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.

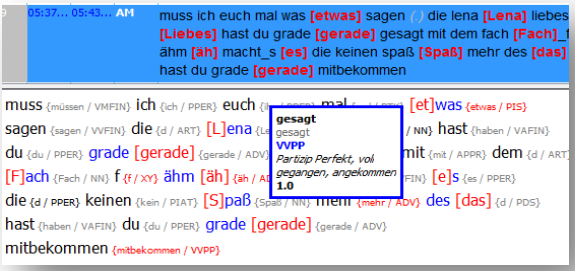

Abbildung 2: Anzeige der Wahrscheinlichkeiten und des Tags in OrthoNormal (Version 0.6)

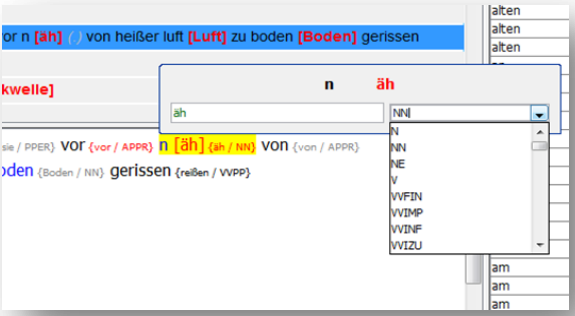

Abbildung 3: Korrektur der Tags in OrthoNormal (Version 0.6)

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.