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 Zeitungtexten 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:

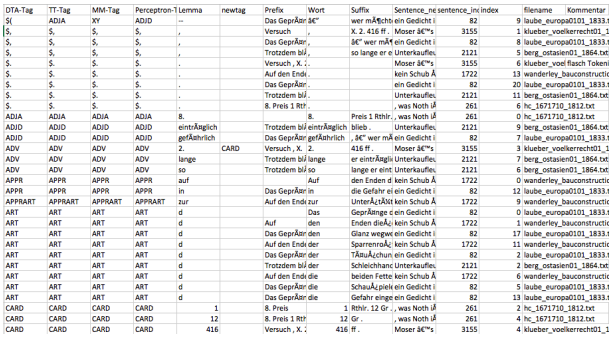
- Eine erste Evaluation des POS-Tagging des DTA für den Zeitraum 1800-1930 im Vergleich mit der gegenwärtigen Generation der POS-Tagger;
- Der heuristische Aufweis von interessanten Fällen, die Forschungsdesiderate für Linguistik und Literaturwissenschaft aufzeigen;
- Die Verbesserung des Sprachmodells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



DTA-Tag	TT-Tag	MM-Tag	Perceptron-T	Lemma	newtag	Prefix	Wort	Suffix	Sentence_ne	sentence_in	index	filename	Kommentar
$(	ADJA	XY	ADJD	--		Das GeprÃ¤n	â€"	wer mÃ¶chti	ein Gedicht i	82	9	laube_europa0101_1833.t	
$,	$,	$,	$,	,		Versuch	,	X. 2. 416 ff .	Moser â€™s	3155	1	klueber_voelkerrecht01_1	
$.	$.	$.	$.	.		Das GeprÃ¤n	,	â€" wer mÃ¶	ein Gedicht i	82	8	laube_europa0101_1833.t	
$,	$,	$,	$,	,		Trotzdem bl	,	so lange er e	Unterkauflei	2121	5	berg_ostasien01_1864.txt	
$.	$.	$.	$.	.		Versuch , X.	.		Moser â€™s	3155	6	klueber_voel flasch Tokeni	
$.	$.	$.	$.	.		Auf den Endu	.		kein Schub Ã	1722	13	wanderley_bauconstructio	
$.	$.	$.	$.	.		Das GeprÃ¤n	.		ein Gedicht i	82	20	laube_europa0101_1833.t	
$.	$.	$.	$.	.		Trotzdem bl	.		Unterkauflei	2121	11	berg_ostasien01_1864.txt	
$.	$.	$.	$.	.		8. Preis 1 Rth	.		, was Noth iÃ	261	6	hc_1671710_1812.txt	
ADJA	ADJA	ADJA	ADJA	8.			8.	Preis 1 Rthlr.	, was Noth iÃ	261	0	hc_1671710_1812.txt	
ADJD	ADJD	ADJD	ADJD	eintrÃ¤glich		Trotzdem bl	eintrÃ¤glich	bleib .	Unterkauflei	2121	9	berg_ostasien01_1864.txt	
ADJD	ADJD	ADJD	ADJD	gefÃ¤hrlich		Das GeprÃ¤n	gefÃ¤hrlich	, â€" wer mÃ¶	ein Gedicht i	82	7	laube_europa0101_1833.t	
ADV	ADV	ADV	ADV	2.	CARD	Versuch , X.	2.	416 ff .	Moser â€™s	3155	3	klueber_voelkerrecht01_1	
ADV	ADV	ADV	ADV	lange		Trotzdem bl	lange	er eintrÃ¤gli	Unterkauflei	2121	7	berg_ostasien01_1864.txt	
ADV	ADV	ADV	ADV	so		Trotzdem bl	so	lange er eint	Unterkauflei	2121	6	berg_ostasien01_1864.txt	
APPR	APPR	APPR	APPR	auf			Auf	den Enden d	kein Schub Ã	1722	0	wanderley_bauconstructio	
APPR	APPR	APPR	APPR	in		Das GeprÃ¤n	in	die Gefahr ei	ein Gedicht i	82	12	laube_europa0101_1833.t	
APPRART	APPRART	APPRART	APPRART	zur		Auf den Endu	zur	UnterÃ¿tÃ¼t	kein Schub Ã	1722	9	wanderley_bauconstructio	
ART	ART	ART	ART	d			Das	GeprÃ¤nge d	ein Gedicht i	82	0	laube_europa0101_1833.t	
ART	ART	ART	ART	d		Auf	den	Enden dieÃ¿	kein Schub Ã	1722	1	wanderley_bauconstructio	
ART	ART	ART	ART	d		Das GeprÃ¤n	den	Glanz wegw	ein Gedicht i	82	17	laube_europa0101_1833.t	
ART	ART	ART	ART	d		Auf den Endu	der	SparrenroÃ¿	kein Schub Ã	1722	11	wanderley_bauconstructio	
ART	ART	ART	ART	d		Das GeprÃ¤n	der	TÃ¤uÃ¿chun	ein Gedicht i	82	2	laube_europa0101_1833.t	
ART	ART	ART	ART	d		Trotzdem bl	der	Schleichhanc	Unterkauflei	2121	2	berg_ostasien01_1864.txt	
ART	ART	ART	ART	d		Auf den Endu	die	beiden Fette	kein Schub Ã	1722	6	wanderley_bauconstructio	
ART	ART	ART	ART	d		Das GeprÃ¤n	die	SchauÃ¿piele	ein Gedicht i	82	5	laube_europa0101_1833.t	
ART	ART	ART	ART	d		Das GeprÃ¤n	die	Gefahr einge	ein Gedicht i	82	13	laube_europa0101_1833.t	
CARD	CARD	CARD	CARD	1		8. Preis	1	Rthlr. 12 Gr .	, was Noth iÃ	261	2	hc_1671710_1812.txt	
CARD	CARD	CARD	CARD	12		8. Preis 1 Rth	12	Gr .	, was Noth iÃ	261	4	hc_1671710_1812.txt	
CARD	CARD	CARD	CARD	416		Versuch , X. 2	416	ff .	Moser â€™s	3155	4	klueber_voelkerrecht01_1	

Abbildung 1: Screenshot POS-Annotationstool (Ausschnitt)

Dem automatischen wie händischen Tagging lag das Tagset des STTS (Schiller et al. 1999) zugrunde. Wo angezeigt, wurden im Projekt zusätzliche Regeln für der Handhabung des STTS-Manuals vereinbart und dokumentiert. Hierzu gehört auch u.a. die systematische Einbindung eines korpusbasierten Wörterbuchs (http://www.duden.de/) bei Eigennamen und Fremdwörtern.

Das Tagging wurde in drei Phasen durchgeführt. Primäres Ziel des Taggings war es, die Genauigkeit von *moot* auf der genannten Stichprobe gegen manuelles und automatisches Tagging (TT, MarMot, Perceptron) zu evaluieren. Die Phase I diente neben der Erarbeitung und Ergänzung des Tagging-Manuals und dem Einarbeiten der Coder einer ersten quantitativen und qualitativen Analyse des *moot*-Taggings. Die untersuchte Stichprobe umfasste N=100 randomisiert extrahierte Sätze (N=3.635 Tokens). Jedes Token wurde bezüglich des zugewiesenen POS-Tag durch alle Coder unabhängig überprüft und ggf. korrigiert.

In Phase II wurden dieselben Coder, Tagger und dieselbe Software eingesetzt, ebenso wie die in Phase 1 erarbeiteten Tagging-Guidelines. Anders als in der ersten Phase wurden jedoch nicht ganze Sätze, sondern jeweils einhundert Token pro POS-Kategorie aus dem DTA extrahiert. Dadurch wurde eine Ungleichverteilung der einzelnen POS-Kategorien, welche in natürlichen Sätzen gegeben ist (vgl. Evert 2006; Kilgarriff 2005), vermieden. Für fünfundfünfzig POS-Kategorien des STTS wurden jeweils n=100 Wort/Token-POS-Paare sortiert nach STTS-Tag annotiert. Dies entspricht einer Grundgesamtheit von N=5.500 Tokens. Jedes Token wurde von zwei Codern unabhängig annotiert.

In der anschließenden Diskussionsphase wurden die strittigen Fälle besprochen und finale Annotationen erarbeitet. Zu diesem Zweck wurden Statistiken für die Tags (über Coder und Tagger) analysiert und Nichtübereinstimmung der vergebenen Tags identifiziert. Für die statistische Evaluation wurde die Interrater-Reliabilität als *Agreement* und Cohen‘s Kappa berechnet (Package „irr“ in R Version 3.3). Darüber hinaus wurde für die erste Tagging-Phase ein Fleiss’-Kappa für die Coder berechnet (dies steht für Phase 2 noch aus).

Ergebnisse

Die Ergebnisse in Tabelle 1 basieren für Phase I auf den finalen Annotationen und für Phase II zum momentanen Zeitpunkt auf etwa der Hälfte der finalen Annotation. Sie zeigen für *moot* verglichen mit dem Referenzstandard eine Gesamtgenauigkeit von 90,16% (TreeTagger 80,88%; MarMoT 83,99%; Perceptron 79,75%). Dies ist eine niedrige Gesamtgenauigkeit gemessen an 98,6% zur modernen Standardvarietät (Brants 2000), entspricht aber in etwa den von Scheible et al. (2011) für das Frühneuhochdeutsche erhobenen 91,6%.

		moot	TreeTagger	MarMoT	Perceptron
Phase I	Mittelwert Coder	93,44	84,75	86,43	86,23
	Referenzstandard	92,93	85,47	87,18	86,35
Phase II	Mittelwert Coder	84,93	67,87	72,32	62,34
	Referenzstandard	82,94	68,73	72,29	62,33
Kombinierter Wert Phase I & II*	Mittelwert Coder	89,17	75,92	79,11	73,72
	Referenzstandard	90,16	80,88	83,09	79,76

Tabelle 1: *Genauigkeit der Tagger bzgl. Coder (Mittelwert über Coder vor Diskussion) und Referenzstandard (nach Diskussion) in Prozent*

*Es handelt sich nicht um den Mittelwert von Phase I und II sondern um eine gewichtete Statistik, die die unterschiedlichen Stichprobengrößen (zum Zeitpunkt der Rechnung) einbezieht.

Die Übereinstimmung zwischen Codern vor der Diskussion und Referenzstandard ist hingegen vergleichsweise hoch, auch wenn sie in der zweiten Tagging-Phase etwas abfällt (*Agreement* Phase I = 95,47 – 98,13%, *Agreement* Phase II = 95,56 – 96,22%, Cohen‘s Kappa Phase I = 0,95 – 0,98, Cohen‘s Kappa Phase II = 0,95 – 0,96). Gleiches gilt für die Interrater-Reliabilität (Übereinstimmung zwischen Codern vor Diskussion) obwohl die Differenz zwischen den beiden Phasen größer ist (*Agreement* Phase I = 94,14 – 95,20%, *Agreement* Phase II = 89,45 – 92,64%, Cohen‘s Kappa Phase I = 0,94 – 0,95, Cohen‘s Kappa Phase II = 0,89 – 0,92). Das Fleiss’ Kappa weist mit 0,94 einen hohen Wert auf. Die Coder annotieren in beiden Phasen also genauer als *moot*.

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, Indefinitpronomina). Dabei war die Analyse der Disagreements eine produktive Heuristik, um (computer-)linguistisch und literarturwissenschaftlich 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 Taggervergleich 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.*