch die Konvergenz sprachlicher, technischer und struktureller Eigenschaften. In Beispiel 85 ist beispielsweise das Token „cornelsenverlag“ zugleich (i) ein Eigenname (in der „realen Welt“), (ii) der Name eines anderen Twitter-Nutzers (vermutlich der Twitter-Dependance des betreffenden Verlags), (iii) ein durch <@> gekennzeichneter Adressierungsausdruck, mit welchem die Adressierung des Tweets an den genannten Nutzer angezeigt wird, und (iv) ein Hyperlink, der dem Ausdruck vom System aufgrund der Verbindung mit dem <@>-

Zeichen automatisch hinzugefügt wird und der auf die Profilseite des adressierten Nutzers verweist. Durch die Kombination mit dem <@>-Zeichen fungiert der gesamte Ausdruck <@cornelsenverlag> darüber hinaus zugleich (v) als ein Kommando an das System, den Tweet in die persönliche Timeline des Nutzers „cornelsenverlag“ zu integrieren; loggt sich der Nutzer das nächste Mal in Twitter ein, findet er den Tweet dort vor.

(85) Tweet-Nachricht auf twitter.com:

*Erfreulich, dass ich Vertreter der **@cornelsenverlag** e auf Veranstaltungen wie dem **#slml13** der **@werkstatt_bpb** antreffe. Das macht Mut.*

Analoges gilt in Beispiel 85 für den Adressierungsausdruck „@werkstatt_bpb“ und für das Hashtag „#slml13“, mit welchem vom Autor des Tweets eine thematische Vernetzung mit den Tweets anderer Nutzer erzeugt wird, die sich ebenfalls auf das Thema „slml13“ beziehen.[4]

Vermutlich dürfte es sinnvoll sein, Einheiten dieser Art auf einer der Wortarten-annotation vor- oder nachgeordneten Verarbeitungs- und Modellierungsebene als Einheiten zu beschreiben, die zwischen der rein sprachlichen Ebene der Kommunikation und der hypermedialen Struktur der betreffenden Kommunikationsplattformen vermitteln und in denen die strategische Nutzung von Möglichkeiten zur technischen Vernetzung mit Formen der sprachlichen Referenzierung (von Themen und Adressaten) konvergieren.[5]

4 Automatische Verarbeitung sprachlicher Besonderheiten in der internetbasierten Kommunikation: Datengestützter Problemaufriss

In diesem Abschnitt geben wir einen Aufriss typischer Probleme bei der automatischen Wortartenannotation von Sprachdaten aus Genres internetbasierter Kommunikation mit Verarbeitungswerkzeugen, die von linguistischen Anwendern „off the shelf“ über die Oberfläche der webbasierten Analyseplattform *WebLicht* (https://weblicht.sfs.uni-tuebingen.de/) genutzt werden können.

Wir werden zeigen, dass nur ein kleiner Teil der Herausforderungen, die mit der automatischen linguistischen Analyse solcher Daten verbunden sind, mit dem Fehlen geeigneter Kategorien im verwendeten POS-Tagset zu tun hat. Ein großer Teil der sprachlichen Einheiten ist auf Basis vorhandener POS-Kategorien – zumindest theoretisch – schon jetzt zutreffend klassifizierbar. Aufgrund nicht-standardkonformer Formmerkmale werden aber in vielen Fällen Tokens, die bei einer intellektuellen Klassifikation problemlos einer gängigen POS-Kategorie zugeordnet werden könnten, nicht zuverlässig erkannt. Weitere Verarbeitungsprobleme ergeben sich auf der Ebene der automatischen Tokenisierung und resultieren in Segmentierungen, deren Resultate sich auf höheren Verarbeitungsebenen nicht sinnvoll linguistischen Kategorien zuordnen lassen.

Zunächst beschreiben wir das für unsere Verarbeitungsexperimente zusammengestellte Evaluationsdatenset und die verwendeten Verarbeitungswerkzeuge. Anschließend fassen wir die im Zuge unserer Experimente festgestellten Verarbeitungsprobleme zu Problemtypen

zusammen. Dabei wird deutlich, auf welchen Ebenen des Verarbeitungsprozesses Bemühungen zur Verbesserung der Verarbeitungsergebnisse ansetzen können (und sollten) und für welche der in Abschnitt 3 vorgestellten Phänomene eine Erweiterung des POS-Tagsets erforderlich erscheint. In Abschnitt 5 werden wir davon ausgehend dann konkrete Vorschläge für Modifikationen und Erweiterungen des „Stuttgart-Tübingen Tagset“ (STTS) formulieren.

4.1 Evaluationsdatenset und verwendete Sprachverarbeitungswerkzeuge

Das Evaluationsdatenset umfasst manuell zusammengestellte Belegsammlungen für ausgewählte Typen sprachlicher Phänomene in der internetbasierten Kommunikation. Berücksichtigt sind verschiedene Phänomentypen aus den in Abschnitt 3 unterschiedenen Phänomenbereichen *Geschriebene Umgangssprache*, *IBK-typische Akronyme* und *IBK-spezifische interaktive Einheiten*. Die Belege für die untersuchten Phänomene sind zu gleichen Teilen dem Dortmunder Chat-Korpus (Beißwenger 2013) und Diskussionsseiten der deutschsprachigen Wikipedia entnommen. Für jedes der Genres *Chat* und *Wikipedia-Diskussion* enthält das Datenset Subsets mit jeweils 100 Belegen für jeden der untersuchten Phänomentypen. Insgesamt umfasst das Evaluationsdatenset somit 1.000 Belege für das Vorkommen der untersuchten Phänomene (Tabelle 1).

Phänomentyp		**Belege**		
		Chat	Wikipedia-Diskussionen	DWDS
III.1+ III.3	Geschriebene Umgangssprache I: Umgangssprachlich fundierte Wortschreibungen und umgangssprachliche Lexik	100	100	100
III.2	Geschriebene Umgangssprache II: Kontraktierte Formen	100	100	
V	IBK-typische Akronyme	100	100	
VI.1	IBK-spezifische interaktive Einheiten I: Emoticons	100	100	
VI.2	IBK-spezifische interaktive Einheiten II: Aktionswörter	100	100	
Belege gesamt (nur IBK)		**1.000**		

Tab. 1: Zusammensetzung des Evaluationsdatensets.

Für den ersten Phänomentyp wurden zusätzlich in gleichem Umfang standardsprachliche Pendants im Kernkorpus des 20. Jahrhunderts des Projekts „Digitales Wörterbuch der Deutschen Sprache“ (DWDS, Geyken 2007) erhoben. Ein Beleg enthält im Kontext eines Nutzerbeitrags (Posting) mindestens eine Instanz des jeweils untersuchten Phänomentyps (Beispiel 86). Nur für diesen Ausdruck wird im Folgenden der Output der Sprachverarbeitungswerkzeuge untersucht und bewertet.

(86) Beleg für die Verwendung eines Aktionsworts (**räusper**) aus dem Dortmunder Chat-Korpus nach Verarbeitung mit dem TreeTagger:

```
*räusper*/ADJA  Hömma/NN  woher/PWAV  kommste/VVFIN  denn/ADV ?/$.
Ick/ADJD  bin/VAFIN  aus/APPR  Do-Stadt/NN ,/$,  net/ADJD
aus/APPR  Berlin./NE
```

Die Verarbeitung des Testdatensets wurde mit Werkzeugen durchgeführt, die in der webbasierten Annotationsumgebung *WebLicht* (https://weblicht.sfs.uni-tuebingen.de/, Hinrichs/Zastrow/Hinrichs 2010) zur Verfügung stehen. *WebLicht* wurde im Rahmen des D-SPIN-Projekts maßgeblich am Seminar für Sprachwissenschaft der Universität Tübingen entwickelt und wird derzeit im Rahmen des Projekts CLARIN-D weiter ausgebaut.[6] Die Umgebung ermöglicht einen einfachen, Webservice-basierten Zugriff auf eine Vielzahl gängiger Sprachverarbeitungswerkzeuge, die somit nicht mehr lokal installiert und konfiguriert werden müssen, sondern direkt online aufgerufen und auf Daten angewendet werden können.

Die Tokenisierung und das POS-Tagging der Evaluationsdaten erfolgte in zwei voneinander unabhängigen Durchgängen mithilfe des TreeTaggers (Schmid 1994) und des POS-Taggers aus dem OpenNLP-Projekt (Modell: MaxEnt, Trainingsdaten: TIGER-Korpus; http://opennlp.apache.org) sowie den jeweils zugehörigen Tokenisierern (Abb. 3). Die Ergebnisse der Tokenisierung wurden zunächst separat evaluiert, um solche Probleme im Verarbeitungsprozess zu dokumentieren, die bereits auf Tokenisierungsebene auftreten. Anschließend wurde der Output der Tokenisierer für alle Datensätze manuell überprüft und nachbearbeitet, um als Input für das POS-Tagging normalisierte Daten zur Verfügung stellen zu können.

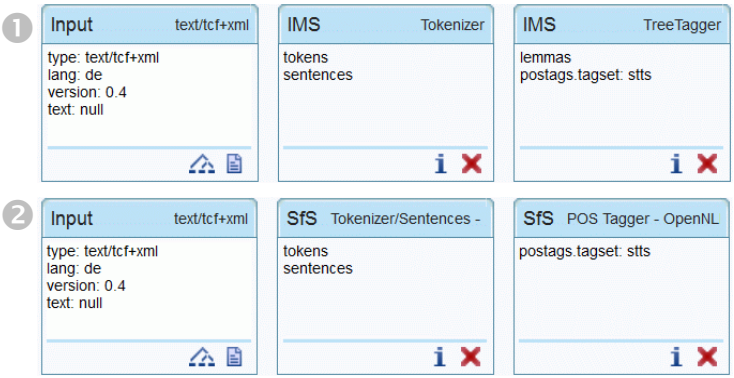


Abb. 3: Verwendete Verarbeitungsketten (Toolchains) und -werkzeuge, zusammengestellt über *WebLicht*.

4.2 Testergebnisse und Problemtypen

Bei den Tests der oben vorgestellten Verarbeitungsketten auf unseren Evaluationsdaten haben sich im Wesentlichen drei Einbruchstellen herauskristallisiert, an denen – teilweise in

Abhängigkeit zu bestimmten Phänomentypen – Verarbeitungsfehler entstehen und die somit für die Qualität der Tagging-Ergebnisse von zentraler Bedeutung sind. Ausgehend von diesen Einbruchstellen, die sich auf unterschiedliche Ebenen des Verarbeitungsprozesses und die darin genutzten Ressourcen beziehen, werden im Folgenden drei Problemtypen abgeleitet und auf der Basis der Testergebnisse konkretisiert.

4.2.1 Segmentierungsprobleme

Sprachliche Formen, die prinzipiell mithilfe vorhandener Kategorien im STTS klassifizierbar wären, werden teilweise schon bei der Wort- und Satzgrenzenerkennung als Einheiten repräsentiert, die sich nicht sinnvoll weiter analysieren lassen. Im einfachen Fall sind die Segmentierungsprobleme hauptsächlich durch die Auslassung von Spatien verursacht, die in Abschnitt 3 (Phänomentyp I.1) als typisches Schnellschreibphänomen beschrieben wurden. Da das Vorhandensein von Spatien für gängige Tokenisierer ein zentrales Kriterium für die Identifizierung von Tokengrenzen darstellt, kommt es in solchen Fällen entsprechend zu nicht sinnvollen Segmentierungen. Bei den Datensets zur geschriebenen Umgangssprache und zu IBK-typischen Akronymen, in denen 1–11% der Tokens vom Tokenisierer nicht korrekt zugeschnitten werden (s. Tabelle 2), stellen fehlende Spatien die Hauptursache für Segmentierungsfehler dar. Segmentierungsprobleme aufgrund fehlender Spatien illustriert Beispiel 87.

Eklatant schlechter werden die Segmentierungsergebnisse dann, wenn nicht nur die Tokengrenzen zum Problem werden, sondern wenn das Token als solches vom Tokenisierer nicht zuverlässig konstituiert werden kann. Im Falle von Emoticons wie z.B. „:-))“ oder „=o)“ werden bestimmte Typen von Zeichen – nämlich Interpunktions-, Buchstaben- und Sonderzeichen – zu Einheiten kombiniert, die nach den Segmentierungsregeln der Tokenisierer z.T. bereits selbst Tokenstatus haben können. Entsprechend führt die Anwendung dieser Regeln in vielen Fällen zu einer zu feinkörnigen Segmentierung, bei der Emoticons in Sequenzen von Interpunktions- und alphanumerischen Zeichen zerlegt werden (Beispiel 88). Im Chat-Subset wurden weniger als die Hälfte der untersuchten Formen dieses Typs korrekt segmentiert, im Subset mit den Wikipedia-Diskussionen sogar weniger als ein Viertel (s. Tab. 2).

Bei den Belegen mit Aktionswörtern, zu denen wir Inflektive und mehrteilige Inflektivkonstruktionen zählen, bereitete die Repräsentation der paarigen Asteriske, die diese Einheiten häufig umschließen (z.B. „*freu*“ oder „*baff bin*“), Probleme. Bei fast allen Belegen im Evaluationsdatenset werteten die Tokenisierer die Asteriske nicht als eigene Tokens, sondern als Bestandteile der adjazenten sprachlichen Ausdrücke (Beispiel 89).

(87) Segmentierungsprobleme: Der gesuchte Ausdruck „biste“ wurde aufgrund fehlenden Spatiums nicht als Token konstituiert, weiterer Fehler im Kontext (Output des Tokenisierers aus der TreeTagger-Toolchain):

```
wieso <token ID="t155">stoeps?biste</token> losgerannt einkaufen
udn ahst vergessen dich anzuziehen <token ID="t163">vorher?*G*
</token>
```

(88) Segmentierungsprobleme: Segmentierung eines Emoticons (Output des Tokenisierers aus der TreeTagger-Toolchain):

```
<token ID="t64">:</token>
<token ID="t65">)</token>
<token ID="t66">)</token>
<token ID="t67">)</token>
```

(89) Segmentierungsprobleme: Segmentierung eines mehrteiligen Aktionsworts (Output des Tokenisierers aus der TreeTagger-Toolchain):

```
<token ID="t946">*ins</token>
<token ID="t947">Bett</token>
<token ID="t948">fall*</token>
```

Phänomene	TreeTagger	OpenNLP-Tagger	Datenset
Geschriebene Umgangssprache I: Umgangssprachlich fundierte Wortschreibungen und umgangssprachliche Lexik	99 von 100	100 von 100	Wikipedia-Diskussionen
	91 von 100	92 von 100	Chat
	100 von 100	100 von 100	DWDS
Geschriebene Umgangssprache II: Kontraktierte Formen	100 von 100	100 von 100	Wikipedia-Diskussionen
	96 von 100	92 von 100	Chat
IBK-typische Akronyme	98 von 100	98 von 100	Wikipedia-Diskussionen
	89 von 100	92 von 100	Chat
IBK-spezifische interaktive Einheiten I: Emoticons	23 von 100	22 von 100	Wikipedia-Diskussionen
	48 von 100	45 von 100	Chat
IBK-spezifische interaktive Einheiten II: Aktionswörter	9 von 100	9 von 100	Wikipedia-Diskussionen
	0 von 100	0 von 100	Chat

Tab. 2: Korrekt segmentierte Instanzen von Emoticons und Aktionswörtern.

Die Segmentierungsprobleme in Bezug auf Emoticons und Aktionswörter waren insofern erwartbar, als die verwendeten Tokenisierer bislang nicht für den Umgang mit Phänomenen dieser Art angepasst wurden. Eine sinnvolle Tokenisierung solcher Einheiten ist keine triviale Aufgabe. Abhängig davon, ob eine Kombination aus Interpunktions- und alphanumerischen Zeichen in konventioneller Weise oder als Emoticon verwendet ist, muss entweder die Einheit als ganze oder jedes einzelne Zeichen als Token repräsentiert werden. Zudem existieren zu Emoticons homonyme Zeichenkombinationen, bei denen ein Teil der Sequenz den Status von Interpunktionszeichen, ein anderer Teil eine Funktion hat, die sich weder als Interpunktionszeichen noch als Emoticon fassen lässt. Beispiel 90 ist Storrer (2013) entnommen und entstammt der Wikipedia. Die Zeichenfolge repräsentiert ein noch ausstehendes Fußballergebnis; dass es sich weder um ein Emoticon noch um eine reine Sequenz von Interpunktionszeichen handelt, lässt sich nur durch eine Analyse des Kontexts disambiguieren.

(90) *Niederlande − Finnland* ***-:- (-:-)*** *Samstag, 29. August 2009, 17:30*
(Beispiel aus Storrer 2013)

4.2.2 Klassifizierungsprobleme

Von den *Segmentierungsproblemen*, die im Wesentlichen durch Schnellschreibphänomene und die für Tokenisierer „irreguläre" Nutzung von Interpunktions- und Sonderzeichen für den Aufbau IBK-spezifischer interaktiver Einheiten verursacht werden, unterscheiden wir die *Klassifizierungsprobleme*. Klassifizierungsprobleme ergeben sich auf der Ebene des POS-Tagging und bestehen darin, dass bestimmte, vom Tokenisierer korrekt als Tokens konstituierte Einheiten aufgrund nicht-standardkonformer Formmerkmale nicht mit dem Tag für die POS-Kategorie versehen werden können, der sie angehören.

Klassifizierungsprobleme ergeben sich im untersuchten Datenset zum einen für den Bereich der geschriebenen Umgangssprache (Phänomentyp III), zum anderen für die IBK-typischen Akronyme (Phänomentyp V).

Um die Klassifikation umgangssprachlich fundierter Wortschreibungen und umgangssprachlicher Lexik in unserem Evaluationsdatenset mit der Klassifikation entsprechender standardsprachlicher Formen zu vergleichen, haben wir ein Datenset mit standardsprachlichen Entsprechungen aus dem DWDS-Korpus zusammengestellt (s. Tabelle 1). Während die POS-Tagger die untersuchten Wortformen im DWDS-Datensatz in 87% (TreeTagger) bzw. 83% der Fälle (OpenNLP-Tagger) korrekt klassifizieren, erreichen sie für die entsprechenden Formen im IBK-Dastenset eine Genauigkeit von nur 34% und 44% (für die Belege aus Wikipedia-Diskussionen) bzw. 13% und 15% (für die Belege aus dem Dortmunder Chat-Korpus) (Tabelle 3). Beispiel 91 zeigt umgangssprachlich fundierte Wortschreibungen aus einer Wikipedia-Diskussionsseite und aus dem Dortmunder Chat-Korpus im Vergleich zu einer standardsprachlichen Entsprechung aus dem DWDS-Korpus plus die vom TreeTagger für die Wortformen jeweils vergebenen Tags:

(91) *Schaden kann **dat/ADJD** ja nich*
(Beispiel aus den Wikipedia-Diskussionen)

*syno **det/ADJA** is to wenig*
(Beispiel aus dem Dortmunder Chat-Korpus)

***das/PDS** ist schon kraftraubend*
(Beispiel aus dem DWDS-Korpus)

Im Falle der IBK-typischen Akronyme werden von den Taggern nur maximal 21% der untersuchten Formen korrekt zugeordnet (exemplarische Fehlanalysen s. Beispiel 92). Für Akronyme sind im STTS zwar keine eigenen Tags vorgesehen, die STTS-Guidelines sehen dafür aber eine Verfahrensweise vor:

> Abgekürzte Wortformen werden getaggt wie die ausgeschriebene Form. Mehrteilige, nicht durch Spatien getrennte Abkürzungen werden entsprechend ihrer syntaktischen Funktion klassifiziert. (Schiller et al. 1999: 9)

Um die Anwendung dieses Verfahrens auf IBK-typische Akronyme zu optimieren, müssen die Tagger auf den Umgang mit solchen Akronymen trainiert werden (z.B. auf Basis einer Liste).

(92) Klassifizierungsprobleme: POS-Tagging bei IBK-typischen Akronymen (Output des TreeTaggers):

***kA/NN** was das kostet*

***btw/ADJA** ... ward ihr denn auch alle fleißig wählen gewesen?*

***re/VVFIN** alle die ich eben schon begrüßt hatte.*

***Imho/NE** sind hier recht viele sogenannte Spaßformen aufgeführt*

Phänomentyp	TreeTagger	OpenNLP-Tagger	Datenset
Geschriebene Umgangssprache	34 von 100	44 von 100	Wikipedia-Diskussionen
	13 von 100	15 von 100	Chat
	87 von 100	83 von 100	DWDS
IBK-typische Akronyme	8 von 100	21 von 100	Wikipedia-Diskussionen
	10 von 100	17 von 100	Chat

Tab. 3: Korrekt vergebene POS-Tags für Fälle geschriebener Umgangssprache und für IBK-typische Akronyme.

4.2.3 Kategorienprobleme

Während im Falle von *Klassifizierungsproblemen* vorhandene Kategorien aufgrund von Formeigenschaften der analysierten Tokens nicht korrekt zugeordnet werden können, beruht die inkorrekte Klassifizierung von Tokens im Falle von *Kategorienproblemen* darauf, dass die relevanten Einheiten im Tagset überhaupt nicht kategorial repräsentiert sind. Entsprechend lassen sich Tokens, die diesen Einheiten zugehören, mit gängigen POS-Taggern also – erwartbar – noch gar nicht sinnvoll klassifizieren. Die Ursache des Problems liegt dabei auf der Ebene des Tagsets und nicht auf der Ebene der Tagger.

Kategorienprobleme bestehen im Falle der internetbasierten Kommunikation für zwei Typen von Einheiten: zum einen für die IBK-spezifischen interaktiven Einheiten (Emoticons, Aktionswörter, Adressierungen; Phänomentyp VI), zum anderen für die umgangssprachlichen kontraktierten Formen (Phänomentyp III.2). Letztere stellen zwar keinen IBK-spezifischen Phänomentyp dar, insofern sie auch in der gesprochenen Umgangssprache hoch frequent sind; in redigierten Texten, auf deren Verarbeitung die gegenwärtige Version des STTS primär optimiert ist, kommen sie aber bestenfalls in Ausnahmefällen vor.

Die jeweils 100 Instanzen von Emoticons und Aktionswörtern in unserem Evaluationsdatenset werden von den beiden Taggern mit ganz unterschiedlichen POS-Tags versehen:

Emoticons erhalten vom TreeTagger in Chats und Wikipedia-Diskussionen in den meisten Fällen Adjektiv- oder Nomen-Tags (*ADJA/D*, *NN/NE*). Der OpenNLP-Tagger hingegen behandelt die Einheiten häufig wie konventionelle Interpunktionszeichen – in den Belegen aus Wikipedia-Diskussionen sogar überwiegend, in den Chat-Belegen in etwa ähnlich häufig wie Nomen-Tags (Tabelle 4).

Aktionswörter, insbesondere einfache Inflektive, werden vom TreeTagger in 41% der Fälle als Verbformen klassifiziert. Existiert zu einem Aktionsausdruck eine homonyme Imperativform (z.B. **sing**, homonym zu *sing!*), wird diese in etwa der Hälfte der Fälle auch mit dem *VVIMP*-Tag versehen (Tabelle 5). Auch der OpenNLP-Tagger klassifiziert Aktionswörter häufig als Verbformen, vergibt allerdings am häufigsten (in 30% der Fälle) das Tag *XY* für „Nichtwort“, das gemäß STTS-Guidelines „bei größeren Symbolgruppen, [...] sowie Kombinationen aus Ziffern und Zeichen, die sich nicht als *CARD* oder *ADJA* einordnen lassen“, das Mittel der Wahl darstellt (Schiller et al. 1999: 74f.). Weiterhin vergeben beide Tagger für Aktionswörter häufig Adjektiv-Tags (Tabelle 4).

Phänomentyp	TreeTagger		OpenNLP-Tagger		Datenset
Emoticons	ADJA/D	52 von 100	$./$,/$(	68 von 100	Wikipedia-Diskussionen
	NN/NE	43 von 100	NN/NE	23 von 100	
	VVFIN	3 von 100	VV*	5 von 100	
	ADJA/D	52 von 100	NN/NE	40 von 100	Chat
	NN/NE	43 von 100	$./$(	36 von 100	
	CARD	4 von 100	XY	15 von 100	
Aktionswörter	VV*	41 von 100	XY	30 von 100	Wikipedia-Diskussionen
	NN/NE	32 von 100	VV*	24 von 100	
	ADJA/D	25 von 100	ADJA/D	22 von 100	
	VV*	41 von 100	XY	46 von 100	Chat
	ADJA/D	32 von 100	VV*	23 von 100	
	NN/NE	26 von 100	ADJA/D	20 von 100	

Tab. 4: Am häufigsten vergebene POS-Tags für Instanzen von Emoticons und Aktionswörtern.

Phänomentyp	TreeTagger	OpenNLP-Tagger	Datenset
Aktionswörter, zu denen es im verbalen Paradigma homonyme Imperativformen gibt	34 von 58	0 von 58	Wikipedia-Diskussionen
	28 von 47	1 von 47	Chat

Tab. 5: Vergebene VVIMP-Tags zu denjenigen Aktionswörtern, zu denen eine homonyme Imperativform existiert.

Auch für die Annotation umgangssprachlicher kontraktierter Formen gibt es im STTS derzeit noch keine geeignete Kategorie. Zwar existiert mit *APPRART* eine Kategorie für „Präpositionen mit inkorporiertem Artikel"; die Beschreibung der Kategorie (Schiller et al. 1999: 67) nennt als typische Vertreter aber Formen wie *am*, *zur*, *zum* und *ans*, die (auch) standardsprachlich etabliert sind, d.h. ihre umgangssprachliche Markiertheit verloren haben und überdies einen hohen Grammatikalisierungsgrad aufweisen. Zudem ist die Kategorie auf das Kontraktionsmuster Präposition + Artikel festgelegt. Umgangssprachlich werden Verschmelzungen aber auch nach diversen weiteren Mustern gebildet (vgl. die exemplarische Liste in Abschnitt 3, Phänomentyp III.2).

Das Subset „Kontraktierte Formen" aus unserem Evaluationsdatenset umfasst ausschließlich Formen mit verbaler erster Komponente. Am häufigsten werden diese von den beiden getesteten Taggern mit Tags für Verbformen versehen (*VVFIN*, *VAFIN*, *VMFIN*, *VVIMP*; Tabelle 6), wenngleich der TreeTagger fast ebenso häufig auch das NN-Tag vergibt. Eine künftige erweiterte Version des STTS sollte – wie dies auch von Gimpel et al. (2011) für das Englische unternommen wurde – eine eigene Kategorie für kontraktierte Formen vorsehen. Eine solche Kategorie könnte nicht nur für die linguistische Analyse von Sprachdaten aus Genres internetbasierter Kommunikation, sondern auch für das Tagging von Korpora mit Transkripten gesprochener Sprache von Nutzen sein.

Phänomentyp	TreeTagger		OpenNLP-Tagger		Datenset
Kontraktierte Formen	VVFIN	35 von 100	VVFIN	26 von 100	Wikipedia-Diskussionen
	VVIMP	1 von 100	VMFIN	10 von 100	
	NN	34 von 100	VAFIN	6 von 100	
	VVFIN	41 von 100	VVFIN	34 von 100	Chat
	VMFIN	6 von 100	VAFIN	11 von 100	
	NN	32 von 100	VMFIN	6 von 100	

Tab. 6: Am häufigsten vergebene POS-Tags für Instanzen von kontraktierten Formen.

5 Lösungsperspektiven und Vorschläge zur Modifikation des Stuttgart-Tübingen-Tagset (STTS) für die Annotation von Korpora internetbasierter Kommunikation

Die Ergebnisse aus den in Abschnitt 4 beschriebenen Tests zur automatischen Wortartenannotation von Sprachdaten aus Genres internetbasierter Kommunikation haben gezeigt, dass Probleme an unterschiedlichen Stellen des automatischen Verarbeitungsprozesses auftreten können:

1) **Segmentierungsprobleme** ergeben sich dadurch, dass auf der Ebene der automatischen Tokenisierung Zeichenfolgen als Tokens konstituiert werden, die in Verarbeitungsprozessen wie dem POS-Tagging, die tokenisierte Daten als Input nutzen, nicht sinnvoll weiter analysiert werden können. Wird das Tokenisierungsergebnis nicht zunächst manuell normalisiert, ergeben sich dadurch am Ende der Verarbeitungskette Fehler bei der Zuordnung von POS-Kategorien, die nicht dem POS-Tagger, sondern dem Tokenisierer anzulasten sind.

   Typische Ursachen für Segmentierungsprobleme sind die irreguläre Verwendung von Spatien sowie die Nutzung von Interpunktions- und Sonderzeichen für die Bildung von Emoticons und für die Kennzeichnung von Aktionswörtern.

   Lösungsmöglichkeiten für Probleme bei der automatischen Segmentierung sind entweder ein Training der Tokenisierungswerkzeuge auf den Umgang mit nicht-standardkonformer Schriftlichkeit und IBK-spezifischen lexikalischen Besonderheiten oder eine Normalisierung der Tokenisierungsergebnisse.

2) **Klassifizierungsprobleme** ergeben sich auf der Ebene des POS-Tagging und bestehen darin, dass bestimmte Tokens, für die im verwendeten POS-Tagsets geeignete Kategorien existieren, aufgrund nicht-standardkonformer Formmerkmale nicht mit dem entsprechenden Tag versehen werden können. Um Klassifizierungsprobleme als solche identifizieren zu können, muss sichergestellt sein, dass die Input-Daten für den Tagger eine Tokenisierung aufweisen, die frei von Segmentierungsproblemen ist.

   Ursachen für Klassifizierungsprobleme in Sprachdaten aus Genres internetbasierter Kommunikation sind u.a. Phänomene geschriebener Umgangssprache auf der Ebene der Orthographie und der Lexik sowie IBK-typische Akronyme. Fehler beim POS-Tagging, die sich durch Schnellschreibphänomene, Phänomene der graphischen Nachbildung suprasegmentaler Elemente der gesprochenen Sprache oder Verfremdungsschreibungen ergeben, fallen ebenfalls unter diesen Problemtyp.

Lösungsmöglichkeiten für Probleme bei der automatischen POS-Klassifizierung sind entweder ein Training der POS-Tagger auf den Umgang mit den entsprechenden Phänomenen oder eine Normalisierung der Input-Daten in einer dem POS-Tagging vorgeordneten Aufbereitungsphase.

3) Im Fall von **Kategorienproblemen** beruht die inkorrekte Zuordnung von POS-Kategorien zu Tokens darauf, dass für die Zielkategorien im Tagset keine Tags vorgesehen sind. Einheiten in Sprachdaten internetbasierter Kommunikation, für die im STTS bislang keine geeigneten Kategorien existieren, sind die IBK-spezifischen interaktiven Einheiten (mit den Subtypen Emoticons, Aktionswörter und Adressierungen) sowie umgangssprachliche kontraktierte Formen (*haste*, *biste*, *willste*, *machstes*; *aufm*; *isn* usw.).

   Durch Kategorienprobleme verursachte „Fehler" bei der Zuordnung von POS-Tags lassen sich nur durch eine Erweiterung des Tagsets lösen. Im Falle der kontraktierten Formen ist alternativ auch eine Normalisierung der Daten in einem Vorverarbeitungsprozess denkbar, bei dem die kontraktierten Formen entsprechend ihren standardsprachlichen Pendants (künstlich) aufgelöst werden (z.B. *biste* ⇨ *bist du*, *machstes* ⇨ *machst du es*, *aufm* ⇨ *auf dem*, *isn* ⇨ *ist denn* usw.) – dies aber verbunden mit einem Verlust der Information, dass es sich bei den dann künstlich erzeugten Varianten in den Originaldaten um typisch sprechsprachliche kontraktierte Formen handelt.

Die drei Problemtypen können einander überlagern. Abb. 4 ordnet die in Abschnitt 3 unterschiedenen Typen von sprachlichen Besonderheiten in der internetbasierten Kommunikation den verschiedenen Lösungsperspektiven zu.

Phänomentyp	**Bearbeitung der durch die Phänomene verursachten Verarbeitungsprobleme durch …**			
	Anpassung des Tokenisierers	Anpassung des POS-Taggers	Normalisierung der Tokenisierung	Erweiterung des Tagsets
I. Schnellschreibphänomene				
I.1 Irreguläre Verwendung von Spatien	√		√	
I.2 Tippfehler	√		√	
I.3 Ökonomiebedingte Abweichungen von den Normen für die Groß- und Kleinschreibung		√	√	
II. Graphische Nachbildung suprasegmentaler Elemente der gesprochenen Sprache		√	√	
III. Geschriebene Umgangssprache				
III.1 Umgangssprachlich fundierte Wortschreibungen		√	√	
III.2 Umgangssprachliche kontraktierte Formen			√	√
III.3 Umgangssprachliche Lexik		√		
IV. Verfremdungsschreibungen		√	√	
V. IBK-typische Akronyme		√		
VI. IBK-spezifische interaktive Einheiten				
VII.1 Emoticons	√	√		√
VII.2 Aktionswörter	√	√		√
VII.3 Adressierungen	√	√		√

Abb. 4: Sprachliche Besonderheiten in der internetbasierten Kommunikation und Perspektiven ihrer Behandlung bei der Verarbeitung mit Werkzeugen für die automatische linguistische Analyse.

Im Folgenden formulieren wir Vorschläge zur Modifikation und Erweiterung des STTS in Hinblick auf die Behandlung solcher Einheiten in der internetbasierten Kommunikation, die zum gegenwärtigen Stand der Kunst beim POS-Tagging Kategorienprobleme verursachen.

5.1 Emoticons, Aktionswörter, Adressierungen

Emoticons, Aktionswörter und Adressierungen können als durch die internetbasierte Kommunikation hervorgebrachte (Emoticons, Adressierungen) bzw. für eine breite Nutzung in schriftlicher interpersonaler Kommunikation adaptierte (Aktionswörter) Erweiterungen des lexikalischen Inventars gelten. Um sie in einem Kategoriensystem für das Tagging von Wortarten darstellbar zu machen, sind sie sinnvollerweise nicht als eine in der Luft hängende neue Kategorie zu konstituieren, sondern zu vorhandenen Kategorien in Beziehung zu setzen. Die Einordnung in einen grammatischen Beschreibungsrahmen sollte dabei den Funktionen Rechnung tragen, die sie in dialogischer Kommunikation übernehmen und mit denen sie die Möglichkeiten schriftlicher Kommunikation spezifisch erweitern.

In Abschnitt 3 (Phänomentyp VI) haben wir – ausgehend von den Vorschlägen in Beißwenger et al. (2012) – Emoticons, Aktionswörter und Adressierungen als IBK-spezifische Erweiterungen der Kategorie der *interaktiven Einheiten* dargestellt, die in der Konzeption der GDS (Zifonun et al. 1997) die Interjektionen (*ach*, *äh*, *mhm*, *ne*, *tja*) und die Responsive (*ja*, *nein*, *okay*) umfasst. Emoticons, Aktionswörter und Adressierungen erweitern diese Kategorie um Einheiten, die auf die Erfordernisse der Handlungskoordination und der emotionalen Kommentierung in dialogischer schriftlicher Kommunikation spezialisiert sind und die die Möglichkeiten schriftlicher Kommunikation im (zeitlichen und sozialen) Nahbereich ausbauen.

Um diese Einheiten in einer erweiterten Version des STTS-Kategoriensystems darstellbar zu machen, schlagen wir eine Restrukturierung desjenigen Systemausschnitts vor, der bislang durch die Kategorien *ITJ* (Interjektion) und *PTKANT* (Antwortpartikel) repräsentiert wird:

1. Die Kategorie *ITJ* verliert ihren Status als Hauptkategorie, die Hauptkategorie *PTK* (Partikeln) bleibt erhalten, wird aber um die Antwortpartikeln reduziert.
2. Auf Ebene der Hauptwortarten wird in Anlehnung an die entsprechende Kategorie der GDS das Konzept der *interaktiven Einheiten* als neue Hauptkategorie eingeführt. Die Kategorie wird durch ein eigenes Hauptkategorien-Tag dargestellt – zum Beispiel *IE*.
3. *ITJ* wird – bei gleichbleibender intensionaler und extensionaler Bestimmung – eine Subkategorie von *IE*. Um die kategoriale Einordnung im Tag selbst anzuzeigen, wird das Tag *ITJ* – entsprechend den Tagnamenkonventionen des STTS – zu *IEITJ* erweitert (lies: „Haupttyp: Interaktive Einheit, Subtyp: Interjektion“).
4. *PTKANT* wird – bei gleichbleibender intensionaler und extensionaler Bestimmung – ebenfalls eine Subkategorie von *IE*. Um die Kategorie der interaktiven Einheiten im STTS konsistent zur entsprechenden Kategorie der GDS darzustellen, wird der Terminus „Antwortpartikel“ ersetzt durch den Terminus „Responsiv“ (*RSP*). Um zudem

die kategoriale Einordnung im Tag selbst anzuzeigen, wird dafür das Tag *IERSP* vergeben (lies: „Haupttyp: Interaktive Einheit, Subtyp: Responsiv“).

Mit den Modifikationen 1–4 ist die Kategorie der interaktiven Einheiten im STTS etabliert. Die dabei vorgenommenen Restrukturierungen des Kategoriensystems sind aus unserer Sicht insofern moderat, als die schon existierenden Kategorien „Interjektion“ und „Antwortpartikel“ nur umbenannt, in ihrem Zuschnitt aber nicht verändert werden. Das STTS wird durch diese Restrukturierung vorbereitet, in einem weiteren Schritt die Emoticons, die Aktionswörter und die Adressierungen als IBK-spezifische Erweiterungen der interaktiven Einheiten aufzunehmen:

5. Einführung der neuen Kategorien *IEEMO* (Emoticons), *IEATW* (Aktionswort) und *IEADR* (Adressierung) als Subkategorien zu *IE* und auf gleicher Ebene wie *IEITJ* und *IERSP*.

Die entsprechenden Ausschnitte aus dem STTS-Kategoriensystem vor und nach der vorgeschlagenen Restrukturierung sind in Abb. 5 gegenübergestellt.

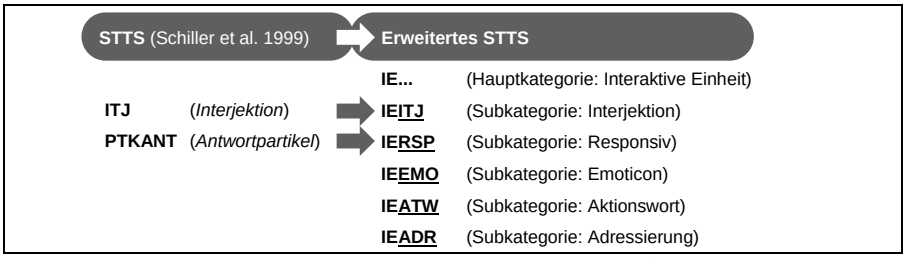


Abb. 5: Restrukturierung des STTS zur Darstellung IBK-spezifischer interaktiver Einheiten.

5.2 Umgangssprachliche kontraktierte Formen

Für die Darstellung umgangssprachlicher kontraktierter Formen (vgl. Phänomentyp III.2 in Abschnitt 3) bieten sich zwei verschiedene Lösungsvarianten an: eine einfache, bei welcher für Formen dieser Art eine neue Hauptkategorie (z.B. *KTR*) ohne weitere Subdifferenzierungen eingeführt wird, die kontraktive Formen aller möglichen Bildungsmuster umfasst. Diese einfache Lösung hätte den Vorteil, dass mögliche künftige Muster der Bildung von Verschmelzungen, die gegenwärtig nicht oder nur schwach produktiv sind, keiner erneuten Änderung des Tagsets bedürften, um im STTS dargestellt werden zu können. Der Nachteil dieser Lösung liegt entsprechend darin, dass das verwendete Tag wenig grammatische Strukturinformation enthält.

Die komplexere Lösungsvariante orientiert sich an der schon vorhandenen STTS-Kategorie *APPRART*, die (stark grammatikalisierte und auch standardsprachlich etablierte) Verschmelzungen des Typs Präposition+Artikel beschreibt und bei der sich das Bildungsmuster (Präposition+Artikel) aus der Benennung des Tags ableiten lässt. Nicht ablesen lässt sich aber, dass es sich bei den damit beschriebenen Einheiten um Verschmelzungen handelt.

Die komplexere Lösungsvariante für die Darstellung umgangssprachlicher kontraktierter Formen im STTS strebt an, wie im Falle von *APPRART* Strukturinformation zur Verschmelzung im Tag selbst zu kodieren. Darüber hinaus soll zusätzlich der Verschmelzungscharakter der beschriebenen Einheiten im Tag angezeigt werden. Dafür sind die folgenden Erweiterungen des Tagsets erforderlich:

1. Einführung einer neuen Hauptkategorie für umgangssprachliche kontraktierte Formen. Die Kategorie wird durch ein eigenes Hauptkategorien-Tag dargestellt – zum Beispiel *KTR*.
2. Einführung einer Reihe von Subkategorien für jeden einzelnen Strukturtyp der unter *KTR* erfassten Einheiten. Die Tags für die Subkategorien sind zusammengesetzt aus dem Namen der Hauptkategorie (*KTR*) als erstem Segment und den Namen derjenigen STTS-Kategorien als weiteren Segmenten, aus deren Einheiten die jeweiligen kontraktierten Formen gebildet sind. Für die in Abschnitt 3 exemplarisch aufgeführten Bildungsmuster ergeben sich somit z.B. die folgenden Kategorien und Tags:

– Präposition + Artikel (*innem, aufm, ausn*): [7]	*KTRAPPRART*
– Adverb + Artikel (*nochn*):	*KTRADVART*
– Konjunktion + Personalpronomen (*fallste, obse*):	*KTRKOPPER*
– Auxiliarverb + Personalpronomen (*haste, biste*):	*KTRVAPPER*
– Vollverb + Personalpronomen (*machste, gehste, denkste*):	*KTRVVPPER*
– Vollverb + zwei Personalpronomina (*machstes, gibstes*):	*KTRVVPPERPPER*
– Kopulaverb + Personalpronomen (*warens*):	*KTRVAPPER*
– Modalverb + Personalpronomen (*kannste, willste, sollste*):	*KTRVMPPER*
– Auxiliarverb + Abtönungspartikel (*isn*):	*KTRVAPTK*

Die vergebenen Tags wären in diesem Fall ungleich informativer – aber auch deutlich komplexer – als im Falle der zuerst vorgestellten Variante.

6 Fazit und Ausblick

Ziel dieses Beitrags war es, aus linguistischer Sicht die zentralen konzeptuellen Grundlagen bereitzustellen, die für eine Optimierung von Verfahren des Wortartentaggings für Zwecke der Annotation von Korpora zu Genres internetbasierter Kommunikation benötigt werden. Hierzu wurde zum einen eine Typologie sprachlicher Besonderheiten präsentiert, hinsichtlich derer sich die schriftliche Sprachverwendung in der internetbasierten Kommunikation charakteristischerweise von der Schriftlichkeit redigierter Texte (z.B. Zeitungstexte) unterscheidet (Abschnitt 3). In einem zweiten Schritt wurde eine Typologie von Problemen skizziert, die sich beim Wortartentagging von Daten aus Chats und aus Wikipedia-Diskussionsseiten ergeben (Abschnitt 4). Die Probleme betreffen unterschiedliche Ebenen und Ressourcen des Verarbeitungsprozesses; eine Optimierung automatischer Verfahren muss sowohl auf der Ebene der Tokenisierung wie auch auf der Ebene des POS-Taggings ansetzen und bedarf darüber hinaus eines Tagsets, das Erweiterungen für sprachliche Einheiten umfasst, die als spezifisch für die schriftliche internetbasierte Kommunikation gelten können. Ein Vorschlag, wie solche Erweiterungen in

einer modifizierten Version des STTS umgesetzt werden könnten, wurde in Abschnitt 5 formuliert.

Ausgehend von den vorgestellten Phänomen- und Problembeschreibungen und von der vorgeschlagenen STTS-Erweiterung können in einem nächsten Schritt manuell annotierte Trainigsdatensets (Goldstandard) zu Sprachdaten aus Genres internetbasierter Kommunikation aufgebaut und an diesen Sets existierende Tokenisierungs- und POS-Tagging-Verfahren retrainiert oder neu entwickelt werden. Entsprechende Vorhaben werden derzeit in unterschiedlichen Projektzusammenhängen verfolgt – u.a.:

a) im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (*DeRiK*, Beißwenger et al. 2013), in dem derzeit ein linguistisch annotiertes Korpus mit deutschen Sprachdaten aus den wichtigsten Genres internetbasierter Kommunikation aufgebaut wird, das als eine Zusatzkomponente zu den Korpora geschriebener Sprache im Projekt „Digitales Wörterbuch der deutschen Sprache“ (http://www.dwds.de) konzipiert ist (Kooperationsprojekt des Instituts für deutsche Sprache der TU Dortmund mit der DWDS-Arbeitsgruppe an der Berlin-Brandenburgischen Akademie der Wissenschaften);

b) im DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation“ (*Empirikom*, http://www.empirikom.net), in dem seit Frühjahr 2013 eine Shared Task zur automatischen Tokenisierung und Wortartenannotation von Sprachdaten aus der deutschsprachigen internetbasierten Kommunikation vorbereitet wird (Koordination: Michael Beißwenger);

c) im BMBF-Verbundprojekt „Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining“ (*KobRA*, http://www.kobra.tu-dortmund.de), in dem in einer Kooperation von Germanistischer Linguistik und Künstlicher Intelligenzforschung Verfahren des Maschinellen Lernens u.a. für die Adaption von POS-Tagging-Verfahren an die besonderen Anforderungen von Genres schriftlicher internetbasierter Kommunikation eingesetzt werden sollen (Projektleitung: Angelika Storrer).

Anmerkung

Wir danken den anonymen Reviewerinnen und Reviewern für ihre hilfreichen Anregungen und konstruktiven Kommentare zu einer früheren Version dieses Artikels.

Literatur

ANDROUTSOPOULOS, J./ZIEGLER, E. (2003). „Sprachvariation und Internet: Regionalismen in einer Chat-Gemeinschaft.“ In: Androutsopoulos, J./Ziegler, E. (Hrsg.): ‚Standardfragen‘. Soziolinguistische Perspektiven auf Sprachgeschichte, Sprachkontakt und Sprachvariation. Frankfurt: Peter Lang, 251-279.

AVONTUUR, T./BALEMANS, I./ELSHOF, L./VAN NOORD, N./VAN ZAANEN, M. (2012). „Developing a part-of-speech tagger for Dutch tweets.“ In: Computational Linguistics in the Netherlands Journal 2, 34–51.

BEIßWENGER, M. (2000). Kommunikation in virtuellen Welten: Sprache, Text und Wirklichkeit. Stuttgart: ibidem.

BEIßWENGER, M. (2003). „Sprachhandlungskoordination im Chat." In: Zeitschrift für germanistische Linguistik 31 (2), 198-231.

BEIßWENGER, M. (2007). Sprachhandlungskoordination in der Chat-Kommunikation. Berlin/New York: de Gruyter (Linguistik – Impulse & Tendenzen 26).

BEIßWENGER, M. (2013). „Das Dortmunder Chat-Korpus." In: Zeitschrift für germanistische Linguistik 41, H. 1, 161-164. Erweiterte Fassung online unter http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

BEIßWENGER, M./ERMAKOVA, M./GEYKEN, A./LEMNITZER, L./STORRER, A. (2012). „A TEI Schema for the Representation of Computer-mediated Communication." In: Journal of the Text Encoding Initiative (jTEI), Issue 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

BEIßWENGER, M./ERMAKOVA, M./GEYKEN, A./LEMNITZER, L./STORRER, A. (2013). „DeRiK: A German Reference Corpus of Computer-Mediated Communication." In: Literary and Linguistic Computing 2013 (doi: 10.1093/llc/fqt038).

BEIßWENGER, M./STORRER, A. (Hrsg.; 2005). Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte – Werkzeuge – Anwendungsfelder. Stuttgart: ibidem.

BEIßWENGER, M./STORRER, A. (2005a). „Chat-Szenarien für Beruf, Bildung und Medien." In: Beißwenger, M./Storrer, A. (Hrsg.): Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte – Werkzeuge – Anwendungsfelder. Stuttgart: ibidem, 9-25.

BEIßWENGER, M./STORRER, A. (2012). „Interaktionsorientiertes Schreiben und interaktive Lesespiele in der Chat-Kommunikation." In: Zeitschrift für Literaturwissenschaft und Linguistik 168, 92-124.

BIBER, D. et al. (1999). Longman Grammar of Spoken and Written English. Edinburgh: Pearson Education Limited.

BIBER, D./CONRAD, S./LEECH, G. (2002). Longman Student Grammar of Spoken and Written English. Edinburgh: Pearson Education Limited.

BICK, E. (2010). „Degrees of Orality in Speech-like Corpora: Comparative Annotation of Chat and E-mail Corpora." In: Otoguro, R./Ishikawa. K./Umemoto, H./Yoshimoto, K./Harada, Y. (Hrsg.): Proceedings of the 24th Pacific Asia Conference on Language (PACLIC24). Institute for Digital Enhancement of Cognitive Development, Waseda University, 721-729.

BITTNER, J. (2003). Digitalität, Sprache, Kommunikation. Eine Untersuchung zur Medialität von digitalen Kommunikationsformen und Textsorten und deren varietätenlinguistischer Modellierung. Berlin (Philologische Studien und Quellen 178).

BLAKE, B.J. (2008). All About Language. New York: Oxford University Press.

BRINKER, K. (2001). Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. 5., durchges. u. erg. Aufl. Berlin: Erich Schmitt Verlag (Grundlagen der Germanistik 29).

CRYSTAL, D. (2001). Language and the Internet. Cambridge: Cambridge University Press.

DUDEN-4[5] = DUDEN (1995). Die Grammatik. 5. Aufl. Mannheim: Bibliographisches Institut.

DUDEN-4[7] = DUDEN (2005). Die Grammatik. 7. Aufl. Mannheim: Bibliographisches Institut.

DÜRSCHEID, C. (2004). „Netzsprache – ein neuer Mythos?" In: Beißwenger, M./Hopffmann, L./Storrer, A. (Hrsg.): Internetbasierte Kommunikation (Osnabrücker Beiträge zur Sprachtheorie 68), 141–157.

DÜRSCHEID, C. (2005). „Normabweichendes Schreiben als Mittel zum Zweck." In: Muttersprache 115, 40-53.

DÜRSCHEID, C. (2005a). „Medien, Kommunikationsformen, kommunikative Gattungen." In: Linguistik online 22 (1). WWW-Ressource: http://www.linguistik-online.de/22_05/duerscheid.pdf.

EHLICH, K. (1983). „Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung." In: Assmann, A. et al. (Hrsg.): Schrift und Gedächtnis. Archäologie der literarischen Kommunikation I. München: Fink, 24-43.

EHLICH, K. (1984). „Zum Textbegriff." In: Rothkegel, A./Sandig, B. (Hrsg.): Text – Textsorten – Semantik. Hamburg: Buske, 9-25.

GDS = ZIFONUN, G./HOFFMANN, L./STRECKER, B. (1997). Grammatik der deutschen Sprache. 3 Bde. Berlin: de Gruyter (Schriften des Instituts für deutsche Sprache 7.1-7.3).

GEYKEN, A. (2007). „The DWDS corpus: A reference corpus for the German language of the 20th century." In: Fellbaum, C. (Hrsg.): Collocations and Idioms. London: continuum, 23-40.

GIESBRECHT, E./EVERT, S. (2009). „Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus." In: Alegria, I./Leturia, I./Sharoff, S. (Hrsg.): Proceedings of the 5th Web as Corpus Workshop (WAC5), San Sebastian, Spain. WWW-Ressource: http://cogsci.uni-osnabrueck.de/~severt/PUB/GiesbrechtEvert2009_Tagging.pdf

GIMPEL, K./SCHNEIDER, N./O'CONNOR, B. et al. (2011). „Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments." In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11): short papers - Volume 2, 42–47. WWW-Ressource: http://dl.acm.org/ft_gateway.cfm?id=2002747&ftid=994684&dwn=1&CFID=247642444&CFTOKEN=38951532

GREENBAUM, S. (1996). The Oxford English Grammar. New York: Oxford University Press.

HERRING, S. C. (1999). „Interactional Coherence in CMC." In: Journal of Computer-Mediated Communication 4.4. WWW-Ressource: http://jcmc.indiana.edu/vol4/issue4/herring.html

HERRING, S. C. (Hrsg.) (1996). Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives. Amsterdam/Philadelphia: John Benjamins (Pragmatics and Beyond New Series 39).

HERRING, S. C. (Hrsg.) (2010/11). Computer-Mediated Conversation, Part I/II. Special Issue of Language@Internet. WWW-Ressource: http://www.languageatinternet.org/articles/2010

HINRICHS, M./ZASTROW, T./HINRICHS, E. (2010). „WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure." In: Proceedings of the Seventh conference on International Language Resources and Evaluation, Valetta, Malta.

JARBOU, S.O./AL-SHARE, B. (2012). „The Effect of Dialect and Gender on the Representation of Consonants in Jordanian Chat." In: Language@Internet 9. Online: http://www.languageatinternet.org/articles/2012/Jarbou

KILIAN, J. (2001). „T@stentöne. Geschriebene Umgangssprache in computervermittelter Kommunikation. In Chat-Kommunikation." In: Beißwenger, M. (Hrsg.): Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld, Stuttgart: ibidem, 55-78.

KESTEMONT, M./PEERSMAN, C./DE DECKER, B./DE PAUW, G./LUYCKX, K./MORANTE, R./VAASSEN, F./VAN DE LOO, J./DAELMANS, W. (2012). „The Netlog Corpus. A Resource for the Study of

Flemish Dutch Internet Language.“ In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Paris, 1569-1572.

KING, B. W. (2009). „Building and Analysing Corpora of Computer-Mediated Communication.“ In: Baker, P. (Hrsg.): Contemporary corpus linguistics. London: Continuum, 301-320.

KOCH, P./OESTERREICHER, W. (1994). „Schriftlichkeit und Sprache.“ In: Günther, H./Ludwig, O. (Hrsg.): Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung. Band 1. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 12.1), 587-604.

LUCKHARDT, K. (2009). Stilanalysen zur Chat-Kommunikation. Eine korpusgestützte Untersuchung am Beispiel eines medialen Chats. Diss., TU Dortmund. Digitale Ressource: http://hdl.handle.net/2003/26055.

MCARTHUR, T. (Hrsg.) (1998). Concise Oxford Companion to the English Language. Oxford: Oxford University Press.

OWOPUTI, O./O'CONNOR, B./DYER, C./GIMPEL, K./SCHNEIDER, N. (2012). Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical report, Carnegie Mellon University (CMU-ML-12-107). WWW-Ressource: http://ttic.uchicago.edu/~kgimpel/papers/CMU-ML-12-107.pdf

REYNAERT, M./OOSTDIJK, N./DE CLERCQ, O./VAN DEN HEUVEL, H./DE JONG, F.M.G. (2010). „Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus.“ In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Paris: European Language Resources Association (ELRA), 2693-2698.

RITTER, A./CLARK, S./ETZIONI, M./ETZIONI, O. (2001). „Named Entity Recognition in Tweets: An Experimental Study.“ In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), 1524–1534. WWW-Ressource: http://dl.acm.org/ft_gateway.cfm?id=2145595&ftid=1146158&dwn=1&CFID=247642444&CFTOKEN=38951532

SCHIFFRIN, D. (1986). Discourse markers. Cambridge: Cambridge University Press (Studies in interactional sociolinguistics 5).

SCHILLER, A./TEUFEL, S./STÖCKERT, CH./THIELEN, CH. (1999). „Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).“ Technischer Bericht. Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. WWW-Ressource: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

SCHLOBINSKI, P. (2001). „*knuddel – zurueckknuddel – dich ganzdollknuddel*. Inflektive und Inflektivkonstruktionen im Deutschen.“ In: Zeitschrift für germanistische Linguistik 29, 192-218.

SCHMID, H. (1994). „Probabilistic Part-of-Speech Tagging Using Decision Trees.” In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK. WWW-Ressource: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

SCHÖNFELDT, J./GOLATO, A. (2003). „Repair in Chats: A Conversation Analytic Approach.“ In: Research on Language and Social Interaction 36 (3), 241-284.

SCHWITALLA, J. (2006). Gesprochenes Deutsch. Eine Einführung,. 3., neu bearb. Aufl. Berlin: Erich Schmidt Verlag (Grundlagen der Germanistik 33).

STORRER, A. (2000). „Schriftverkehr auf der Datenautobahn. Besonderheiten der schriftlichen Kommunikation im Internet.“ In: Voß, G.G./Holly, W./Boehnke, K. (Hrsg.): Neue Medien im

Alltag: Begriffsbestimmungen eines interdisziplinären Forschungsfeldes. Opladen: Leske + Budrich, 153-177.

STORRER, A. (2001). „Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation." In: Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet. Hrsg. v. Andrea Lehr, Matthias Kammerer, Klaus-Peter Konerding, Angelika Storrer, Caja Thimm und Werner Wolski. Berlin: de Gruyter, 439-465.

STORRER, A. (2007). „Chat-Kommunikation in Beruf und Weiterbildung." In: Der Deutschunterricht, 49-61.

STORRER, A. (2009). „Rhetorisch-stilistische Eigenschaften der Sprache des Internets." In: Fix, U./Gardt, A./Knape, J. (Hrsg.): Rhetorik und Stilistik – Rhetorics and Stilistics. Ein internationales Handbuch historischer und systematischer Forschung. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 31/2), 2211-2226.

STORRER, A. (2012). „Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia". In: Köster, J./Feilke, H./Steinmetz, M. (Hrsg.): Textkompetenzen in der Sekundarstufe II. Freiburg: Fillibach, 277-304.

STORRER, A. (2013). „Sprachstil und Sprachvariation in sozialen Netzwerken". In: Frank-Job, B./Mehler, A./Sutter, T. (Hrsg.): Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW. Wiesbaden: VS Verlag für Sozialwissenschaften, 329-364.

STORRER, A. (im Druck). „Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde." In: Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013.

---

1 http://wiki.itmc.tu-dortmund.de/cmc/

2 http://www.tei-c.org/Activities/SIG/CMC/

3 Zu den kommunikativen Funktionen prosodischer Mittel in der gesprochenen Sprache vgl. z.B. Schwitalla (2006): 59-62.

4 „slml13" ist das Kürzel für eine Veranstaltung der Bundeszentrale für Politische Bildung mit dem Titel „SpeedLab: Mobiles Lernen – unabhängig von Raum und Zeit?", die am 26. April 2013 in Hannover stattfand.

5 Die Modellierung solcher „Linked-Data"-Phänomene steht u.a. auf der Agenda der Special Interest Group „Computer-Mediated Communication" im Rahmen der Text Encoding Initiative (http://www.tei-c.org/), die im Herbst 2013 ihre Arbeit aufgenommen hat.

6 Das Projekt D-SPIN („Deutsche Sprachressourcen-Infrastruktur") war bis 2011 der deutsche Beitrag zum europäischen CLARIN-Projekt („Common Language Resources and Technology Infrastructure"); seit 2011 werden die in D-SPIN begonnen Arbeiten im Nachfolgeprojekt CLARIN-D fortgeführt (vgl. http://www.d-spin.org/ und http://www.clarin-d.de/).

7 Erfasst würden unter dieser Kategorie dann lediglich die *umgangssprachlichen* Formen von Präposition-Artikel-Verschmelzungen (wie z.B. *aufm*, *innem*). Standardsprachlich etablierte Fälle des gleichen Bildungsmusters würden wie bisher unter *APPRART* erfasst.

Ines Rehbein and Sören Schalowski

STTS goes Kiez – Experiments on Annotating and Tagging Urban Youth Language

1 Introduction

The Stuttgart-Tübingen Tag Set (STTS) (Schiller et al., 1995) has long been established as a quasi-standard for part-of-speech (POS) tagging of German. It has been used, with minor modifications, for the annotation of three German newspaper treebanks, the NEGRA treebank (Skut et al., 1997), the TiGer treebank (Brants et al., 2002) and the TüBa-D/Z (Telljohann et al., 2004). One major drawback, however, is the lack of tags for the analysis of language phenomena from domains other than the newspaper domain. A case in point is spoken language, which displays a wide range of phenomena which do not (or only very rarely) occur in newspaper text.

The STTS, as a consequence, does not provide POS tags to capture these phenomena. As a result, other POS categories have been stretched to describe spoken language. For instance, in the Tübingen Treebank of Spoken German (TüBa-D/S) (Stegmann et al., 2000) the tag for interjections has been used to annotate filled pauses and backchannel signals like *uh, mhm*, adjectives like *richtig, gut, hervorragend* (right, good, excellent) when used in isolation, and for question tags. From a linguistic point of view, this practice is unsatisfactory and should be given up in favour of a more adequate description of spoken language phenomena.

In this paper, we present an extension of the STTS for the annotation of spoken language. We describe our new tagset and evaluate its adequacy in a manual annotation experiment. Furthermore, we develop a POS tagger for analysing spoken language data and evaluate its performance on spoken language transcripts as well as on a normalised version of the data.

This paper is structured as follows. In Section 2 we motivate the need for an extension of the STTS and define the new tags. Section 3 describes the data used in our experiments and presents the results of an annotation experiment. We report inter-annotator agreement on the extended tagset and make a proposal for restructuring and integrating the new tags into the STTS. In Section 4 we report on our efforts to develop a tagger for spoken language data, describing the tagging architecture and basic features used in our experiments. Section 5 focusses on adapting the tagger to spoken language, especially on addressing the out-of-vocabulary (OOV) problem of our data. We conclude and outline future work in Section 6.

JLCL 2013 – Band 28 (1) – 199-227

POS	TiGer	TüBa-D/Z	TüBa-D/S
PTKANT	7.5	26.2	279.7
ADV	27.1	71.6	46.7
ITJ	0.4	0.0	0.0
NN	1.8	2.4	0.0
KON	0.0	2.0	0.0
TOTAL	36.8	102.2	326.4

Table 1: Distribution of *ja* (yes) in different corpora, normalised by corpus size

2 Extensions to the STTS tag set

This section describes our extensions to the STTS for the annotation of spoken language. We first motivate the need for additional POS tags for analysing spoken language data. We review related work and argue that extending an existing annotation scheme is preferable to developing a new scheme tailored towards the specific needs of spoken language. Then we present our new tags and describe their usage.

2.1 (Why) do we need new tags?

A major pitfall for the annotation of spoken language is the danger of carrying over annotation guidelines from standard written text which, at first glance, seem to be adequate for the description of spoken language, too. Only at second glance does it become obvious that what looks similar at first may not necessarily be the same.

A case in point is *ja* (yes), which in written text mostly occurs as a modal particle in the middle field, while in spoken language occurrences of *ja* in utterance-initial position constitute the more frequent type. Table 1 shows the distribution of *ja* in two German newspaper copora, the TiGer treebank (Brants et al., 2002) and the TüBa-D/Z (Release 8) (Telljohann et al., 2004), and in a corpus of spoken dialogues, the TüBa-D/S (Stegmann et al., 2000). In TiGer and TüBa-D/Z, most instances of *ja* are in the middle field, annotated as ADV, while the utterance-initial instances in the TüBa-D/S are assigned the tag PTKANT (answer particle). Motivated by the difference in distribution, we take a closer look at these instances and observe that many of the utterance-initial cases are in fact discourse markers (Example 1).

(1) **ja** wer bist du denn ?
PTCL who are you then ?
And who are you now?

Other phenomena which cannot be analysed using the STTS inventory are filled pauses, question tags and backchannel signals. In the TüBa-D/S, filled pauses have been removed from the corpus, while question tags have been analysed as interjections (Example 2), as have backchannel signals (Example 3).

(2) es war doch Donnerstag , **ne** ?
is was however Thursday , no ?

It was Thursday, wasn't it?

(3) **mhm** ja das ist bei mir ganz offen
uh-huh yes this is for me totally open
Uh-huh, yes, I'm quite flexible.

We argue that these instances should not be analysed as interjections, as done in the TüBa-D/S, but should be assigned a new POS tag. In the next section, we report on related work on the annotation of word classes in spoken language corpora.

2.2 Related work

A number of spoken language corpora already exist, annotated with parts of speech. However, not much work has been done on developing or extending POS tagsets for the annotation of spoken language. Many corpora use POS tagsets originally developed for written language or make only minor changes to the tagset.

The Tübingen Treebank of Spoken German (TüBa-D/S), for instance, uses the STTS which had been developed for the annotation of written language. The spoken part of the BNC applies a tagset with around 60 tags but does not encode spoken language phenomena on the POS level. Hesitations in the BNC are not considered to be linguistic words and are thus annotated as *unclassified items*, as are non-English words, special typographical symbols and formulae. Discourse markers, on the other hand, such as backchannel signals and question tags, are subsumed under the interjection label.

The Switchboard corpus (Godfrey et al., 1992), a corpus of spontaneous conversations of English telephone bandwidth speech, follows the tagging guidelines of the Penn Treebank POS tagset, which was developed for annotating written (newspaper) text.[1] Switchboard only introduced minor changes to the original tagset. They added the BES and HVS tags to distinguish between *is* and *has* when being contracted and reduced to *'s*. Another extension is the XX tag used for marking partial words where the full word form can not be recovered from the context. Similarly to the BNC, different discourse markers are treated as interjections.

One example for a POS tagset specifically designed for annotating spoken language is the one developed for the Spoken Dutch Corpus (Oostdijk, 2000). The hierarchical tagset distinguishes 10 major word classes, while the whole tagset provides more than 300 fine-grained morpho-syntactic tags (Eynde et al., 2000). Despite its detail, the tagset does not encode differences between different markers of discourse but, similar to the BNC, analyses these items as interjections.

Two noteworthy exceptions to the simple re-use of schemes developed for written language are the Vienna-Oxford International Corpus of English (VOICE) (Breiteneder et al., 2006), a corpus of English as a lingua franca, and the the Christine corpus (Sampson, 2000), which is one of the first treebanks of spoken language data.

[1] The Switchboard corpus does, however, provide a fine-grained annotation of disfluencies on the syntactic level, covering phenomena such as non-sentential units and restarts.

VOICE adapts the Penn Treebank POS tagset by adding 26 new categories to the tagset. Some of them describe properties of spoken discourse (e.g. discourse markers, response particles, and formulaic items like greetings), while others add non-verbal information (e.g. breathing or laughter). Other additional tags distinguish between contracted verb forms, similar to Switchboard.

The Christine corpus uses a much more fine-grained POS tagset with more than 400 morpho-syntactic tags tailored to the analysis of spoken language. The POS tags in the Christine corpus allow one to annotate discourse phenomena such as filled pauses, backchannel signals and question tags, to distinguish between different types of swearwords, to annotate formulaic expressions like greetings, or to encode onomatopoeia and forms of echoism. The tagset also distinguishes between different types of pragmatic units, such as apologies, responsives, and positive and negative answers.

In the next section, we present our own work on extending the STTS for the annotation of spoken language.

2.3 Extensions to the STTS for spoken language annotation

Our approach to extending the STTS is purely *data-driven*. We started annotating the data using the original STTS tagset, and only when encountering phenomena which could not be described within the STTS tagset, we introduced new tags. We tested these provisional tags on new data and refined our classification, merging classes which proved to be difficult for the human annotators. As a result, we ended up with 11 additional tags for the annotation of spoken language phenomena (Table 2).

POS	**description**	**example**	literal translation
PAUSE	*pause, silent*	so ein **(-)** Rapper	such a (-) rapper
PTKFILL	*particle, filler*	ich **äh** ich komme auch .	I er I come too
PTKINI	*particle in utterance-initial position*	**ja** kommst du denn auch ?	PART come you then too
PTKRZ	*backchannel signal*	A: ich komme auch . B: **hm-hm** .	A: I come too B: uh-huh
PTKQU	*question particle*	du kommst auch . **Ne** ?	you come too . no ?
PTKONO	*onomatopoeia*	das Lied ging so **lalala** .	the song went so lalala
PTKPH	*placeholder particle*	er hat **dings** hier .	he has thingy here
VVNI	*uninflected verb*	seufz	sigh
XYB	*unfinished word, interruption*	ich **ko** #	I co #
XYU	*uninterpretable*	**(unverständlich)** #	(uninterpretable) #
$#	*unfinished utterance*	ich ko **#**	I co #

Table 2: Additional POS tags for spoken language data

2.3.1 Hesitations

The first two tags encode silent pauses and filled pauses. The PAUSE tag is used for silent (unfilled) pauses which can occur at any position in the utterance.

(4) das ist irgend so ein **(-)** Rapper
that is some such a rapper
That is some rapper.

The PTKFILL tag is used for filled pauses which can occur at any position in the utterance.

(5) das ist irgend so ein **äh** Rapper
that is some such a uh rapper
That is some uh rapper.

2.3.2 Discourse particles

The following tags are used for the annotation of discourse particles. We use the term *discourse particles* in a theory-neutral sense as an umbrella term for a variety of particles and discourse markers frequently used in spoken language.

The PTKINI tag is assigned to particles such as *ja* (yes), *na* (there, well) when used as a discourse marker in an utterance-initial position. In contrast to interjections, these particles do not carry stress. They have been described in the literature as *Eröffnungssignale* (opening signals) (Schwitalla, 1976) or *Gliederungssignale* (discourse structuring signals) in utterance-initial position (Schwitalla, 2006), or as discourse markers in the pre-prefield (Auer and Günthner, 2005).

(6) **ja** wer bist du denn ?
PTCL who are you then ?
And who are you now?

Please note that most occurrences of *ja* (yes) in the middle field are modal particles (Example 7) which are assigned the ADV label (adverb) in the German treebanks. Occurrences of *ja* in utterance-initial position, on the other hand, are discourse markers and thus should be treated differently (also see Meer (2009) for a discussion on the different word classes of *ja*).

(7) die hat **ja** auch nicht funktioniert .
that one has PTCL also not worked .
That one didn't work, either.

The PTKRZ tag is used for backchannel signals. We define backchannel signals as plain, non-emotional reactions of the recipient to signal the speaker that the utterance has been received and understood.

(8) A: stell dir das mal vor !
A: imagine you this PTCL VERB PTCL !
Imagine that !

(9) B: **m-hm**
B: uh-huh

Preliminary annotation experiments showed a very low inter-annotator agreement for the distinction between answer particles and backchannel signals for *ja*. There is, in fact, an overlap of meaning which makes a clear distinction between the function of *ja* as an answer particle and a backchannel signal infeasible. To support consistency of annotation, we always label *ja* as answer particle and not as backchannel signal.

The PTKQU tag is used for question tags such as *ne* (no), *gell* (right) or *wa* (what) added to the end of a positive or negative statement. Please note that we do not annotate adjectives like *okay, richtig* (okay, right), interrogative pronouns like *was* (what) or conjunctions like *oder* (or) as PTKQU, as both classes show distributional differences. Instances of *okay, richtig, was, und, oder* in the context of a question are still annotated as adjectives, interrogative pronouns or conjunctions, respectively.

(10) wir treffen uns am Kino , **ne** ?
we meet REFL at the cinema , no ?
We'll meet at the cinema, right ?

Du kommst auch , **wa** ?
You come too , what ?
You'll come too, right?

2.3.3 Other particles

The PTKONO tag is used for labelling onomatopoeia and forms of echoism.

(11) das Lied ging so **lalalala**
The song went like lalalala

(12) **eieieieia** !

(13) **bam , bam , bam** !

(14) interessant , **bla bla** .
interesting , bla bla .

The PTKPH tag is used as a placeholder when the correct word class cannot be inferred from the context. Example (15), for instance, has a number of possible readings. In (a), the correct POS tag would be noun (NN), while in (b) we would assign a past participle (VVPP) tag. The placeholder might also stand for a whole VP, as in (c).

(15) er hat **dings** hier .
he has thingy here .

a. er hat MP3-Player$_{nn}$ hier .
he has MP3 player here .

b. er hat gewonnen$_{vvpp}$ hier .
he has won here .

c. er hat (Schuhe gekauft)$_{vp}$ hier .
he has shoes bought here .

2.3.4 Uninflected verb forms

We use the tag VVNI to annotate non-inflected verb forms (Teuber, 1998). Non-inflected auxiliaries (VANI) and modal verbs (VMNI) are also possible forms but very rarely occur in spoken language. They do, however, occur in computer-mediated communication (CMC).

(16) ich muss noch putzen . **seufz** !
I must still clean . sigh !
I still have to clean. Sigh!

(17) gleich haben wir Mathe . **gähn** !
soon have we math . yawn !
We have math right now. Yawn!

2.3.5 Non-words

The STTS provides the XY tag for the annotation of non-words. We add two new subclasses to distinguish between different types of non-words.

1. uninterpretable material (XYU)
2. unfinished words (XYB)
3. other (XY)

The XYU tag is used for lexical material which is uninterpretable, mostly because of poor audio quality of the speech recordings or because of code-switching.[2]

(18) wir waren gestern bei **(fremdsprachlich)**.
we were yesterday at (FOREIGN).
Yesterday we've been at (FOREIGN).

The XYB tag is used for abandoned words.

[2]For foreign language material in the data which can be understood and transcribed we use the FM tag provided by the STTS.

(19) ich habe **gest** # heute komme ich nicht .
I have yest # today come I not .
I have yest- # I won't come today.

The XY tag is used for all non-words which do not fit one of the categories above. This category is more or less consistent with the XY category in the STTS where it is used for non-words including special symbols.

2.3.6 Punctuation

The **$#** tag is a new punctuation tag used to mark interrupted or abandoned utterances. These can (but do not necessarily) include unfinished words, as in Example (20).

(20) sie war gest #
she was yest #

2.3.7 Extensions to the STTS – Conclusion

The corpora presented in Section 2.2 made different decisions regarding the question what kind of information should be encoded on the POS level. Some of them try to restrict the tagset to word classes which can be defined purely on a grammatical level (TüBa-D/S, BNC, Switchboard, Spoken Dutch Corpus), others choose to also include rich pragmatic information (VOICE, Christine). While it is hard to stick to a purely grammatical distinction – the STTS, for instance, uses a mix of grammatical, distributional and semantic criteria for defining different word classes – the latter approach is not uncontroversial, either. Pragmatic categories are often vague and ill-defined, thus compromising the consistency of the annotations. It can also be argued that they provide a very different type of information which should not be encoded on the word class level.

While that point is well taken, we would still like to include pragmatic information, which is highly relevant for the analysis of discourse, in the corpus. We consider the annotation layers of the corpus not as the final product but as a database which allows us to generate different views on the data (which would correspond to different corpus versions of the same data, one subsuming all discourse particles under the *interjection* label, another one also including the pragmatic tags on the same level or projecting those to a new annotation layer). Our reasons for encoding pragmatic information on the POS level are mostly practical ones. This way of proceeding allows for swift annotation without the need for a second pass over the data, it results in a more compressed, thus more clearly arranged presentation of the data (whereas adding yet another corpus layer would give us a more confusing view), and, finally, it also facilitates corpus queries.

3 Annotation experiments

This section reports on an annotation experiment with human annotators using the extended tagset. We describe the data we used in the experiments and report numbers for inter-annotator agreement (IAA) between the annotators. Based on a detailed error analysis for the new POS tags we present our proposal for integrating the new tags into the STTS.

3.1 Data: KiDKo – The Kiezdeutsch-Korpus

The data we use in our experiments is taken from the Kiezdeutsch-Korpus (KiDKo) (Wiese et al., 2012). Kiezdeutsch (*'hood German*) is a variety of German used in informal peer-group communication, spoken by adolescents from multilingual urban neighbourhoods.

The data was collected in the first phase of project B6 "Grammatical reduction and information structural preferences in a contact variety of German: Kiezdeutsch" as part of the SFB (Collaborative Research Centre) 632 "Information Structure" in Potsdam. It contains spontaneous peer-group dialogues of adolescents from multiethnic Berlin-Kreuzberg (around 48 hours of recordings) and a supplementary corpus with adolescent speakers from monoethnic Berlin-Hellersdorf (around 18 hours of recordings). The current version of the corpus includes the audio signals aligned with transcriptions. The data was transcribed using an adapted version of the transcription inventory GAT basic (Selting et al., 1998), often referred to as minimal GAT, including information on primary accent and pauses. Additional annotation layers (POS, Chunking, Topological Fields) are work in progress.[3]

The transcription scheme has an orthographic basis but, in order to enable investigations of prosodic characteristics of the data, it also tries to closely capture the pronunciation, including pauses, and encodes disfluencies and primary accents. In addition, we are adding a level of orthographic normalisation where non-canonical pronunciations and capitalisation are reduced to standard German spelling. This annotation layer enables us to use standard NLP tools for semi-automatic annotation.[4] It also increases the usability of the corpus as it allows one to find all pronounciation variants of a particular lexeme. The normalisation is done in a semi-automatic way. We copy the text from the transcription layer, automatically correcting frequent deviations from the orthographic norm based on dictionaries and frequency lists. The remaining changes are carried out manually during the transcription process.

Figure 1 shows an example transcript from the KiDKo, displaying the transcription and the normalisation layer, the POS tags and the layers for non-verbal information. Uppercase letters on the transcription layer mark the main accent of the utterance. The equals sign is used to encode the tight continuation of a word form with a following

[3] The first release of KiDKo is scheduled for spring 2014 and will include the transcribed data as well as the normalisation and POS annotation layers.

[4] Please note that we still have to adapt these tools to our data. However, without the normalisation the manual effort to correct these tags would be much higher.

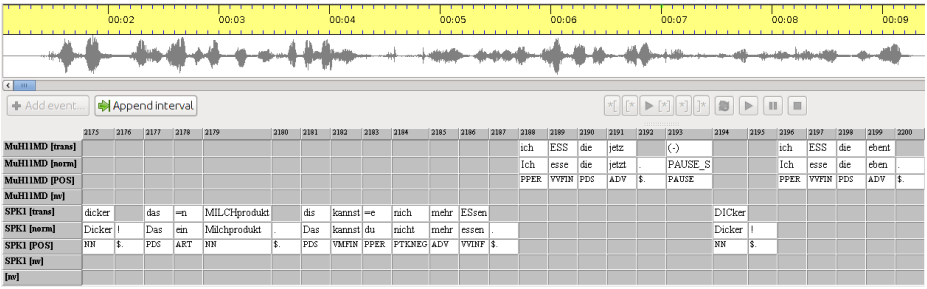


Figure 1: Screenshot KiDKo sample of a short dialogue between two speakers (MuH11MD, SPK1) in EXMARaLDA (transcription, normalisation, POS and non-verbal layer)

form, where one of the two forms (or both of them) is reduced (e.g. das =n "this a", kannst =e "can you").

We would like to emphasise that linguistic annotations not only provide a description but also an interpretation of the data. This is especially true for the annotation of learner data, where the formulation of target hypotheses has been discussed as a way to deal with the ambiguity inherent to a learner's utterances (Hirschmann et al., 2007; Reznicek et al., 2010). When annotating informal spoken language, we encounter similar problems. Adding an orthographic normalisation to the transcription might be seen as a 'poor man's target hypothesis' where decisions made during the annotation become more transparent.

3.2 Inter-annotator agreement

Inter-annotator agreement for human coders on the core STTS is quite high. For instance, Rehbein et al. (2012) report a percentage agreement of 97.9% and a Fleiss' κ of 0.978 for two human annotators on the target hypotheses of essays written by advanced second language learners of German.

In a preliminary annonation experiment with three human annotators, we obtained a percentage agreement of 96.5% and a κ of 0.975 on a small test set (3,415 tokens) from the KiDKo, using our extended tagset. This shows that the extended tagset does not result in a decrease in accuracy for the manual annotation.

However, due to the small size of the test set, some of the new tags only occurred infrequently in the sample. To provide a more meaningful evaluation for the new tags, we created a new test set, focussing only on the discourse particles *answer particles, backchannel signals, question tags, fillers, utterance-initial particles* (PTKANT, PTKRZ, PTKQU, PTKFILL, PTKINI) and on *onomatopoeia* and *placeholders* (PTKONO,

POS	freq.	# agr.	% agr.
PTKANT	903	809	89.59
PTKQU	296	255	86.15
PTKFILL	126	112	88.89
PTKINI	121	116	95.87
PTKONO	61	55	90.16
PTKRZ	33	15	45.45
PTKPH	13	8	61.54
avg.	1553	1370	88.22

Table 3: IAA for two human annotators on the discourse particles

PTKPH).[5] We took a subpart of the corpus with 39,583 tokens which had already been annotated with POS by one annotator. A second annotator then assigned POS tags to all instances which she considered to be one of the discourse particles listed above. Candidate instances for the second annotation were identified using heuristics based on automatically assigned tags by two versions of the TreeTagger, one using the tagging model trained on the TiGer corpus and the second one using a tagging model adapted to the new tags.[6] The accuracy of the two tagging models on the KiDKo data is not very high. The two taggers do, however, produce systematic errors on the new domain which allows us to detect instances of discourse particles without having to look at too many tokens. In the evaluation, we only include those instances which had been assigned one of the discourse particle tags by at least one of the annotators. Table 3 shows detailed results for these tags.

For all particle tags we observe an agreement which is below the average agreement on the whole tagset. This is not surprising, as these tags do encode pragmatic information and are thus much harder to define and operationalise than most of the other POS tags.

For some categories, we obtained an acceptable agreement of close to or over 90% (PTKINI, PTKANT, PTKFILL, PTKONO). For the question tags (PTKQU), many disagreements were caused by one annotator assigning the PTKQU tag to conjunctions like *oder, und* (or, and) when used in the context of a question, while the second annotator assigned the KON tag to these instances (Example 21), as intended by the guidelines. This problem can easily be solved by revising the guidelines and making explicit which tokens should be interpreted as a question tag and which should not.

(21) du musst au wir müssen AUFhören **oder** (transcript)
Du musst au wir müssen aufhören . Oder ? (normalisation)
you have to sto we have to stop . or (literal translation)

[5] Silent pauses and $# have not been included in the testset because they are not ambiguous. The XYB/XYU tags do not encode linguistically interpretable categories but should be considered as a technical device to deal with material which, partly caused by low audio quality and partly caused by phenomena of social interaction, otherwise could not be analysed.

[6] The adapted model was also trained on the TiGer corpus but uses an extended dictionary which includes word-POS pairs for the most frequent tokens in the KiDKo data.

You have to sto- we have to stop, right? (free translation)

Only low agreement could be achieved on the placeholder particles. Here the annotators disagreed on whether instances of the following kind are ambiguous between different POS tags or not (Example 22). According to the guidelines, Example 22 should be analysed as a placeholder because the placeholder slot could be filled with a noun (Example 22 a) or a verb (Example 22 b). Unfortunately, the annotators are not always aware of these ambiguities but tend to settle on the most probable interpretation of the utterance, thus overlooking other possible readings.

(22) wenn ich das hier äh (-) **DINGS** (-) äh a (-) wenn ich das ANschalte
Wenn ich das hier äh **dings** äh a Wenn ich das anschalte
when I this here uh thingy uh a when I this turn on

a. Wenn ich das hier äh **Schalter** äh a (-)
when I this here uh button uh a

b. Wenn ich das hier äh **anschalte** äh a (-)
when I this here uh turn on uh a

The hardest task for the human annotators was the distinction between answer particles, backchannel signals and fillers. For illustration, consider Example 23, where one annotator interpretated the first particle, *hm*, as a filler (PTKFILL) and the second one, *'hmhm*,[7] as an answer particle, while the second annotator analysed both particles as backchannel signals (PTKRZ). In Example 24, on the other hand, annotator one interpreted the token as a backchannel signal while annotator two annotated it as an answer particle (PTKANT).

(23) A: schmeckt ja dann bestimmt SCHEIße (-) **hm** **'hmhm**
A: Schmeckt ja dann bestimmt scheiße . **Hm** . **Hm-hm** .
A: tastes PTC then surely shit hm m-hm
A: This surely will taste like shit. Hm. M-hm.

(24) B: als wenn man da DROgn vertickt aufm schulhof (-) A: **hm**
B: Als wenn man da Drogen vertickt aufm Schulhof . A: **Mh**
B: as if one there drugs sells at the schoolyard A: mh
B: As if one would sell drugs on the schoolyard. A: Mh .

It is not clear whether these distinctions can be operationalised sufficiently to enable a reliable annotation. One might ask whether it is advisable to encode pragmatic differences like those in a part-of-speech tagset or whether these fine-grained annotations should be transferred to a separate layer of annotation, and should be subsumed under the answer particles on the part-of-speech level.

[7] The apostrophe indicates a glottal stop.

coarse	fine	
	DMANT	answer particles (PTKANT)
	DMITJ	interjections (ITJ)
DM	DMQU	question particles
discourse markers	DMRZ	backchannel signals
	DMFILL	filler
	DMINI	utterance-initial discourse particle
PTKONO		onomatopoeia
PTKPH		placeholder
VANI/VMNI/VVNI		uninflected verbs

Table 4: Possible integration of the new tags in the STTS

As a compromise, we propose the following classification of the new tags, shown in Table 4. A coarse-grained POS tag for discourse markers could ensure a reliable, consistent annotation, while a more fine-grained classification can be used when a more detailed analysis is wanted. The DM tag would now comprise the former STTS tags for answer particles (PTKANT) and interjections (ITJ) as well as question particles, backchannel signals, fillers and utterance-initial discourse particles. In addition to the STTS tags for separable verb particles (PTKVZ), the particle *zu* with an infinite verb form (PTKZU) and a particle with an adjective or adverb (PTKA), we now have the placeholder particle (PTKPH) and the particle for onomatopoeia and forms of echoism (PTKONO).[8] The non-inflected verb forms (VANI/VMNI/VVNI) are part of the STTS verb paradigm, as indicated by the prefix VA/VM/VV.

4 Developing a POS tagger for Kiezdeutsch

While automatic POS tagging of canonical, written text from the newspaper domain might appear to be a solved problem with accuracies in the high nineties (Schmid, 1994, 1995; Schmid and Laws, 2008), a very different picture emerges when looking at text from other domains. Applying a tagger trained on newspaper text to spoken language data or to user-generated content from the web will result in a substantial decrease in accuracy. The use of informal language, creative inventions of new words and a high number of variants for existing word forms in combination with a non-canonical syntax result in data sparseness and causes problems for automatic processing. For spoken language, disfluencies such as hesitations, filled pauses, repeated material or repairs further add to the problem.

This section describes our efforts to develop a POS tagger for spontaneous multiparty dialogues of Kiezdeutsch. The data used in our experiments includes 18 different transcripts, where each transcript has between two and seven speakers. Table 14 (Appendix) shows the distribution of POS tags in the manually annotated data from the

[8] It is open to discussion whether onomatopoeia and placeholders should be integrated as particles or as subordinate XY elements.

corpus	tagging ambiguity
KiDKo (normalised)	1.10
KiDKo (transcripts)	1.11
TiGer	1.03

Table 5: Tagging ambiguity for KiDKo (normalised and transcribed layer) and for an equally-sized portion of the TiGer corpus

KiDKo (training/development/test sets, normalised; 28,827 tokens) and, for comparison, in a test set of the same size from the TiGer treebank. As expected, we can observe crucial differences between the data sets. Average sentence length in the newspaper corpus is much longer than the average length of utterances in KiDKo, as indicated by the higher number of sentence delimiters (\$., \$#). The exact numbers, however, should be interpreted with care, as the question of how to segment spoken language is by no means clear.[9]

Our guidelines for segmentation are motivated by our goal of maintaining interoperability and comparability with corpora of written language on sentential utterances in the data. We follow the terminology of Fernandez and Ginzburg (2002) and define *sentential utterances* as utterances containing a finite verb, while we call utterances without a fininte verb *non-sentential utterances*. We thus base our unit of analysis on structural properties of the utterance and, if not sufficient, also include functional aspects of the utterance as criteria for segmentation. The latter results in what we call the principle of the *smallest possible unit*, meaning that when in doubt whether to merge lexical material into one or more units, we consider the speech act type and discourse function of the segments. Example (25) illustrates this by showing an utterance including an imperative (*Speak German!*) and a check question (*Okay?*). It would be perfectly possible to include both in the same unit, separated by a comma. However, as both reflect different speech acts, we choose to annotate them as separate units of analysis.

(25) Rede auf Deutsch ! Okay ?
Speak on German ! Okay ?
Speak German! Okay?

Other striking differences between KiDKo and TiGer include the higher number of attributive adjectives, adpositions, articles, nouns and proper names in TiGer, while in KiDKo we observe a higher frequency of adverbs, personal and demonstrative pronouns, finite auxiliaries and imperatives and, of course, of answer particles.

The tagging ambiguity (number of tags per token) for the KiDKo is 1.10 for the normalised transcripts and 1.11 for the original transcripts (Table 5). For comparison, the tagging ambiguity for an equally-sized portion from the TiGer treebank is 1.03.

[9] See, e.g., Crookes (1990); Foster et al. (2000) for a discussion on the adequate unit of analysis for investigations of spoken language.

We divided the manually annotated data into a training set, a development set and a test set. The split of the data was done as follows. First we created 10 bins. Then we processed the data utterance-wise and put the first utterance in bin1, the second utterance in bin2, and so forth. As a result, we ended up with three bins holding 475 utterances each and seven bins with 474 utterances each. From this, we took the first 1,500 of the 4,743 utterances for the development set (9,210 tokens) and the next 1,500 utterances (8,935 tokens) for the test set. The remaining 1,743 utterances (10,682 tokens) were used as training data.

The transcribed version of the data has fewer tokens than the normalised version because the transcripts do not include punctuation. For the transcripts, the development set includes 7,059 tokens, the test set 6,827 tokens and the training set 8,231 tokens. Unless stated otherwise, all results reported throughout the paper are on the development set.

Before developing our own tagger, we want to establish a baseline by testing how well a state-of-the-art tagger for German, trained on newspaper text, performs on the spoken language data.

4.1 Baseline

For our baseline we use the TreeTagger (Schmid, 1994, 1995), a probabilistic POS tagger using Markov models and decision trees. We use the tagger and the German tagging model out-of-the-box with no adaptations and apply it to the spoken language data.

Preparing the input for the tagger is not straightforward, as we have multi-party dialogues with overlapping speech events. In this work we pay no attention to the temporal ordering of the utterances but use randomised sets as training/development/test data (see Section 4 above). Proceeding like this means that we lose important context information, while a more sophisticated way of presenting the data to the tagger might improve results. We will explore this in future work.

	accuracy			
	transcript		*normalised*	
	dev	*test*	*dev*	*test*
baseline 1: original TreeTagger				
extended tagset	42.54	42.48	74.53	73.67
core STTS tags only	51.48	51.71	86.03	85.29
baseline 2: re-trained TreeTagger				
extended tagset	58.42	59.90	91.76	91.56
core STTS tags only	53.49	55.28	92.22	92.08

Table 6: Baseline results for the TreeTagger on spoken language data for the original transcripts and for the normalised version of the transcripts

Table 6 shows the baseline results for the TreeTagger on the transcribed dialogues as well as on a normalised version of the development and test set. In the first row results are given for all tags in the extended tagset. Please note that this setting is rather unfair as many of the POS tags are not included in the original STTS tag set and are thus unknown to the tagger. The second row presents results on those POS tags which are included in the STTS tag set. Results show the importance of the manual normalisation of the transcripts. For the extended tagset as well as for the core STTS tags, we achieve an accuracy of more than 30% higher than on the original transcripts.

The second baseline shows results for the TreeTagger which was re-trained on the KiDKo training set. Results show that even a small amount of manually annotated data is enough to obtain a substantial increase in accuracy both for the transcripts as well as for the normalised version of the data. The better results on the extended tagset as compared to the STTS-only setting can be explained by the PAUSE tag, which is unambiguous and occurs with a high frequency in the data.

Please note that the accuracy on the original STTS tags even on the normalised transcripts is still substantially lower than the one obtained on in-domain data from German newspaper text where we can expect results in the range of 96 to 98%.

4.2 Base tagger and feature tuning

We develop our tagger based on Conditional Random Fields (CRF) (Lafferty et al., 2001), using the CRFsuite package[10] (Okazaki, 2007) which provides a fast implementation of CRF for sequence labelling.

This section describes the features used in our experiments. We train the tagger on a small feature set, using features like word form, word length, or the number of digits in a word (see Table 7 for the whole feature set). In addition, we use prefix/suffix features (the first/last n characters of the input word form) as well as feature templates which generate new features of word ngrams where the input word form is combined with preceding and following word forms. Example 26 illustrates how these templates work. For instance, for the third token in (26), *irgend*, our templates extract the features in Table 8.

(26) Das ist irgend so ein äh Rapper . (normalisation)
this is some such a uh rapper .
This is some uh rapper.

Table 9 (01a) presents results for the different settings for prefix/suffix size, starting from 4 up to 10. Results show a slight increase with larger prefix/suffix sizes. The differences between prefix/suffix sizes of 5 to 10, however, are not statistically significant.

As the transcripts include uppercase letters for marking the main accent of an utterance, we run a further experiment where we transform all word forms in the

[10] We run all our experiments with the default parameter setting (1st-order Markov CRF with dyad features, training method: limited memory BFGS).

feature	description
wrd	word form
len	word length
cap	is the word form capitalised? $\{0, 1\}$
anonym	number of X in word form
upper	number of upper case in wrd
digit	number of digits in wrd
	prefix/suffix features
pre N	prefix: first N characters of word form (from 1 to N)
suf N	suffix: last N characters of word form (from 1 to N)
	ngram features
ngrams	different ngram combinations
2grams	adjacent word forms: $w_{-2}w_{-1},\ w_{-1}w_0,\ w_0w_1,\ w_1w_2$, context of w_0: $w_0w_{-1},$ ..., $w_0w_{-9},\ w_0w_1,$ ..., w_0w_9
3grams	$w_{-2}w_{-1}w_0,\ w_{-1}w_0w_1,\ w_0w_1w_2$
4grams	$w_{-2}w_{-1}w_0w_1,\ w_{-1}w_0w_1w_2$
5grams	$w_{-2}w_{-1}w_0w_1w_2$

Table 7: Feature set used in our experiments

data to lowercase letters. We expect that results for the transcripts improve while the accuracy on the normalised data should go down.

Table 9 (01b) shows results for the lowercase setting. We observe a significant improvement for the transcribed version of the data (two-sided McNemar test, $p < 0.0001$). To our surprise, the accuracy on the normalised transcripts also shows a slight increase, which again is statistically significant. As the difference between the prefix/suffix sizes of 5 to 10 is not statistically significant, we decided to run all further experiments with a size of 7.

5 Tagger adaptation

This section presents two methods for domain adaptation, both addressing the out-of-vocabulary (OOV) problem in the data. The first approach uses Latent Dirichlet Allocation (LDA) word clusters learned on unannotated data from the social media, namely from Twitter[11] microtext. The second approach relies on knowledge learned from a huge corpus of out-of-domain data, automatically annotated with POS tags. The first approach is motivated by the assumption that computer-mediated communication (CMC) shares some of the properties of face-to-face communication (see, e.g., Biber and Conrad (2009), Chapter 7 for a discussion of similarities and differences of CMC and face-to-face conversation). The second approach uses data from the same domain as the training data, but aims at improving tagging performance on the target domain by reducing the number of unknown words in the data.

[11] `https://de.twitter.com`

feature	value
w_{-2}	das
w_{-1}	ist
w_0	irgend
w_1	so
w_2	ein
$w_{0,-2}$	irgend|Das
$w_{0,-1}$	irgend|ist
$w_{0,1}$	irgend|so
$w_{0,2}$	irgend|ein
$w_{0,3}$	irgend|äh
$w_{0,4}$	irgend|Rapper
$w_{0,5}$	irgend|.
$w_{-2,-1,0}$	Das|ist|irgend
$w_{-1,0,1}$	ist|irgend|so
$w_{-2,-1}$	Das|ist
$w_{-1,0}$	ist|irgend
$w_{0,1}$	irgend|so
$w_{1,2}$	so|ein
$w_{0,1,2}$	irgend|so|ein
$w_{-2,-1,0,1}$	Das|ist|irgend|so
$w_{-1,0,1,2}$	ist|irgend|so|ein
$w_{-2,-1,0,1,2}$	Das|ist|irgend|so|ein

Table 8: Additional features extracted by the templates for *irgend*, Example 26

5.1 Tagger adaptation with LDA word clusters

Word clustering has been used for unsupervised and semi-supervised POS tagging, with considerable success (Biemann, 2006; Søgaard, 2010; Chrupała, 2011; Owoputi et al., 2012). For tagging English Twitter data, Owoputi et al. (2012) apply Brown clustering, a hierarchical word clustering algorithm, to the unlabelled tweets. During clustering, each word is assigned a binary tree path. Prefixes of these tree paths are then used as new features for the tagger.

Chrupała (2011) proposes an alternative to Brown clustering, using LDA. LDA has two important advantages over Brown clustering. First, the LDA clustering approach is much more efficient in terms of training time. Second, LDA clustering produces soft, probabilistic word classes instead of the hard classes generated by the Brown algorithm, thus allowing one word to belong to more than one cluster. Chrupała (2011, 2012) shows that the LDA approach outperforms Brown clustering on many NLP tasks. We

EXP	features	trans.	norm.
01a	pre/suf 4	83.89	93.45
	pre/suf 5	84.29	93.59
	pre/suf 6	84.51	93.59
	pre/suf 7	84.63	93.60
	pre/suf 8	84.68	93.67
	pre/suf 9	84.75	**93.71**
	pre/suf 10	**84.78**	**93.71**
01b	pre/suf 4 lc	86.31	93.52
	pre/suf 5 lc	86.67	93.65
	pre/suf 6 lc	87.19	93.87
	pre/suf 7 lc	87.23	**93.88**
	pre/suf 8 lc	**87.26**	93.87
	pre/suf 9 lc	87.20	93.87
	pre/suf 10 lc	87.25	93.80

Table 9: Results for different feature settings: varying prefix/suffix sizes (01a,b), lowercase word forms (01b)

thus apply the LDA clustering approach using the software of Chrupała (2011)[12] to an unlabelled corpus of Twitter data.

We decided to use Twitter data because it is freely accessible in a digitised format, it provides highly informal communication and includes many phenomena of spoken language, like fillers and repairs (27a), interjections and non-canonical word order (27b) as well as verb-less utterances (27c). While coming from a medially written register, Twitter data can in many respects be considered conceptually oral.[13]

(27) a. Find ich nicht **gut** ... **äh schlimm** !
find I not good ... uh bad !
I don't think it's good ... uh bad!

b. @BrandyShaloo **ah** OK **weil** **ich bin** noch nicht soweit ;)
@BrandyShaloo ah ok because I am still not ready ;)
@BrandyShaloo ah ok, because I'm not ready yet ;)

c. Nächste Woche dann ich wieder mit voller Dröhnung .
next week then I again with full thundering .
Next week it's me again doing the full monty

[12] https://bitbucket.org/gchrupala/lda-wordclass/

[13] For a model of medial and conceptual orality and literacy see Koch and Oesterreicher (1985), and Dürscheid (2003) for an extension of the model to new forms of communication such as email, text messages or chat.

frequency	word form	POS
410873	einen	ART
16550	einen	PIS
8679	einen	ADJA
438	einen	NN
160	einen	VVFIN
144	einen	VVINF

Table 10: Entries for *einen* in the automatically created HGC dictionary (ART: determiner; PIS: indefinite pronoun, substitutive; ADJA: adjective, attributive; NN: noun; VVFIN verb, finite; VVINF: verb, infinite)

The Twitter data was collected from Twitter over a time period from July 2012 to February 2013. We used the Python Tweepy module[14] as an interface to the Twitter Search API[15] where we set the API language parameter to German and used the geocode parameter to define positions in 48 different urban and rural regions all over Germany from where we harvested our data.[16] We ended up with a corpus of 12,782,097 tweets with a unique id.

Before clustering, we normalise the data. Instead of word forms, we use lemmas automatically generated by the TreeTagger (for unknown lemmas, we fall back to the word form). Our corpus contains 204,036,829 tokens and is much smaller than the one used in Owoputi et al. (2012) which includes 847,000,000 tokens of Twitter microtext.

We test different settings for LDA, setting the threshold for the minimum number of occurrences for each word to be clustered to 8, 10, 12 and 20, and induce clusters with 50 and 100 classes. Results for 50 and 100 classes are in the same range but slightly higher for 50 classes. We thus keep this setting and only report results for inducing 50 word classes.

5.2 Tagger adaptation with an automatically extracted dictionary

In our second approach to domain adaptation we stack the tagger with features extracted from an automatically created dictionary. The dictionary was harvested from the Huge German Corpus (HGC) (Fitschen, 2004), a collection of newspaper corpora from the 1990s with more than 204 billion tokens. We automatically POS tagged the HGC using the TreeTagger (Schmid, 1995). For each word form, we add the five most frequently assigned POS tags as new features. To reduce noise, we only included word POS pairs with a POS which had been predicted at least 10 times for this particular word form. As an example, consider the word form *einen* (Table 10). For *einen*, our automatically

[14] `http://pythonhosted.org/tweepy/html`

[15] `https://dev.twitter.com/docs/api/1/get/search`

[16] The reader should be aware that these parameters are not flawless and should be considered as an approximation rather than the plain truth.

method	transcription	normalised
Base tagger (lowercase)	87.23	93.88
domain adaptation 1: LDA		
LDA 50-6	88.66	95.02
LDA 50-8	88.76	94.94
LDA 50-10	88.95	**95.02**
LDA 50-12	**88.96**	95.01
LDA 50-20	88.77	94.93
domain adaptation 2: HGC		
HGC	**90.66**	**95.86**

Table 11: Results for LDA word clusters (trans.: lc, without `cap, upper` ; norm.: lc, but with features `cap, upper`) and for the HGC dictionary on the development set

created dictionary lists the entries in Table 10. We use the first 5 tags, ART, PIS, ADJA, NN and VVFIN, as additional features for the tagger.

5.3 Tagger adaptation – results

Table 11 presents results for the different domain adaptation methods. For convenience, we also repeat the results from Section 4.2 for our base tagger (prefix/suffix size 7).

The results for the different thresholds for the LDA word clustering approach are very close. We obtain an improvement over the base tagger of close to 2% on the transcripts and around 1% on the normalised data. Using the dictionary features from the HGC, we achieve an increase in acccuracy of more than 3% on the transcripts and of nearly 2% on the normalised version of the data.

5.4 Error analysis

To find out whether the two different methods address the same problem, we analyse the most frequent errors for each setting.

Figure 2 shows the impact of the different settings on the predictions for those tags where our base tagger made the most mistakes on the transcripts (nouns: NN, adjectives: ADJD/ADJA, adverbs: ADV, finite/infinite/imperative verbs: VVFIN/VVINF/VVIMP, indefinite substitutive pronouns: PIS, foreign material: FM, verb particles: PTKVZ, demonstrative pronouns: PDS, interjections: ITJ, unfinished words: XYB, proper names: NE, determiners: ART). We see a substantial improvement for the LDA clustering approach on most POS tags. For interjections and proper names, however, the clusters do result in a higher error rate as compared to the base tagger.

The features from the HGC improve the tagging accuracy for all tags over the base tagger, and also lead to an improvement over the LDA approach on most tags. Exceptions are foreign material (FM), unfinished words (XYB) and determiners (ART).

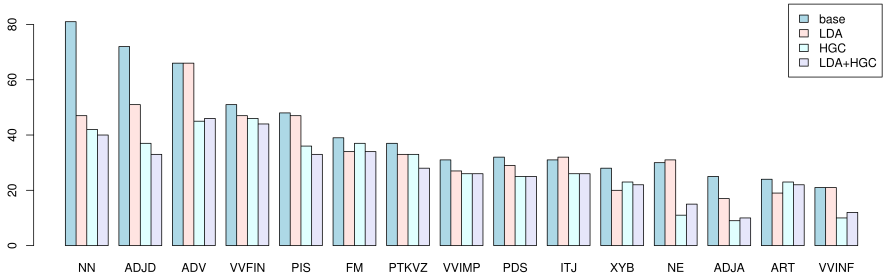


Figure 2: Error reduction for individual POS tags for the most frequent error types on the transcripts

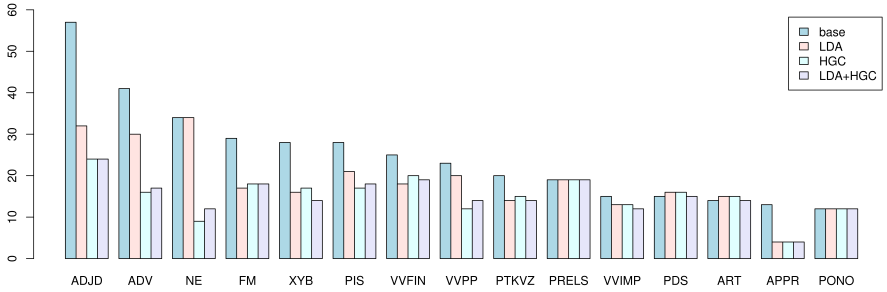


Figure 3: Error reduction for individual POS tags for the most frequent error types on the normalised data

The increased error rate for determiners can be easily explained by the different distribution of ART and PDS, the tag most commonly confused with ART, in the different resources. In the HGC, the ratio between ART and PDS, according to the TreeTagger predictions, is 33:1 while in the KiDKo we have slightly more demonstrative pronouns than determiners. This means that the features from the HGC are biased towards predicting a determiner, which has a negative impact on the accuracy of the ART and PDS tags on the KiDKo data. Additionally, determiners in the transcripts often show a deviating form like *ne* (eine; a) or *=s* (das; the) which causes the tagger to confuse it with an answer particle in the first case and with a personal pronoun in the latter case.

transcription						
POS	**IAA (%)**	**freq.**	**base**	**LDA**	**HGC**	**LDA+HGC**
PTKANT	89.59	434	96.54	96.54	96.54	96.77
PTKQU	86.15	61	75.41	78.69	78.69	78.69
PTKFILL	88.89	135	89.63	90.37	89.63	91.11
PTKINI	95.87	44	70.45	70.45	70.45	70.45
PTKONO	90.16	29	0.00	0.00	0.00	0.00
PTKPH	61.54	10	0.00	0.00	0.00	0.00
PTKRZ	45.45	38	78.95	78.95	78.95	78.95
TOTAL	88.22	751	86.15	86.55	86.42	86.82
normalised						
POS	**IAA (%)**	**freq.**	**base**	**LDA**	**HGC**	**LDA+HGC**
PTKANT	89.59	436	95.18	95.18	95.41	95.87
PTKQU	86.15	61	93.44	93.44	93.44	93.44
PTKFILL	88.89	135	94.81	96.30	96.30	96.30
PTKINI	95.87	44	77.27	77.27	77.27	77.27
PTKONO	90.16	29	0.00	0.00	0.00	0.00
PTKPH	61.54	10	30.00	30.00	30.00	30.00
PTKRZ	45.45	38	89.47	89.47	92.11	92.11
TOTAL	88.22	751	89.11	89.38	89.64	89.91

Table 12: Tagging results for the discourse markers (please note that the IAA was not computed on the same data but on a larger test set for the same POS tags and is thus not directly comparable).

For the two most frequent error tags, nouns (NN) and adjectives (ADJD), the combination of the LDA clustering and the HGC dictionary approach further reduce the error rate, showing that the two methods can provide complementary information. The same is true for VVFIN, PIS and PTKVZ.

Figure 3 shows the same analysis for the normalised data. Here, the most frequent errors made by the base tagger involved adjectives (ADJD), adverbs (ADV), proper names (NE), foreign material (FM), unfinished words (XYB), indefinite substitutive pronouns (PIS), finite verbs (VVFIN), past participles (VVPP), verb particles (PTKVZ), substitutive relative pronouns (PRELS), imperatives (VVIMP), demonstrative pronouns (PDS), determiners (ART), prepositions (APPR) and onomatopoeia and forms of echoism (PTKONO).

On the normalised data, the LDA model again results in a higher error rate for determiners and also causes a higher number of errors on demonstrative pronouns. The HGC model performs worse on foreign material, unfinished words, finite verbs and verb particles. Not surprisingly, the domain adaptation approaches have a higher impact on the original transcripts than on the normalised data.

Table 12 shows results for the discourse marker tags. To give an idea how humans perform on the same tags, we added the scores for IAA from Section 3.2 but would like to remind the reader that these scores have been obtained on a different test set and are thus not directly comparable.

At first glance the tagger does better than our human annotators. This, however, is only true for those tags with a strong most-frequent-sense baseline where the tagger has a strong bias for assigning the most frequent tag. The annoation of *ja* (yes) is a case in point. There are 206 instances of *ja* in the development set. Out of those, 7 instances are used as an interjection, 8 instances as an utterance-initial discourse particle, 29 instances of modal particles positioned in the middle field, and 162 answer particles.

All instances of *ja* as an answer particle have been tagged correctly by the tagger. However, utterance-initial discourse particles are either assigned the ADV or the PTKANT tag, and only one instance of the *ja* interjections received the correct tag while the other 6 had been annotated as PTKANT. This most-frequent-sense strategy results in an overall accuracy for PTKANT, PTKFILL and PTKRZ which is higher than the one of the human annotators. However, it would be wrong to claim that the tagger has learned to distinguish between these tags.

The tags with the lowest frequency have been ignored completely by the tagger when tagging the transcripts (PTKONO, PTKPH). On the normalised data, the tagger does at least tag some instances of the placeholder particle correctly.

5.5 Simple, lexicon-based normalisation

As seen above, there is still a huge gap between the results on the transcripts and those on the normalised version of the data. While our base tagger showed a difference in accuracy of around 6% on the transcripts and on the normalised data, the combination of the word clustering approach and the HGC dictionary method reduced the gap to around 5%.

We now try to further improve results on the transcripts by applying a normalisation dictionary extracted from the KiDKo corpus. Our dictionary has 14,030 entries, sorted by frequency. Our approach is very simple. For each word in the transcripts, we extract its most frequent normalisation and replace the word by its normalised form. We also remove colons (used to indicate lengthened vowels) from the word forms not found in the normalisation dictionary. This very simple approach gives us a further improvement of around 1% on the development set, increasing accuracy for our best setting (LDA + HGC) from 90.93 to 91.94%.

5.6 Final results

Finally, we validate the results on the held-out test set. Table 13 shows results for the different settings on the development and the test set. For convenience, we also repeat the two baselines (TreeTagger) and the results from Section 4.2 for our base tagger (without domain adaptation).

Results on the test set are in the same range as the ones on the development set. Best results are obtained by the combination of the word clustering approach and the HGC dictionary method and, for the transcripts, by applying normalisation using the simple dictionary approach. The last row of Table 13 shows results for our proposal for a coarse-grained classification, subsuming answer particles, interjections, question

	trans.		norm.	
	dev	**test**	**dev**	**test**
Baseline 1 (TreeTagger, original)	42.54	42.48	74.53	73.67
Baseline 2 (TreeTagger, re-trained)	58.42	59.90	91.76	91.56
Base tagger (lowercase)	87.24	88.43	93.95	94.14
LDA 50-10	88.95	89.53	95.02	94.89
HGC	90.66	90.81	95.86	95.66
LDA 50-10 + HGC	90.93	91.09	**95.97**	**95.77**
LDA 50-10 + HGC + normalisation	**91.94**	**91.97**	-	-
LDA 50-10 + HGC + norm., coarse (DM)	92.33	92.37	96.20	95.95

Table 13: Baselines and results for the different approaches on the development and test set

particles, backchannel signals, fillers and utterance-initial discourse particles under the label DM (Table 4). The coarse-grained classification does not have a strong impact on the overall results but slightly increases the accuracies by about half a percent.

6 Conclusions and future work

In this paper, we presented an extension of the STTS for the annotation of spoken language. Our extended tagset captures silent and filled pauses, backchannel signals, question tags, utterance-initial discourse particles, non-inflected verb forms, placeholders for ambiguous material as well as tags for unfinished or uninterpretable words. We also added a new punctuation tag for abandoned or interrupted utterances.

We showed that the new tagset can be applied by human annotators without causing an overall decrease in accuracy. We identified and discussed problematic tags and proposed a two-way classification scheme which comprises a coarse-grained tag for discourse markers, thus allowing one to consistently annotate spoken language data without spending too much time on difficult pragmatic distinctions. The fine-grained classification, paying attention to the distinctions between different discourse markers, can be used, if need be, and the fine-grained tags can always be reduced to the umbrella tag DM for those who do not wish (or trust) the more detailed analysis. Our proposal also includes a restructuring of the STTS, renaming answer particles and interjections and grouping both in the *discourse marker* category.

On the basis of our manual annotations, we developed a CRF-based POS tagger for spoken language. We showed different methods to address the out-of-vocabulary problem of our data and presented tagging accuracies of close to 92% on the original transcripts and of close to 96% on the normalised version of the data.

Much more can be done to improve the tagger. A more sophisticated approach to normalisation, for instance, might take into account the immediate context of the word form, thus reducing noise introduced by faulty normalisations. The LDA word

clustering approach will most probably benefit from a larger amount of unlabelled data, and further experiments on the normalisation of the unlabelled data used as input for the word clustering might also improve results.

Acknowledgements

This work was supported by a grant from the German Research Foundation (DFG) awarded to SFB 632 "Information Structure" of Potsdam University and Humboldt-University Berlin, Project B6: "The Kiezdeutsch Corpus". We gratefully acknowledge the work of our transcribers and annotators, Anne Junghans, Sophie Hamm, Jana Kiolbassa, Marlen Leisner, Charlotte Pauli, Nadja Reinhold and Emiel Visser. We would also like to thank the three anonymous reviewers for their insightful comments.

References

Auer, P. and Günthner, S. (2005). Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In Leuschner, T., Mortelmans, T., and de Groodt, S., editors, *Grammatikalisierung im Deutschen (Linguistik - Impulse und Tendenzen; 9.)*, pages 335–362. Berlin: de Gruyter.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–42.

Breiteneder, A., Pitzl, M.-L., Majewski, S., and Klimpfinger, T. (2006). VOICE recording – Methodological challenges in the compilation of a corpus of spoken ELF. *Nordic Journal of English Studies*, 5(2):161–188.

Chrupała, G. (2011). Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Chrupała, G. (2012). Hierarchical clustering of word class distributions. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 100–104, Montréal, Canada. Association for Computational Linguistics.

Crookes, G. (1990). The utterance and other basic units for second language discourse analysis. *Applied Linguistics*, 11:183–199.

Dürscheid, C. (2003). Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift für angewandte Linguistik*, 38:37–56.

Eynde, F. V., Zavrel, J., and Daelemans, W. (2000). Part of speech tagging and lemmatisation for the Spoken Dutch Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.

Fernandez, R. and Ginzburg, J. (2002). Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*.

Fitschen, A. (2004). *Ein computerlinguistisches Lexikon als komplexes System*. PhD thesis, Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354–375.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 of *ICASSP*, pages 517–520, San Francisco, California, USA.

Hirschmann, H., Doolittle, S., and Lüdeling, A. (2007). Syntactic annotation of non-canonical linguistic structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.

Koch, P. and Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Meer, D. (2009). "Unscharfe Ränder". Einige kategoriale überlegungen zu Konstruktionen mit dem Diskursmarker *ja* in konfrontativen Talkshowpassagen. In Günthner, S. and Bücker, J., editors, *Grammatik im Gespräch*, pages 87–114. Walter de Gruyter, Berlin, New York.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Oostdijk, N. (2000). The spoken Dutch corpus. Overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N. (2012). Part-of-speech tagging for Twitter: Word clusters and other advances. Technical Report CMU-ML-12-107, Carnegie Mellon University.

Rehbein, I., Hirschmann, H., Lüdeling, A., and Reznicek, M. (2012). Better tags give better trees - or do they? In *Proceedings of Treebanks and Linguistic Theory (TLT-10)*.

Reznicek, M., Walter, M., Schmidt, K., Lüdeling, A., Hirschmann, H., Krummes, C., and Andreas, T. (2010). *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin.

Sampson, G. (2000). *English for the computer: the SUSANNE corpus and analytic scheme*. Clarendon Press, Oxford.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT-Workshop: From Text to Tags*.

Schmid, H. and 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 '08.

Schwitalla, J. (1976). Dialogsteuerung. Vorschläge zur Untersuchung. In Berens, F. J., Jäger, K.-H., Schank, G., and Schwitalla, J., editors, *Projekt Dialogstrukturen. Ein Arbeitsbericht*, pages 73–104. Hueber, München.

Schwitalla, J. (2006). *Gesprochenes Deutsch. Eine Einführung*. Erich Schmidt Verlag, Berlin.

Selting, M., Auer, P., Barden, B., Bergmann, J., Couper-Kuhlen, E., Günthner, S., Quasthoff, U., Meier, C., Schlobinski, P., and Uhmannet, S. (1998). Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173:91–122.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, ANLP '97, pages 88–95, Washington D.C., USA.

Søgaard, A. (2010). Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics.

Stegmann, R., Telljohann, H., and Hinrichs, E. W. (2000). Stylebook for the German treebank in VERBMOBIL. Technical Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.

Telljohann, H., Hinrichs, E., and Kübler, S. (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*, pages 2229–2235.

Teuber, O. (1998). fasel beschreib erwähn – Der Inflektiv als Wortform des Deutschen. *Germanistische Linguistik*, 26(6):141–142.

Wiese, H., Freywald, U., Schalowski, S., and Mayr, K. (2012). Das KiezDeutsch-Korpus. Spontansprachliche Daten Jugendlicher aus urbanen Wohngebieten. *Deutsche Sprache*, 2(40):797–123.

7 Appendix

POS	TiGer	KiDKo	POS	TiGer	KiDKo
$(	1159	226	PRF	172	77
$.	1424	4744	PROAV	160	98
$#	n.a.	846	PTKINI	n.a.	63
$,	1539	891	PTKA	15	20
ADJA	1760	222	PTKANT	1	648
ADJD	566	887	PTKNEG	169	352
ADV	1187	2498	PTKONO	n.a.	39
APPO	12	2	PTKFILL	n.a.	212
APPR	2410	690	PTKPH	n.a.	12
APPRART	465	79	PTKQU	n.a.	94
APZR	13	0	PTKRZ	n.a.	62
ART	3124	627	PTKVZ	159	258
CARD	580	137	PTKZU	158	25
FM	27	104	PWAT	4	6
ITJ	0	592	PWAV	54	207
KOKOM	71	44	PWS	15	202
KON	713	596	TRUNC	36	5
KOUI	33	3	VAFIN	819	1368
KOUS	244	224	VAIMP	0	4
NE	1821	478	VAINF	121	35
NN	5856	1540	VAPP	51	6
PAUSE	n.a.	2475	VMFIN	258	334
PDAT	105	55	VMIMP	0	1
PDS	111	537	VMINF	18	3
PIAT	172	128	VVFIN	1090	1032
PIS	175	321	VVIMP	11	351
PTKINI	n.a.	1	VVINF	424	410
PPER	454	2292	VVIZU	53	6
PPOSAT	188	196	VVPP	586	536
PPOSS	2	15	XY	24	0
PRELAT	13	2	XYB	n.a.	115
PRELS	205	47	XYU	n.a.	733

Table 14: Distribution of POS tags in a subset of TiGer (28,827 tokens) and in KiDKo (normalised training/development/test sets; 28,827 tokens)

Author Index

Thomas Bartz
Institut für Deutsche Sprache und Literatur
Technische Universität Dortmund
thomas.bartz@uni-dortmund.de

Kathrin Beck
Seminar für Sprachwissenschaft
Universität Tübingen
kbeck@sfs.uni-tuebingen.de

Michael Beißwenger
Institut für Deutsche Sprache und Literatur
Technische Universität Dortmund
michael.beisswenger@uni-dortmund.de

Simon Clematide
Institut für Computerlinguistik
Universität Zürich
simon.clematide@cl.uzh.ch

Stefanie Dipper
Sprachwissenschaftliches Institut

Ruhr-Universität Bochum
dipper@linguistics.rub.de

Karin Donhauser
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
karin.donhauser@staff.hu-berlin.de

Ulrich Heid
Institut für Informationswissenschaft und Sprachtechnologie
Universität Hildesheim
heid@uni-hildesheim.de

Erhard Hinrichs
Seminar für Sprachwissenschaft
Universität Tübingen
erhard.hinrichs@uni-tuebingen.de

Thomas Klein
Abteilung für Germanistische Linguistik
Universität Bonn
thomas.klein@uni-bonn.de

Sandra Kübler
Department of Linguistics
Indiana University, Bloomington, IN
skuebler@indiana.edu

Sonja Linde

Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
lindeson@cms.hu-berlin.de

Wolfgang Maier
Institut für Sprache und Information
Universität Düsseldorf
maierw@hhu.de

Stefan Müller
Max Weber Stiftung
Bonn
mueller@maxweberstiftung.de

Ines Rehbein
Institut für Germanistik
Universität Potsdam
irehbein@uni-potsdam.de

Marc Reznicek
Facultad de Filología
Universidad Complutense de Madrid
mreznice@ucm.es

† Sören Schalowsky
Institut für Germanistik
Universität Potsdam

Thomas Schmidt

Institut für Deutsche Sprache
Mannheim
thomas.schmidt@ids-mannheim.de

Angelika Storrer
Institut für Deutsche Sprache und Literatur
Technische Universität Dortmund
angelika.storrer@uni-dortmund.de

Heike Telljohann
Seminar für Sprachwissenschaft
Universität Tübingen
telljohann@sfs.uni-tuebingen.de

Yannick Versley
Institut für Computerlinguistik
Ruprechts-Karls-Universität Heidelberg
versley@cl.uni-heidelberg.de

Klaus-Peter Wegera
Germanistisches Institut
Ruhr-Universität Bochum
klaus-peter.wegera@ruhr-uni-bochum.de

Swantje Westpfahl
Institut für Deutsche Sprache
Mannheim
westpfahl@ids-mannheim.de

Thomas Zastrow
Rechenzentrum der Max-Planck-Gesellschaft Garching (RZG)
thomas.zastrow@rzg.mpg.de

Heike Zinsmeister
Institut für Germanistik
Universität Hamburg
heike.zinsmeister@uni-hamburg.de