

# Praktische Tagger-Kritik. Zur Evaluation des POS- Tagging des Deutschen Textarchivs

**Herrmann, J. Berenike**

berenike.herrmann@unibas.ch  
Universität Basel, Schweiz

## Einleitung

Der vorliegende Beitrag leistet eine Tool- und Methoden-Kritik der automatischen Auszeichnung von Wortarten (Part of Speech-, bzw. POS-Taggern) an literarischen Texten des 19. und frühen 20. Jahrhunderts. Er geht über eine rein intellektuelle Reflektion hinaus, indem er erste Schritte einer empirischen Evaluation des POS-Tagging des Deutschen Textarchivs (DTA, Berlin-Brandenburgische Akademie der Wissenschaften) und seiner praktischen Verbesserung vorlegt.

Aus der Perspektive der Digitalen Literaturstilistik und des Distant Reading sind Wortarten besonders interessante lexiko-grammatikalische Merkmale, ist ihre Verteilung doch ein wichtiger Indikator für Dimensionen wie Autorstil, Gattung und Register (z.B. Biber / Conrad 2009). POS sind vergleichsweise leicht und scheinbar valide zu bestimmen, gilt doch in der Computerlinguistik das Problem der automatischen Wortartenannotation als gelöst – auch für das Deutsche, wo eine durchschnittliche Erkennungsgenauigkeit bei 95-97% liegt (vgl. Giesbrecht / Evert 2009). Für DH-Anwender scheint es also nahe zu liegen, ihre Korpora komfortabel mit out-of-the-box-Taggern zu annotieren, oder sich bereits annotierter Korpora zu bedienen, wie zum Beispiel des DTA.

Ein genauerer Blick zeigt jedoch, dass Korpora der Geisteswissenschaft, historische wie literarische, von der sprachlichen Varietät abweichen, die den Sprachmodellen der verfügbaren Tagger zugrunde liegt, also Zeitungstexten der Gegenwart (der für das Deutsche frei verfügbare Goldstandard ist derzeit TIGER, ein Korpus von 900.000 Wörtern aus der Frankfurter Rundschau, vgl. Brants et al. 2004). In Nichtstandardvarietäten sinkt die Genauigkeit des POS-Taggings rapide (vgl. z.B. Scheible et al. 2011), und teilweise sind Aussagen über die Annotationsgenauigkeit mangels Referenzstandards gar nicht möglich. Dies betrifft auch das DTA, dessen POS-Tagging bislang nicht systematisch evaluiert wurde. Insgesamt ist die DH-Community also noch recht weit von einem Goldstandard für historische literarische narrative Texte des Zeitraums entfernt.

Unser Beitrag leistet hier einen wichtigen Schritt, indem er erste Ergebnisse zur Einschätzung der Qualität

ebenso wie zur Verbesserung des Annotationstools vorlegt. Ausgehend von dem Ziel unser Korpus der Literarischen Moderne (KOLIMO <http://kolimo.uni-goettingen.de/>) valide mit POS auszuzeichnen, haben wir eine Stichprobe (N= 9.065 ) des DTA manuell nachannotiert. Unsere Methode verbindet einen Tagger-Vergleich mit einer händischen Analyse. Dabei werden folgende Ziele verfolgt:

- Eine erste Evaluation des POS-Tagging des DTA für den Zeitraum 1800-1930 im Vergleich mit der gegenwärtigen Generation der POS-Tagger;
- Der heuristische Aufweis von interessanten Fällen, die Forschungsdesiderate für Linguistik und Literaturwissenschaft aufzeigen;
- Die Verbesserung des Sprachmodells und so eine Domänen-Adaptation der Tagger.

## Studie

### Prozedur

Die Evaluation des POS-Taggings wurde durchgeführt auf einer randomisierten Stichprobe des DTA, die aufgrund unseres Forschungsinteresses auf narrative Texte mit Publikationsdatum ab 1800 beschränkt war, wobei sowohl fiktionale wie auch nicht-fiktionale Texte berücksichtigt wurden (ausschlaggebend waren die Metadaten zur Erstveröffentlichung und Gattung im Header des DTA). Die Grundgesamtheit der aus dem DTA entnommenen Stichprobe umfasste N= 64.924.458 Tokens, die der händisch annotierten Tokens umfasste n= 9.065 Tokens/ POS-Tags, also 0,014%). Die Stichprobe wurde in ihrer tokenisierten und normalisierten Form aus dem DTA übernommen (vgl. DTA). Der Taggervergleich nutzte neben dem DTA-Tagger *moot* (Jurish / Würzner 2013) den TreeTagger (Schmid 1994), MarMoT (Müller et al. 2013) sowie den Perceptron-Tagger (Rosenblatt 1958), also solche Tagger, die in der digitalen Textanalyse häufig verwendet werden. Input war für alle Tagger dieselbe Stichprobe aus dem DTA.

Das Tagging wurde durch vier studentische Hilfskräfte besorgt, wobei iterative Analysen und finale Annotation durch die PI betreut wurden. Mit einem Skript wurden csv-Tabellen erstellt, die die Tokens (fortlaufende Wortformen, inklusive Interpunktion) und POS-Tags in einem *Keyword-in-Context*-Format präsentieren. Abbildung 1 zeigt einen Ausschnitt der Ansicht des Annotationstools: jede Zeile enthält neben dem Token (Wort) die jeweiligen POS-Tags, das Lemma, den linken und rechten Satzkontext, sowie einen größeren Satzkontext, Angaben zu Werktitel, Autor, und Erscheinungsdatum. Bei der Analyse wurde jeweils nur die Abweichungen zum (von *moot* zugewiesenen) DTA-Tag händisch in eine gesonderte Zelle (*newtag*) eingefügt, ebenso wie ein fakultativer Kommentar des Coders.

Obwohl die Gesamtergebnisse noch ausstehen, könnte der Unterschied zwischen den Phasen tentativ damit erklärt werden, dass Phase II mehr problematische Tags annotiert, die eine niedrigere Distribution haben und im per-Satz-Tagging seltener auftreten. Für die einzelnen POS-Kategorien variiert die Genauigkeit zwischen 0% und 100%, wobei *moot* die höchste mittlere Genauigkeit (Mittelwert = 88,65%) und niedrigste Streuung (Standardabweichung = 17,25) aufweist (TreeTagger =

67,05 ± 28,52, MarMoT = 73,41 ± 27,49, Perceptron = 62,66 ± 31,37). Eine detaillierte Analyse der einzelnen POS-Tag-Kategorien zeigt, dass *moot* in den meisten, aber nicht allen, POS-Kategorien die besten Ergebnisse erzielt (vgl. Tabelle 2).

	moot	TreeTagger	MarMoT	Perceptron
ADJA	94,5	93	93,5	95
ADJD	85,37	79,67	78,05	76,42
ADV	81,2	72,93	75,56	68,42
NE	75,25	61,87	63,55	87,63
NN	92,81	93,46	92,32	91,67

Tabelle 2: Genauigkeit einiger STTS-Tags über Tagger (in Prozent)

## Diskussion

In den Gruppendiskussionen konnten Probleme identifiziert werden, die vornehmlich bei den Taggern lagen (z.B. bei Abkürzungen, Relativpronomen). Es traten aber auch Fälle auf, in denen die STTS-Guideline nicht präzise genug ist (z.B. bei Vergleichspartikeln, Possessivpronomen, Indefinitpronomen). Dabei war die Analyse der Disagreements eine produktive Heuristik, um (computer-)linguistisch und literaturwissenschaftlich interessante Fälle aufzuwerfen. So scheint gerade in literarischen Fällen eine Ambiguität (etwa zwischen Adjektiv und Verb bei Partizipien) geradezu intentional. Ähnlich und in „Bravo! Warum denn nicht? Bravo! Und wieder Bravo!“ (Kafka, *Der Prozess*), welches als Konjunktion, aber auch als Diskurspartikel interpretiert werden kann.

Insgesamt zeigt unsere praktische Taggerkritik, dass auch eine scheinbar gelöste NLP-Aufgabe wie die Wortartenauszeichnung kein Solitär ist, auf den geistes- und literaturwissenschaftliche Projekte ohne genauere Prüfung bauen sollten. Unsere Ergebnisse zeigen trotz der hochqualitativen Vorverarbeitung des DTA eine Fehlerrate von ca. 9% an, die allerdings stark nach POS-Tag variiert. Die diachrone wie synchrone Heterogenität des literarischen Diskurses führt generische POS-Tagger bislang fast zwangsläufig an ihre Grenzen, durch historische Sprachformen, aber auch die Vielfalt der Gattungen, Erzähltechniken und kreative Lexik und Syntax. Zukünftig bieten sich hier wohl zwei Wege an: zum einen die fortlaufende Verbesserung von generischen Tools, zum anderen gerade aber auch die Feinabstimmung der Tools für spezifische Anwendungen, mit flexibel ansprechbaren Tagging- und Sprachmodellen. So haben wir unsere Annotation an *moot* zurückgespielt, um das spezifische Sprachmodell zu verbessern. Die Ergebnisse unseres Taggervergleichs deuten zudem für bestimmte Tags auf die Nützlichkeit eines Ensemble-Taggings hin, bei

dem verschiedene Algorithmen verschränkt werden (van Halteren et al., 2001).

## Bibliographie

- Biber, Douglas / Conrad, Susan** (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.
- Brants, Thorsten** (2000): "Inter-annotator agreement for a German newspaper corpus", in: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2000/pdf/333.pdf>.
- Brants, Sabine / Dipper, Stefanie / Eisenberg, Peter / Hansen-Schirra, Silvia / König, Esther / Lezius, Wolfgang / Rohrer, Christian / Smith, George / Uszkoreit, Hans** (2004): "TIGER: Linguistic interpretation of a German corpus", in: *Research on Language and Computation*, 2(4): 597–620.
- Berlin-Brandenburgische Akademie der Wissenschaften**. *Deutsches Textarchiv*. <http://www.deutschestextarchiv.de/> [Letzter Zugriff 13.01.2018].
- Giesbrecht, Eugenie / Evert, Stefan** (2009). "Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus", in: Alegria, I. / Leturia, I. / Sharoff, S. (eds.): *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.
- Evert, Stefan** (2006): "How random is a corpus? The library metaphor", in: *Zeitschrift für Anglistik und Amerikanistik* 54.2: 177-190.
- Jurish, Bryan / Würzner, Kay-Michael** (2013). "Word and sentence tokenization with Hidden Markov Models", in: *JLCL* 28(2): 61-83.
- Kilgarriff, Adam** (2005): "Language is never, ever, ever, random", in: *Corpus linguistics and linguistic theory* 1(2): 263-276.
- Müller, Thomas / Schmid, Helmut / Schütze, Hinrich** (2013): "Efficient higher-order CRFs for morphological Tagging", in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.
- Rosenblatt, Frank** (1958): "The perceptron: A probabilistic model for information storage and organization in the brain", in: *Psychological Review*, 65(6): 386.
- Scheible, S. / Whitt, R. J. / Durrell, M. / Bennett, P.** (2011): "A gold standard corpus of Early Modern German" in: *Proceedings of the 5th Linguistic Annotation Workshop* 124–128.
- Schmid, Helmut** (1994). "Probabilistic part-of speech Tagging using decision trees", in: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Halteren, H. / Daelemans, W. / Zavrel, J.** (2001): "Improving accuracy in word class tagging through the combination of Machine Learning systems", in: *Computational Linguistics*, 27.