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.

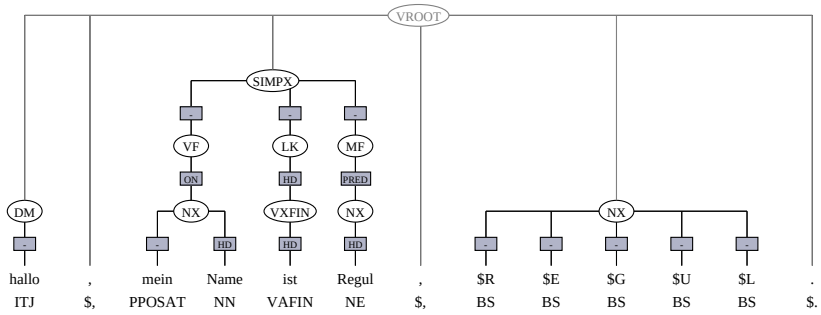

Abbildung 1: Baumbeispiel aus der TüBa-D/S

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.

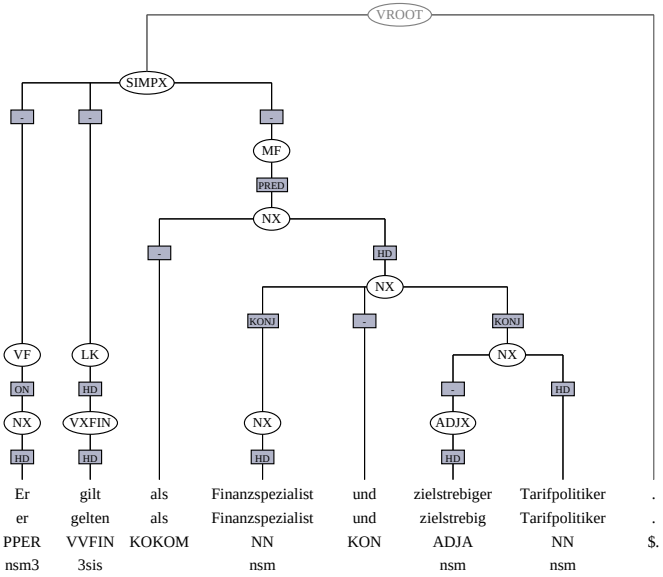

Abbildung 2: POS-Tagging des nichtkomparativen *als* in der TüBa-D/Z

(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

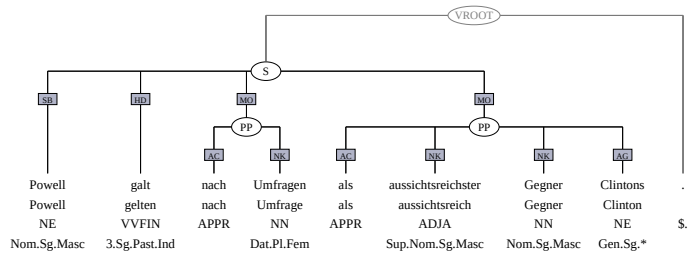

Abbildung 3: POS-Tagging des nichtkomparativen *als* in der TIGER-Baumbank

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.

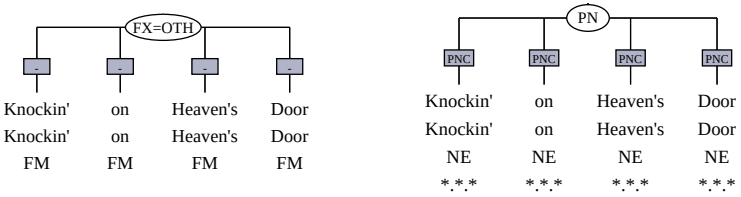

Abbildung 4: Named Entity in der TüBa-D/Z (links) und der TIGER-Baumbank (rechts)

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).

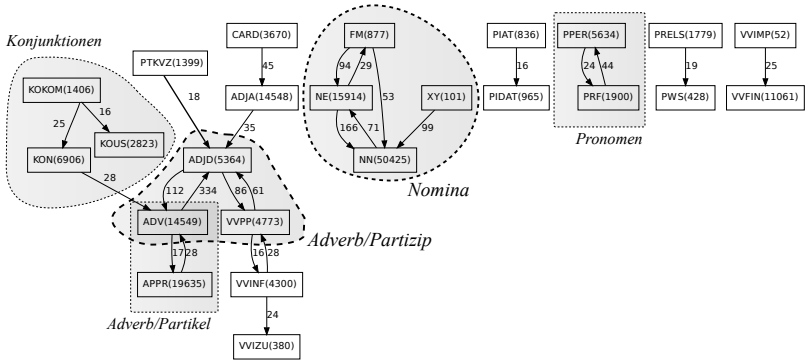

Abbildung 5: Häufigste POS-Änderungen zwischen Release 1 und Release 8

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:

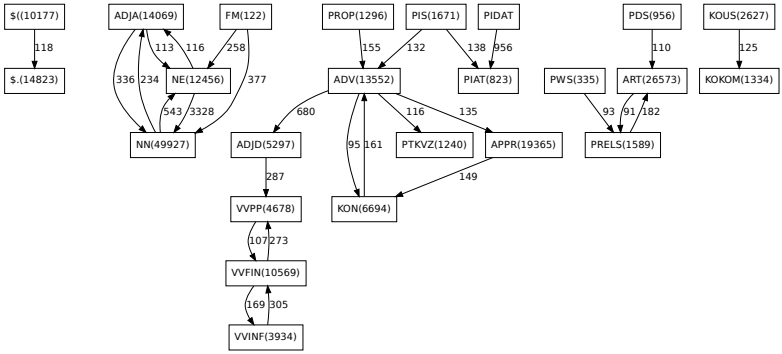

Abbildung 6: Häufigste POS-Fehler des TreeTagger

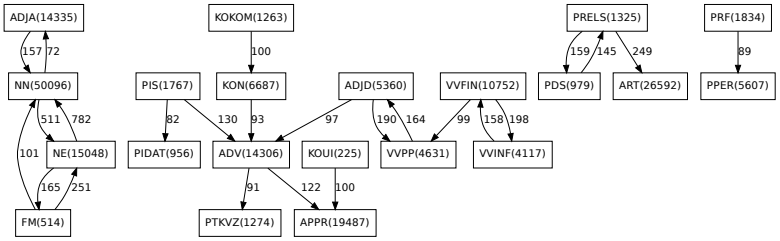

Abbildung 7: Häufigste POS-Fehler des PCFG-Parsers

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.