Stand und Perspektiven des Wortarttagsets STTS: Einführung in den Workshop

Ulrich Heid

Als Einführung in die Thematik des Workshops gehen wir von den Zielen und Aufgaben der CLARIN-D - Standorte Tübingen und Stuttgart aus, zu denen die Verbesserung und nachhaltige Zurverfügungstellung von Sprachressourcen gehört.

Nach einer Kurzcharakteristik von STTS und der Nennung einiger seiner "Varianten" wird anhand der Fallstudie von Giesbrecht/Evert 2009 kurz darauf eingegangen, welche Arten von Fragen sich für die Überarbeitung von STTS stellen können: dies sind einerseits linguistische Fragen, die mit der Klassifikation von Wörtern im Kontext und, abstrakter, mit wünschenswerten oder möglichen Wortartenklassifizierungen überhaupt zu tun haben; andererseits Fragen, die sich aus der Verwendung einer gegebenen Tagging-Technologie ergeben: beide Arten von Problemen können Einfluss auf die Gestaltung eines Tagsets haben; allerdings will man nicht im Tagset technologie-spezifische Lösungen vorschreiben.

Den Abschluss bildet eine kurze Diskussion von Zielen und Randbedingungen für die im Workshop und in der Folgezeit geplante Arbeit zur Dokumentation und ggf. Ergänzung von STTS.

Referenz

Eugenie Giesbrecht und Stefan Evert. Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In I. Alegria, I. Leturia, and S. Sharoff, editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spanien, 2009.

STTS & CLARIN-D

Kathrin Beck

CLARIN-D hat das Ziel, linguistische Daten nachhaltig für die Forschungsgemeinschaft bereitzustellen. Ein Aspekt hiervon ist die aufführliche Dokumentation der Daten. CLARIN-D empfiehlt, Annotationskategorien eines Tagsets standardisiert im *DATA Category Registry* ISOcat zu dokumentieren (www.isocat.org). In diesem Beitrag stellen wir ISOcat kurz vor und skizzieren anhand des Tagsets des *Corpus Gesproken Nederlands* (lands.let.kun.nl/cgn/ehome.htm), wie das STTS in ISOcat repräsentiert werden kann.

Modifikationsvorschläge zu STTS – Stand der bisherigen Diskussion

Heike Zinsmeister

Im Dezember 2004 fand der 1. STTS-Workshop in Tübingen statt. Dieser Beitrag fasst die Ergebnisse der Diskussionen anhand vom damaligen Einladungsschreiben (Frank H. Müller), einem Handout (Klatt 2004), dem Protokoll des Workshop (2005) und weiteren Notizen zusammen. Ergänzt wird der Bericht durch Referenzen auf Kesselmeier und von Könemann (2010), die einzelne Problemfälle des STTS-Tagsets ausführlich in Bezug auf ihre linguistische Gültigkeit diskutieren und teilweise ergänzende Tests für die Annotation vorschlagen.

Referenzen

Katja Keßelmeier und Anneli von Könemann. 2010. Kategorisierungsprobleme bei der Wortarten-Annotation von Textkorpora. *Bochumer Linguistische Arbeiten Bla 2.* (http://www.linguistics.ruhr-uni-bochum.de/bla/002-kesselmeier-vonKoenemann2010.pdf)

Stefan Klatt. 2004. Anmerkungen zur aktuellen STTS-Version von 1999. Handout für den 1. STTS-Workshop, Tübingen.

Protokoll. 2005. Ergebnisse des STTS-Workshops in Tübingen am 9. Dezember 2004. Entwurf vom 19. März 2005 (Erstentwurf vom 20. Dezember 2004 von Daniel Hüttl, Tübingen).

Wortartentagging der Tübinger Ressourcen nach STTS – Erfahrungen mit verschiedenen Textgenres

Kathrin Beck, Erhard Hinrichs, Heike Telljohann & Yannick Versley
Universität Tübingen

Überblick über die Tübinger Ressourcen, die nach dem originalen STTS-Tagset getaggt sind

Wortartentagging mit größtmöglicher Anlehnung an das STTS-Tagset wie 1999 definiert (Schiller et al. 1999) (einzige Änderungen: PAV in PROP umbenannt, BS (Buchstabe) in TüBa-D/S hinzugefügt):

- Tübinger Baumbank des Deutschen / Zeitungskorpus – TüBa-D/Z (Telljohann et al. 2012)
- Tübinger Partiell Geparstes Korpus des Deutschen / Zeitungskorpus – TüPP-D/Z (Müller 2004)
- Tübinger Baumbank des Deutschen / Spontansprache – TüBa-D/S (Stegmann et al. 2000)

Mit TreeTagger (Schmid 1995) automatisch annotiert:

- Tübinger Baumbank des Deutschen / Diachrones Corpus – TüBa-D/DC (Hinrichs und Zastrow 2012)

Mit RFTagger (Schmid und Laws 2008) & MaltParser (Hall et al. 2006) automatisch annotiert:

- web-news (Versley und Panchenko 2012)

Bedarfsanalyse

In den oben genannten Tübinger Korpora von Zeitungssprache, gesprochener Sprache und diachroner Literatur verschiedener Genres konnten alle Token eindeutig einem STTS-Tag zugeordnet werden. Es gab keinen nennenswerten Bedarf an weiteren, bisher nicht enthaltenen Tags (einzige Ausnahme: 'BS' in TüBa-D/S) oder an feineren Unterscheidungen der vorhandenen Tags.

Alle darüber hinaus gehenden von uns benötigten Annotationen haben wir in weiteren Annotationsebenen kodiert, z.B. Morphologie, Lemmata, Eigennamen-Ebene usw.

Für das Zusammenspiel der einzelnen Annotationsebenen hat es sich bisher bewährt, dass sich das POS-Tagset auf Wortartenkennzeichnung beschränkt. Morphosyntaktische Tagger wie z. B. der RFTagger (Schmid und Laws 2008), die in einem Schritt feinere Unterscheidungen produzieren, verwenden in der Regel ein hierarchisches Tagset, das in fast allen praktischen Anwendungen in ein STTS-konformes POS-Tag und weitere morphologische Information gesplittet wird.

Modifikations- und Ergänzungsvorschläge

Momentan hat sich das STTS-Tagset in seiner aktuellen Form als de-facto-Standard fürs Deutsche etabliert. Um die damit annotierten Ressourcen und die damit trainierten linguistischen Werkzeuge interoperabel zu halten, schlagen wir vor, das STTS-Tagset für Texte in "Standardsprache" des Gegenwartsdeutschen unverändert beizubehalten. Die andernfalls notwendige Datenkuration wäre sehr aufwendig und würde sicher nur unvollständig vollzogen werden (können). Das sollte nur geschehen, wenn zwingende Gründe anstehen.

Durch die in den letzten Jahren verstärkte Ausweitung der linguistischen Annotation auf Texte, die nicht der Standardsprache des Gegenwartsdeutschen entsprechen, wie z.B. historische Texte oder Chat-Sprache hat sich ein Bedarf an einem veränderten oder erweiterten Tagset entwickelt. Wenn sich bei diesen Sprachvarianten die Zuweisungsrichtlinien der Tags ändern oder wenn die aktuell bestehenden Tags nicht ausreichen, z.B. um Emoticons zu klassifizieren, wäre der Anlass gegeben, spezialisierte Tagsets zu entwickeln. Wenn sie linguistisch motiviert und maschinell erlernbar sind, wären sie sicherlich eine sinnvolle Ergänzung zum Standard-Tagset.

Referenzen

Johan Hall, Joakim Nivre, und Jens Nilsson. Discriminative classifiers for deterministic dependency parsing. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, 316–323. 2006.

Erhard Hinrichs and Thomas Zastrow. Linguistic Annotations for a Diachronic Corpus of German. In: Linguistic Issues in Language Technology, Vol. 7. 2012. http://elanguage.net/journals/lilt/article/view/2689

Frank Henrik Müller. Stylebook for the Tübingen Partially Parsed Corpus of Written German (TÜPP-D/Z). Seminar für Sprachwissenschaft, Universitüt Tübingen, Germany. 2004. http://www.sfs.uni-tuebingen.de/tupp/dz/stylebook.pdf

Anne Schiller, Simone Teufel, Christine Stöckert und Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht, Universitäten Stuttgart und Tübingen. 1999. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Irland. 1995. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Helmut Schmid und Florian Laws. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In Proceedings of COLING 2008.

Rosmary Stegmann, Heike Telljohann und Erhard W. Hinrichs. Stylebook for the German Treebank in Verbmobil. Verbmobil-Report 239, Seminar für Sprachwissenschaft, Universität Tübingen. 2000. http://www.sfs.uni-tuebingen.de/resources/stylebook_vm_ger.pdf

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister und Kathrin Beck. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft, Universität Tübingen, Germany. 2012. http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf

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

Übersicht über die Tübinger Korpora:
http://www.sfs.uni-tuebingen.de/corpora.shtml

POS-Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)

Swantje Westpfahl & Thomas Schmidt
IDS Mannheim

Das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) ist ein im Aufbau befindliches Gesprächskorpus des Deutschen. Es besteht aus Aufnahmen und Transkriptionen authentischer Gespräche aus unterschiedlichsten Situationen. FOLK wird der wissenschaftlichen Öffentlichkeit über die Datenbank für Gesprochenes Deutsch (DGD 2.0) zugänglich gemacht.

In unserem Beitrag stellen wir eine erste Untersuchung zu einem POS-Tagging von FOLK mit Hilfe des TreeTaggers (Schmid 1995) nach dem STTS-Tagset vor. Der Beitrag umfasst erstens eine kurze Präsentation des FOLK-Annotations-Workflows. Zweitens diskutieren wir, welche besonderen Probleme sich bei der Anwendung des STTS-Tagsets auf Transkriptionen von Spontansprache stellen.

Referenzen

DGD 2.0: http://dgd.ids-mannheim.de:8080/dgd/pragdb.dgd_extern.sys_desc

FOLK: http://agd.ids-mannheim.de/folk.shtml

Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Irland. 1995. URL: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Erweiterung des STTS für gesprochene Sprache

Ines Rehbein, Sören Schalowski und Heike Wiese

Universität Potsdam
E-mail: {irehbein|soeren.schalowski|wiese}@uni-potsdam.de

In diesem Vortrag präsentieren wir eine Erweiterung des Stuttgart-Tübingen Tagsets (STTS) [1] für die Annotation von gesprochener Sprache. Die neu eingeführten Tags erfassen Besonderheiten mündlicher Kommunikation wie z.B. gefüllte und nicht-gefüllte Pausen, Abbrüche, Rezeptionspartikeln und Fragepartikeln (siehe Tabelle 1 für eine Auflistung der neu eingeführten Tags) und sind kompatibel mit dem Basis-Tagset des STTS, dem Quasi-Standard für die Annotation von Wortarten in kanonischer geschriebener Sprache.

Das Basis-Tagset des STTS wurde unverändert übernommen, während für Phänomene, die vorwiegend in gesprochener Sprache, selten jedoch in kanonischer geschriebener Sprache vorkommen, neue Wortarten-Tags eingeführt wurden (z.B. PTKREZ für Rezeptionspartikeln und PTKFILL für gefüllte Pausen). Andere Erweiterungen betreffen Wortformen, die in beiden Registern vorkommen, wie z.B. die Partikel *ja*, die in Zeitungstexten vorwiegend als Modalpartikel im Mittelfeld auftritt und als ADV annotiert wird. In gesprochener Sprache hingegen gibt es eine vielfältige Verwendung von *ja*. Am häufigsten tritt *ja* in äußerungsinitialer Position auf, wo es entweder als Antwortpartikel fungiert (1a) oder als Diskursmarker analysiert werden kann (1b). Solche distributionellen und funktionalen Unterscheidungen betrachten wir als Evidenz für die Einführung einer neuen Wortartenkategorie, im Beispiel von *ja* in (1b) die einer unspezifischen Partikel (PTK).

(1) a. **Ja** *PTKANT*, ich will auch ein Eis.

b. **Ja** *PTK* wer bist du denn ?

Unser erweitertes Tagset ermöglicht eine adäquatere Beschreibung der Charakteristika von Spontansprache, des Weiteren gewährleistet unser Ansatz die Interoperabilität mit existierenden linguistischen Ressourcen geschriebener Sprache und damit die Möglichkeit der Durchführung komparativer Korpusstudien. Darüber hinaus können neu annotierte Sprachdaten mit vorhandenen Trainingsdaten geschriebener Sprache kombiniert werden, um Systeme zur automatischen Verarbeitung natürlicher Sprache an die neue Domänen anzupassen. Erste Experimente zeigen, dass das erweiterte Schema mit hinreichender Verlässlichkeit annotiert und von automatischen Wortartentaggern gelernt werden kann.

POS	**Beschreibung**	**Beispiel**
INFL	*Inflektiv*	Morgen schreiben wir Mathe . Seufz !
PAUSE	*stille Pause*	
PTKFILL	*gefüllte Pause*	Ich **äh** ich komme auch .
PTK	*unspezifische Partikel*	**Ja** kommst Du denn auch ?
PTKREZ	*Rezeptionspartikel*	A: Ich komme auch . B: **Hm-hm** .
PTKONO	*Onomatopoeium*	Das Lied ging so **lalala** .
PTKQU	*Fragepartikel*	Du kommst auch . **Ne** ?
PTKPH	*Platzhalter*	Er hat **dings** hier .
XYB	*Wortabbruch*	Ich **ko** #
XYU	*unverständlich*	**(unverständlich)** #
$#	*abgebrochene Äußerung*	Ich ko **#**

Tabelle 1: Neu eingeführte POS-Tags für gesprochene Sprache

Literatur

[1] Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart, Universität Tübingen, 1995.

Tagging von Modeblogs

Gertrud Faaß, Universität Hildesheim

Hintergrund. Im Rahmen eines korpuslinguistischen Einführungsseminars[1] erhielten Studierende des Studiengang Internationale Kommunikation und Übersetzen (4. Sem.) die Aufgabe, ein Spezialkorpus aus einem beliebigen Themenbereich (monolingual) aus dem Web zu sammeln, es aufzubereiten und es in Bezug auf Differenzen zur Allgemeinsprache zu untersuchen. In Bezug auf die im vorgesehenen Workshop zu behandelnden Fragestellungen ist besonders eine der Arbeiten relevant: ein Korpus mit Blog-Beiträgen zum Thema Mode (ca. 57.000 tokens), erstellt von Frau Raissa Khattab. Das Korpus ist ausreichend dokumentiert und anonymisiert und kann für weitere Untersuchungen zur Verfügung gestellt werden.

emoticons. Für emoticons läßt sich sagen, dass die grundsätzliche Problematik überwiegend im Tokenisierungsprozess liegt[2]. emoticons (also z.B. *smileys* wie " ;)") werden nicht als solche erkannt und in Folge voneinander getrennt. Die im Text vorkommenden emoticons wurden daher mit einem vorverarbeitenden Shell-Skript transkribiert. Auch emoticon-ähnliche Symbole wie z.B. ❤ wurden vorab in verarbeitbare Zeichenketten transkribiert, siehe Tabelle 1

emoticon	***code***	***emoticon***	***code***
;), ;)*	WINKSYMB	:), :)*	SMILESYMB
<3	HEARTSYMB	:D*	BIGSMILESYMB
❤	HEARTSYMB	:-*	KISSSYMB

Table 1: *emoticons*/Symbole und ihre Transkription

In online verfügbaren, ähnlichen Korpora (z.B. das Corpuseye *English chat corpus*) scheinen solche Zeichenfolgen in ein Dollarzeichen ($) geändert worden zu sein, wir schlagen dagegen die Transkription sowie ein eigenes tag vor, um (vorher transkribierte) emoticons z.B. mit EMOT oder als Unterkategorie einer SYMB(ol)-Kategorie annotieren zu können. Damit könnten zukünftig auch spezifischere Untersuchungen, z.B. im Bereich Semantik, durchgeführt werden.

Tagging weiterer Ebenen. Das EAGLES-Konzept sieht die Annotation weiterer, spezifischerer linguistischer Merkmale im Rahmen besonderer Anforderungen, z.B. der Erforschung besonderer Phänomene vor. Im MODEBLOGS-Korpus fiel auf, dass Diminutive signifikant häufiger auftraten als z.B. in Webkorpora (SDEWAC), z.B. *Nagelstäbchen, Schleifchen*, etc. Wir würden dieses Phänomen gerne weitergehend untersuchen und daher das tagset entsprechend mit Unterkategorien erweitern, z.B. NN.DIM oder ähnliche.

[1] SoSe2012: *Korpuslinguistisches Experimentieren mit authentischen Texten*

[2] Der Tokenisierungsprozess ist nicht Thema dieses Workshops. Wir planen einen Beitrag für die DGfS-CL-Postersession in 2013 einzureichen, in dem unser Vorgehen sowie Vorschläge für die Integration von vorverarbeitenden Prozessen in den Tokenisierungsprozess Thema sein wird.

Thomas Bartz, Michael Beißwenger & Angelika Storrer
Technische Universität Dortmund

STTS 2.0: Überlegungen zur Modifikation/Erweiterung von STTS für das Tagging von Korpora zur internetbasierten Kommunikation

Das interaktionsorientierte, dialogische Schreiben in sozialen Netzwerken, Chats, Twitter, Online-Foren etc. unterscheidet sich auf verschiedenen sprachlichen Analyseebenen (Orthographie, Morphologie, Wortbildung, Syntax und Lexik) von der redigierten Schriftsprache in Zeitungstexten, in der Fachliteratur und in der Belletristik (vgl. Storrer 2009). Ein charakteristisches Merkmal des interaktionsorientierten Schreibens im Internet ist die Orientierung am Duktus der gesprochenen Sprache mit einem vergleichsweise höheren Anteil an Gesprächspartikeln, Interjektionen und Responsiven als in geschriebener Standardsprache. Ein zweites Charakteristikum internetbasierter Kommunikation ist, dass Wortschreibungen nicht immer den orthographischen Normen entsprechen: Man findet Kompetenz- und Performanzfehler, bewusste Verfremdungen (*froi, grinz*), Verschriftung von umgangssprachlicher Lautung von Einzelwörtern (*wech, is, net*) oder von Wortgruppen (*isse, haste* etc.). Als typische Stilmarker der sog. „Netzsprache" gelten Kurzformen unterschiedlicher Art (*g, vlt; lol, awk, imho; CU, 4U* etc.), Emotikons, Inflektive (**lach*, knuddel, indenarmnehm*) und Adressierungselemente (*@heike*, *@Moderator*). Alle genannten Elemente sind Bestandteile schriftlicher Äußerungssequenzen, denen beim POS-Tagging von Korpora, die auch Daten aus Chats oder Postings in Twitter, Wikis und Foren enthalten, Tags zugewiesen werden sollten.

Mit unserem Beitrag möchten wir Überlegungen dazu anstellen, welche Modifikationen und Erweiterungen des STTS-Tagsets für die Analyse von Korpora zur internetbasierten Kommunikation wünschenswert wären. Wir nutzen hierfür Datenbeispiele aus dem Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation" (*DeRiK*, Beißwenger et al. 2012) und aus dem Dortmunder Chat-Korpus (http://www.chatkorpus.tu-dortmund.de). Die Überlegungen greifen Vorschläge auf, die im Rahmen des Projekts *DeRiK* für die TEI-konforme Annotation von „interaction signs" entwickelt wurden, eine Kategorie, die sich an der Wortarteneinteilung der „Grammatik der deutschen Sprache" (Zifonun et al 1997) orientiert. Wir möchten abschließend diskutieren, inwiefern derartige Erweiterungen/Modifikationen nicht nur für die korpusgestützte Erforschung von stilistisch-rhetorischen Merkmalen internetbasierter Kommunikation, sondern auch für andere Forschungsbereiche interessant sein könnten.

Referenzen:

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): DeRiK: A German Reference Corpus of Computer-Mediated Communication. In: Proceedings of Digital Humanities 2012. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/derik-a-german-reference-corpus-of-computer-mediated-communication/

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (in press): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI). Schema files, enconig examples & documentation online: http://empirikom.net/bin/view/Themen/CmcTEI

Storrer, Angelika (2009): Rhetorisch-stilistische Eigenschaften der Sprache des Internets. In: Fix, Ulla; Gardt, Andreas; Knape, Joachim (Hgg.): Rhetorik und Stilistik. Ein internationales Handbuch historischer und systematischer Forschung. Berlin/New York: de Gruyter, S. 2211-2226.

Zifonun, Gisela/Ludger Hoffmann/Bruno Strecker u.a. (Hrsg.) (1997): Grammatik der deutschen Sprache. 3 Bde. Berlin/New York.

Marc Reznicek
Deutsche Lernerwortarten in Falko.
Was Mehrebenen-POS-tags leisten können.

Wortartenklassifikationen basieren klassischerweise auf Merkmalen unterschiedlicher linguistischer Ebenen: Für das STTS werden neben syntaktisch-distributioneller auch morphologische und lexikalische Information berücksichtigt. Während diese Einzelinformationen in Standardvarietäten idealerweise konvergieren (zu Abweichungen siehe u.a. Hirschmann (erscheint)) zeigt Lernersprache systematische Widersprüche auf den einzelnen Ebenen (siehe Díaz-Negrillo u. a. 2010).

(1)
Lerner: Viele Kriminal Aktivitäten passiert Jeden Tag in der **Heutzutager** Gesellschaft.
ZH: Viel kriminelle Aktivität passiert jeden Tag in der **heutigen** Gesellschaft.
(FalkoEssayL2v2.3:kne19_2006_07)

Für das Token „Heutzutager“ im Lernersatz 1widersprechen sich die Informationen aller drei Ebenen:

- Distribution: ADJA
- Morphologie: NN??
- Lexik: ADV

Widersprüche sind in jedweger Kombination zu finden:

Dist = Morph ≠ Lex

(2) Biologischen Verpflichtungen, die Realität der Schwangerheit schaffen nicht mehr ein **glase** Hemmung vor Frauen zu den Topjobs des Welts.
(FalkoEssayL2v2.3:fkb033_2008_07)

Lex = Morph ≠ Dist

(3) Es gibt doch freundliche Kriminale wie Robin Hood aber die meisten sind **Geldhunger** Männer, der noch mehr Geld und Mag haben wollen.
(FalkoEssayL2v2.3:sa010_2006_09)

Lex = Dist ≠ Morph

(4) Lerner: Ich bin in meinem dritten Jahr an der Universität Stellenbosch und muss **beschlossen** was ich nächstes Jahr machen werde.
(FalkoEssayL2v2.3:sa005_2006_09)

Das STTS-Annotationshandbuch empfiehlt in diesem Fall das Tag für die „richtige Wortform“ zu verwenden (Schiller u. a. 1999, S. 10). In den meisten Fällen existieren allerdings konkurrierende Korrekturmöglichkeiten oder Zielhypothesen (siehe u.a. Lüdeling 2008; Reznicek, Lüdeling und Hirschmann [erscheint]). So ist im Beispielsatz 5 nicht klar, welche der beiden Zielhypothesen besser ist.

(5)

Lerner: Man könnte optimistisch sein, und hofft, dass nichts schlimmes passiert, aber Kriminalität hat kein bestimmte **geschmackt**.
ZHa: Man könnte optimistisch sein und hoffen, dass nichts Schlimmes passiert, aber Kriminalität hat keinen bestimmten **Geschmack**.
ZHb: Man könnte optimistisch sein und hoffen, dass nichts Schlimmes passiert, aber Kriminalität hat keinem Bestimmten **geschmeckt**.
(FalkoEssayL2v2.3:sa008_2006_09)

Statt für die „korrekte Zuweisung" eines Tags plädieren wir dafür, das STTS so zu überarbeiten, dass es die Möglichkeit bietet, Wortarteninformationen auf unterschiedlichen linguistischen Ebenen unabhängig voneinander zu beschreiben und zu kombinieren. Dieses Vorgehen würde es erlauben, „to uniformly characterize well-formed language patterns as well as erroneous learner language resisting a single POS characterization" (Díaz-Negrillo u. a. 2010, S. 13).

Literatur

Schiller, Anne u. a. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS: Technical Report*. URL: `http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf`.

Lüdeling, Anke (2008). „Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora". In: *Fortgeschrittene Lernervarietäten*. Hrsg. von Maik Walter und Patrick Grommes. Bd. 520. Linguistische Arbeiten. Tübingen: Max Niemeyer Verlag, S. 119–140.

Díaz-Negrillo, Ana u. a. (2010). „Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT". In: *Language Forum*. URL: `http://purl.org/dm/papers/diaz-negrillo-et-al-09.html`.

Hirschmann, Hagen (erscheint). „Modifikation im Erwerb des Deutschen als Fremdsprache". Diss. Berlin: Humboldt-Universität zu Berlin.

Reznicek, Marc, Anke Lüdeling und Hagen Hirschmann ([erscheint]). „Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture". In: *Automatic Treatment and Analysis of Learner Corpus Data*. Hrsg. von Ana Díaz-Negrillo. John Benjamins.

DDDTS — ein POS-Tagset für historische Daten

Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum

Im Rahmen der DFG-Projekte *Referenzkorpus Altdeutsch (750–1050)* und *Referenzkorpus Mittelhochdeutsch (1050–1350)* wird ein ausgewogenes historisches Korpus des Deutschen erstellt. U.a. werden alle Wortformen morphologisch analysiert, mit POS-Tags versehen und mit historischen Lemma-Formen lemmatisiert.

Für die POS-Annotation wurde das Tagset *DDDTS* entworfen, das in der Präsentation beim STTS-Workshop vorgestellt werden soll. DDDTS orientiert sich am STTS, unterscheidet sich aber von ihm in mehrerlei Hinsicht: (i) Es benennt einige Tags des STTS systematischer, (ii) es führt ausgewählte zusätzliche Unterscheidungen ein und (iii) es nimmt spezifische Anpassung bzw. Ergänzungen für historische Daten vor. Zur Illustration sind nachfolgend einige Beispiele angeführt.

(i) Systematisierung einiger Tagnamen

Artikel: Die besondere Stellung des Artikel im Gegensatz zu Determinern im Allgemeinen wird im STTS durch ein eigenes Tag, ART, kenntlich gemacht. Im DDDTS wird diese grundsätzliche Unterscheidung beibehalten, allerdings stellen Artikel keine eigene Wortart dar, sondern eine Unterklasse der Determiner-Ausdrücke.

Pronomen: Pronomen werden im STTS unter der Hauptkategorie P zusammengefasst. Im DDDTS wird ein Großteil der Pronomen jedoch als Unterklasse der Determinativ-Ausdrücke analysiert. "Echte" Pronomen sind die, die ausschließlich "intransitiv", d.h. in der Position einer NP vorkommen können, bzw. im Falle der Pronominaladverbien: in der Position einer PP.

(ii) Zusätzliche Unterscheidungen

Artikel: Im STTS wird kein Unterschied zwischen definitem und indefinitem Artikel gemacht, im DDDTS gibt es dafür verschiedene Unterklassen, D und I.

Kardinalzahlen: Das STTS kennt nur ein Tag für Kardinalzahlen: CARD. Im Gegensatz dazu wird im DDDTS unterschieden zwischen attributiver und substantivischer Verwendung.

(iii) Historische Anpassungen bzw. Ergänzungen

Die Subklassifizierung der Adjektive unterscheidet sich von der im STTS in mehreren Punkten.

Attributiv, prädikativ, adverbiell: Im STTS werden anhand der Form zwei Verwendungen unterschieden: ADJA für attribute (und dann meist: flektierte) Adjektive, ADJD für prädikative oder adverbielle (= nicht flektierte) Adjektive.

Im Gegensatz zum heutigen Deutsch unterscheiden sich prädikative und adverbielle Adjektive in früheren Sprachstufen: Prädikative Adjektive sind endungslos (z.B. *snell*, 'schnell'), adverbielle sind suffigiert (*snello* (ahd), *snelle* (mhd)). Adverbielle Adjektive werden daher als ADV getaggt.

Ein weiterer Unterschied zum heutigen Deutsch ist, dass attributive Adjektive vor- und nachgestellt sein können (ADNJ).

Substantivisch: Zusätzlich führen wir eine Unterklasse für substantivische Adjektive ein (im STTS: NN). Als substantivisch gelten alle attributiven Adjektive, die ohne Kopfnomen auftreten. Im STTS wird die Unterscheidung zwischen substantivischen und elliptischen Adjektiven anhand der Groß- bzw. Kleinschreibung vorgenommen (*er ist der* ***Größte/NN*** vs. *sie werden als* ***letzte/ADJA*** *geheuert*, STTS S. 19). Da dieses Unterscheidungsmerkmal auf ältere Sprachstufen nicht anwendbar ist, werden diese Fälle unterschiedslos als "substantivisch" analysiert (ADJS).