Marc Reznicek, Heike Zinsmeister

STTS-Konfusionsklassen beim Tagging von Fremdsprachlernertexten

1 Motivation

Für viele aktuelle Fragestellungen der Zweit- und Fremdspracherwerbsforschung („L2-Erwerbsforschung“) sind Lernerkorpora unverzichtbar geworden. Sie stellen Texte von L2-Lernern[1] zur Verfügung, oftmals ergänzt durch vergleichbare Texte von Muttersprachlern der Zielsprache.

Beschränkten sich Analysen der Lernerkorpusforschung in den ersten Jahren hauptsächlich auf einzelne Wortformen (vgl. Granger, 1998), hat sich das Forschungsinteresse beständig hin zu komplexeren grammatischen Kategorien entwickelt. Dazu zählen u.A. die Untersuchung tiefer syntaktischer Analysen (Dickinson und Ragheb, 2009; Hirschmann et al., 2013, u.a.) oder die Strategien der Markierung von Kohärenzrelationen (z.B. Breckle und Zinsmeister, 2012). Derartige Analysen bauen dabei nur selten auf der Textoberfläche selbst auf, sondern setzen i.d.R. die Annotation von Wortarten für jedes Texttoken voraus und ggfs. weitere, darauf aufbauende Annotationsebenen.

Annotationen dienen generell immer der Suche nach Klassen in den Daten, die anhand der Oberflächenformen allein nicht leicht zugänglich wären (im Kontext von Lernerkorpora vgl. Díaz-Negrillo et al., 2010). Ist man z.B. an einer Analyse von Possessivpronomen interessiert, würde man bei einer Korpussuche, die nur Zugriff auf die Wortformen selbst hat, bei der ambigen Form *meinen* neben Beispielen für das Possessivpronomen (1) auch alle Belege für die gleichlautende Verbform (2) finden. Das Suchergebnis wäre also sehr ‘unsauber’, da die Wortform selbst keinen Aufschluss über ihre Interpretation gibt. Eine Annotation mit Wortarten würde die beiden Lesarten disambiguieren und damit die Rückgabe der Suchanfrage präziser machen. Die Rückgabe würde weniger ungewünschte Lesarten enthalten, die man andernfalls bei der Ergebnissichtung manuell ausschließen müsste. Kurz gesagt, eine Suchanfrage auf Wortarten-annotierten Daten ist für den Nutzer effizienter als eine Suche auf reinen Wortformen.

(1) Ich schäme mich noch immer für **meinen** Einsatz in Gorleben.
(tiger.release.dec05.0.302)[2]

(2) Viele Spanier **meinen** ohnehin, daß die „Transicion“ erst dann vollendet ist, wenn es einen reibungslosen Regierungswechsel gibt.
(tiger.release.dec05.1068)

[1] Der Konsistenz und Lesbarkeit halber verwendet dieser Text die maskuline Form von *Lerner, Sprecher* usw. Selbstverständlich schließen die Nennungen auch weibliche Lerner, Sprecher usw. ein.

[2] Alle Korpusbeispiele sind auffindbar über `https://korpling.german.hu-berlin.de/annis3`.

Für die Wortartenannotation des Deutschen werden von vielen Projekten die 54 Wortartenklassen des sog. kleinen Stuttgart-Tübingen-Tagsets STTS (Schiller et al., 1999) genutzt, welches sich dadurch als eine Art Standard etabliert hat. Da das STTS für die Analyse von geschriebener Standardsprache entwickelt wurde (z.B. für Zeitungstexte), stellt sich die Frage, in wie weit es auch für die Analyse von lernersprachlichen Texten einsetzbar ist.

Beispiel (3) illustriert typische Probleme, mit denen man sich bei der Wortartenanalyse von Lernertexten auseinandersetzen muss. Zum Beispiel existiert der Ausdruck *Kriminal* zielsprachlich nur als gebundenes Morphem, wird hier aber selbstständig genutzt. Das Wort *heutzutage* ist zielsprachlich ein Adverb, das nicht flektiert und nicht attribuierend wie ein Adjektiv zusammen mit einem Nomen verwendet werden kann, hier aber genau in dieser Verwendung erscheint und zudem großgeschrieben ist, was zielsprachlich außer am Satzanfang nur Substantiven und Substantivierungen vorbehalten ist.[3]

(3) Jeden Tag viele **Kriminal Aktivitäten** passiert in der **Heutzutager** Gesellschaft. (FalkoEssayL2v2.4)

In dieser Studie soll nun untersucht werden, inwieweit sich die Besonderheiten geschriebener Lernersprache wie in Beispiel (3) illustriert auf das STTS abbilden lassen, um dann Vorschläge dafür zu präsentieren, wie man mit problematischen Fällen umgehen kann.

Um beispielhaft zu illustrieren, welche Wortartentags ein automatischer Tagger nichtzielsprachlichen Strukturen in Lernertexten zuweist, zeigt Beispiel (4) den Satz aus Beispiel (3) erneut diesmal erweitert mit automatisch zugewiesenen STTS-Tags.

Für das Token *Heutzutager* deutet die Information des lexikalischen Stammes *heutzutage* eindeutig auf ADV (Adverb), Großschreibung und *-er* -Suffix hingegen legen ein NN (normales Nomen) nahe. Bezogen auf das Auftreten eines einzelnen Tokens ist allerdings zwischen ART (Artikel) *der* und dem Nomen *Gesellschaft* in der Zielsprache distributionell sehr stark ADJA (attribuierendes Adjektiv) bevorzugt.

(4) Jeden Tag viele Kriminal/NN Aktivitäten passiert/VVPP in der/ART Heutzutager/NN Gesellschaft/NN. (TreeTagger@FalkoEssayL2v2.4)

Für die Anwendbarkeit des STTS auf Lernervarietäten ergeben sich aus diesen Beobachtungen drei Fragestellungen.

- Welche Strukturen in Texten deutscher Lerner werden durch das STTS nicht adäquat abgedeckt?

- Wie lassen sich Lücken in der Beschreibbarkeit lernersprachlicher Strukturen durch das zielsprachliche System aufdecken?

[3] Beispiel (3) weist weitere Eigenheiten auf z.B. dass das Vorfeld vor dem finiten Verb *passiert* mit zwei Konstituenten besetzt ist und dass *passiert* und das Subjekt im Numerus nicht kongruieren.

- Kann das STTS so erweitert oder anders angewendet werden, dass die quantitative Vergleichbarkeit zwischen Ziel- und Lernersprache aufrechterhalten wird, ohne die Information sich widersprechender POS-Tag-Hinweise zu eliminieren?

Im weiteren Verlauf dieses Artikels werden wir zuerst weiter auf die allgemeine Problematik der Wortartenklassifizierung von Lernersprache eingehen (Abschnitt 2). In Abschnitt 3 stellen wir unsere Untersuchung zum automatischen Tagging von Lernersprache vor, die ermittelt, inwieweit aktuelle auf Zeitungssprache trainierte Wortarten-Tagger in der Lage sind, die Kategorien des STTS korrekt auf Lernersprache abzubilden. Hierbei führen wir eine quantitative Methode ein, besonders interessante Fälle lernersprachlicher Strukturen zu ermitteln. Auf der Basis einer qualitative Analyse dieser Fälle (Abschnitt 5) diskutieren wir anschließend ausführlich verschiedene Herangehensweisen, den Herausforderungen des Wortartentaggings von Lernersprache gerecht zu werden (Abschnitt 6). Abschnitt 7 fasst den Artikel abschließend kurz zusammen.

2 Wortartentagging

Das Konzept von Wortarten hat einen heterogenen Charakter – ganz unabhängig von der Klassifikationen im Stuttgart-Tübingen-Tagset. Im Folgenden skizzieren wir Faktoren dieser Heterogenität und setzen sie mit den Anforderungen des automatischen Taggings und den deskriptiven Anforderungen von Lernersprache in Beziehung.

2.1 Wortarten als Merkmalsbündel

Wortarten können auf Informationen unterschiedlicher linguistischer Ebenen beruhen. Die gängigen Wortartenklassifikationen in beschreibenden Grammatiken berufen sich auf „das syntaktische Prinzip als primäres Kriterium“ (Helbig und Buscha, 2007, S. 19). Hierbei werden über die Darstellung von Prototypen (Eisenberg, 2004, S. 36) und Tests (vgl. u.a. Helbig und Buscha, 2007, S. 19) Klassen auf der Grundlage von Informationen zumindest dreier Ebenen definiert: lexikalische Information, morphologische Markierung und Distribution im Satz (vgl. Díaz-Negrillo et al., 2010, S.3). Die Gewichtung der unterschiedlichen Ebenen ist dabei nicht immer gleich. Anders als Helbig und Buscha (2007) geht das STTS bei seiner Klassifikation zunächst von einer morphologischen Unterscheidung aus (flektierbar vs. nicht-flektierbar). Distributionelle Eigenschaften werden erst im zweiten Schritt berücksichtigt, aus dem sich dann die Aufteilung der elf Hauptwortarten des STTS ableitet, vgl. Tabelle 1.
Diese Gewichtung wird beispw. bei der Unterteilung der Adjektive und Adverben deutlich. So werden sowohl *schnelle* in (5) als auch *schnell* in (6) als Adjektiv klassifiziert (attributiv vs. prädikativ), da das abstrakte lexikalische Element *schnell* prinzipiell flektierbar ist, während das nicht flektierbare aber mit *schnell* distributionell identische *gestern* in Beispiel (7) zu den Adverben zählt.[4]

[4] Für eine ausführliche Diskussion der problematischen Kategorisierung von Adverben vgl. Hirschmann (2013).

1. Nomina (N)
2. Verben (V)
3. Artikel (ART)
4. Adjektive (ADJ)
5. Pronomina (P)
6. Kardinalzahlen (CARD)
7. Adverben (ADV)
8. Konjunktionen (KO)
9. Adpositionen (AP)
10. Interjektionen (ITJ)
11. Partikeln (PTK)

Tabelle 1: Hauptwortarten des STTS (Schiller et al., 1999, S. 4)

> *Als Adverbien werden nur reine, nicht von Adjektiven abgeleitete, nicht flektierbare Modifizierer von Verben, Adjektiven, Adverbien und ganzen Sätzen verstanden.* (Schiller et al., 1999, S. 56)

(5) Die **schnelle/ADJA** Bearbeitung war erfreulich.

(6) Der Beamte arbeitete **schnell/ADJD**.

(7) Der Beamte arbeitete **gestern/ADV**.

2.2 Automatische Wortartenzuweisung

Automatische Wortartentagger (POS-Tagger)[5] besitzen oft zwei Komponenten: eine, die unmittelbar auf die lexikalische Information im Lexikon des Taggers zugreift, und eine, die ggfs. anschließend aus den alternativen Analysen auswählt. Die lexikalisch-morphologische Komponente ordnet dabei jeder Wortform (bzw. jedem Token) alle möglichen Wortartentags zu, die im Lexikon für diese Wortform aufgelistet sind. Ist eine Form im Lexikon nicht vorhanden, weisen einige Systeme mögliche Tags anhand einer morphologischen Analyse zu.[6] Die Disambiguierungskomponente nutzt, je nach System, entweder Regeln oder Wahrscheinlichkeiten von Tag-Abfolgen ('syntaktische Information')[7] oder auch komplexe Eigenschaftsbündel[8]. Führt keine dieser Methoden zu einem eindeutigen Tag, weichen Tagger auf robuste Lösungen aus, indem sie bsplw. das frequenteste Tag im Kontext übernehmen.

Lexikalische und syntaktische Informationen ermitteln viele Tagger statistisch aus manuell annotierten, sog. Trainingsdaten, bei welchen es sich aus praktischen Gründen oftmals um Zeitungsartikel handelt. Als Konsequenz sind die Tagger im Normalfall

[5] „POS" für Englisch *part-of-speech*

[6] Einer der Gutachter wies darauf hin, dass die hier angedeutete reduzierte morphologische Analyse von vollständigen morphologischen Analysen abgegrenzt werden sollte, wie sie von Systemen wie z.B. Morfette (Chrupala et al., 2008) oder SMOR (Schmid et al., 2004) durchgeführt wird.

[7] Wahrscheinlichkeiten von Tag-Abfolgen (v.a. in Hidden Markov Modellen, vgl. Jurafsky und Martin, 2009, Kapitel 5 & 6)

[8] Conditional Random Fields (Lafferty et al., 2001) erlauben es, Merkmale voneinander unabhängiger Ebenen zu nutzen wie z.B. orthographische und distributionelle Information, sowie ein unterschiedlich weites Fenster an Token vor oder nach dem zu taggenden Token auf den entsprechenden Ebenen zu betrachten.

für Standardsprache in der Form, wie sie in überregionalen Zeitschriften vertreten ist, optimiert.

In prototypischen Fällen sind sowohl lexikalische Information als auch morphologische Markierung und syntaktische Verteilung miteinander kompatibel; in Einzelfällen widersprechen sich diese allerdings wie in Beispiel (8). Während die lexikalische Information für *Wenn* und *Aber* auf eine Konjunktion bzw. Adverb verweist, erlaubt der Kontext zunächst nur (Pro)Nomen (oder Substantivierungen).[9]

(8) Schröder ist überzeugt, daß die SPD „ohne **Wenn** und **Aber** mit der Union um ökonomische Kompetenz konkurriert ..." (tiger_release_dec05_616)

In Fällen wie diesem müsste ein Hierarchisierungsmechanismus angewandt werden, der entscheidet, welcher Informationstyp Vorrang haben soll. In automatischen POS-Taggern ist dies normalerweise nur eingeschränkt möglich, da die lexikalische Information, wie oben beschrieben, oft als erster Filter dient, dessen Ausgabe von den anderen Informationstypen nur disambiguiert, aber nicht vollkommen überschrieben werden kann.

2.3 Wortarten in Lernersprache

Kann die Entscheidung für eines der STTS-Tags in Zeitungssprache (zumindest von menschlichen Experten) noch in hohem Maße eindeutig getroffen werden, ist diese Situation in Lernersprache völlig anders. Lernersprache weicht in vielen Aspekten systematisch von der Zielsprache ab (Selinker, 1972; Corder, 1986), wobei ein Teil dieser Abweichungen, wie in Beispiel (3) in Abschnitt 1 illustriert, auch die Entscheidungen des Wortartentaggings betrifft.

Aus computerlinguistischer Perspektive kann das Tagging von Lernersprache als der Versuch beschrieben werden, trotz fehlerhafter Daten korrekte Tagging-Ergebnisse im Sinne der Zielsprache zu erzielen, d.h. eine "robuste" Analyse auf "verrauschtem" Input zu erzeugen (vgl. van Rooy und Schäfer, 2002).

Studien zur Lernersyntax verfolgen hauptsächlich zwei unterschiedliche Ansätze: Fehleranalyse und kontrastive Interlanguage-Analyse. In der kontrastiven Interlanguage-Analyse (u.a. Granger, 1996, 2008) werden Subkorpora miteinander verglichen um signifikante Frequenzunterschiede auszumachen, meistens werden dabei Lernertexte mit Muttersprachlertexten verglichen.[10] Dieser Vergleich ist nur möglich, wenn über die gleichen Kategorien hinweg verglichen wird, d.h. wenn die Tagsets für Lernersprache und Zielsprache identisch oder zumindest aufeinander abbildbar sind. Vor diesem Hintergrund wird verständlich, weshalb der Versuch, eine eigene Lernersprachengrammatik mit einer eigenen Wortartenklassifizierung zu schreiben, keine großen Verfechter gefunden hat.[11]

[9] Unter "Kontext" ist hier die syntaktische Präpositionalphrase mit Koordination zu verstehen, nicht nur die lineare POS-Abfolge.

[10] Andere Vergleichsgruppen sind z.B. Texte von Lernern unterschiedlicher Muttersprachen, mit denen Transfereffekte nachgewiesen werden können. Vergleiche unterschiedlicher Kompetenzniveaus wiederum erlauben eine entwicklungsbezogene Analyse.

[11] Für andere Nichtstandard-Varietäten, beisplw. gesprochene Sprache, ist dies dagegen geschehen (Hennig und Bücker, 2008).

Im nächsten Abschnitt stellen wir eine Methode vor, um die Frage zu beantworten, wie sich sich Lücken in der Beschreibbarkeit lernersprachlicher Strukturen durch das zielsprachliche System aufdecken lassen.

3 Experiment

Um problematische Fälle der Analyse von Lernersprache mit STTS-Tags zu identifizieren, verwenden wir eine halbautomatische Methode, die auf der Idee des *Ensembletaggings* (van Halteren et al., 2001) beruht.

3.1 Tagging

Ein einzelner Tagger macht bestimmte Fehler beim Tagging. Verschiedene Tagger unterscheiden sich potenziell bei den Fehlern, die sie machen. Diese Beobachtung kann für das automatische Tagging und dessen manuelle Korrektur fruchtbar gemacht werden: Annotiert man denselben Text parallel mit mehreren Taggern, kann man davon ausgehen, dass die Einheitsentscheidung der Tagger wahrscheinlich korrekt ist, Unterschiede hingegen auf potenziell problematische Instanzen hinweisen. Wenn es darum geht, möglichst effizient eine gute Annotation zu erreichen, beschränkt man sich bei der manuellen Nachannotation auf diese Unterschiedsfälle.

In der vorliegenden Untersuchung gehen wir davon aus, dass die Unterschiedsfälle neben reinen Taggingfehlern zusätzlich auf potenzielle Abweichungen in der Lernersprache hinweisen. Um diese beiden Typen von Unterschiedsfällen zu trennen, gleichen wir die Ergebnisse der Lernertexte mit Annotationen vergleichbarer muttersprachlicher Texte ab und betrachten in der weiteren Analyse nur solche Unterschiedsfälle, die für die Lernertexte im Vergleich zu den muttersprachlichen Texten markant waren.

Für das Ensembletagging verwendeten wir drei Wortarten-Tagger, die frei verfügbar und gut dokumentiert vorliegen: den TreeTagger (Schmid, 1994, 1995), den RFTagger (Schmid und Laws, 2008) und den Stanford Tagger (Toutanova und Manning, 2000). Anstelle die Tagger auf die Lernerdaten durch Re-Training und anderen Methoden der Domänenadaptation optimal anzupassen, sollten in der Untersuchung gerade Standardmodule zum Einsatz kommen, die auf zielsprachlichen Daten trainiert worden waren. Die Abweichungen des Tagger-Ensembles an Stellen, bei denen sich die Lernersprache nicht zielsprachlich verhält, soll uns auf Probleme in der Beschreibung durch das STTS aufmerksam machen, siehe dazu die Ergebnisse in Abschnitt 5 und die Diskussion in Abschnitt 6.[12]

[12] Wir danken einem der Gutachter für den Hinweis, dass der TNT-Tagger sehr gute Ergebnisse auf STTS-annotierten Daten liefert (vgl. Giesbrecht und Evert, 2009). In zukünftigen Experimenten sollte dieser Tagger mit berücksichtigt werden. Für die vorliegende qualitative Untersuchung beschränken wir uns jedoch auf die im Text genannten Tagger.

3.2 Datengrundlage

Als Datengrundlage dienen die Texte des Kobalt-Korpus (`www.kobalt-daf.de`). Es handelt sich hierbei um Aufsätze von fortgeschrittenen Deutschlernenden, die die Frage diskutieren: „Geht es der Jugend heute besser als früher?“ Tabelle 2 fasst die Zusammenstellung des Korpus anhand der Erstsprachen der Autoren zusammen.[13]

Texttyp	Erstsprache	ISO 639-3	Textanzahl	Tokenanzahl
L2	Weißrussisch	BEL	20	14.401
L2	Chinesisch (Mandarin)	CMN	20	11.724
L2	Schwedisch	SWE	10	5.537
L1	Deutsch	DEU	20	12.410

Tabelle 2: Zusammenstellung des Kobalt-Korpus nach Erstsprachen (Release 1.4)

3.3 Normalisierung

Das Kobalt-Korpus beinhaltet neben den Originaltexten mehrere Normalisierungsebenen, sog. Zielhypothesen (vgl. Lüdeling et al., 2008; Reznicek et al., 2013). Für die vorliegende Untersuchung ist die Ebene der *grammatischen Zielhypothese* (ZH1) relevant, für die jeder Lernersatz systematisch mit einer morpho-syntaktisch grammatischen Entsprechung annotiert wird. Semantische und pragmatische Abweichungen bleiben bei dieser Ebene unberücksichtigt. Für die eigentliche Datenauswertung verwendeten wir eine vereinfachte Variante der Zielhypothese, bei der Bewegungen ignoriert wurde (ZH0)

Tabelle 3 illustriert die ZH1 und ZH0 für den Satz aus Beispiel (4). Die rechte Spalte zeigt automatisch generierte Differenztags, die auf unterschiedliche Abweichungstypen im Lernertext hinweisen.[14]

In der tokenbasierten Korrektur für die ZH1 werden unter Beibehaltung des finiten Verbs sowohl Wortstellung und Kongruenzbedingungen korrigiert (*viele Kriminal Aktivitäten*) als auch lexikalische Ersetzungen durchgeführt (*Heutzutager*). Es sei denn, man findet einen Kontext und eine Lesart, bei denen keine Korrektur notwendig wäre. *Kriminalaktivität* ist grammatisch möglich und verlangt die geringste Korrektur: Die Anzahl der mechanischen Veränderungen (edit distance) im Vergleich zu *kriminelle Aktivität* ist zwar gleich, die phonologische Abweichung ist allerdings geringer. Daher wird es nicht durch das vielleicht gängigere *kriminelle Aktivität* ersetzt.[15]

[13] Die OnDaF-Testergebnisse (`www.ondaf.de`) der Autoren entsprachen etwa dem Kompetenzniveau B2 nach dem Europäischen Referenzrahmen („oberes Mittelmaß“).

[14] Nach einer Konvention aus der Falko-Annotation, wird bei dem Tag MERGE, wie bei *Kriminal Aktivitäten* im Beispiel, kein zusätzliches CHANGE markiert. Die Differenztags werden in der vorliegenden Studie nicht betrachtet, sollten aber in zukünftigen Studien genauer untersucht werden.

[15] Eine detaillierte Operationalisierung des Konzepts *geringste Korrektur* muss noch geleistet werden.

LT	ZH1	**ZH0**	ZH0-Differenztag
Jeden	Jeden	Jeden	
Tag	Tag	Tag	
viele		**viel**	**CHA**NGE
Kriminal Aktivitäten		**Kriminalaktivität**	**MERGE**
passiert	passiert	passiert	
	viel Kriminalaktivität		
in	in	in	
der	der	der	
Heutzutager	**heutigen**	**heutigen**	**CHA**NGE
Gesellschaft	Gesellschaft	Gesellschaft	
.	.	.	

Tabelle 3: Normalisierung von Lernertext (LT) im Sinne der grammatischen Zielhypothese (ZH1). 'CHANGE'-Tag markieren eine Änderung der Buchstabenkette, 'MERGE' das Verschmelzen von Token. ZH0 entsteht durch die Wiederherstellung der urprünglichen Wortstellung aus ZH1 und liegt der aktuellen Untersuchung zugrunde.

4 Quantitative Analyse

Für die Untersuchung wurden beide Ebenen, der Lernertext als auch die Zielhypothese, mittels Ensembletagging mit STTS annotiert. Bei Nicht-Übereinstimmung der Tagger wurde die Mehrheitsentscheidung ausgegeben. Wenn alle drei Tagger unterschiedliche Tags vorschlugen, wurde auf die TreeTagger-Ausgabe als Defaultlösung zurückgegriffen, weil sie unabhängig die höchste Akkuratheit der drei Tagger auf ZH1 erreichte. Da es für den Lernertext keine „richtige" Annotation gibt, kann es auch keinen unmittelbaren Goldstandard geben, für die Normalisierung (ZH1), die der Standardgrammatik entspricht, aber schon. Daher wurden alle Fälle, bei denen die drei Tagger auf der ZH1 nicht übereinstimmten, manuell von zwei Annotatoren überprüft und ggf. korrigiert. Diese korrigierte Fassung nutzten wir auch als Goldstandard für die Lernertexte. Neben der im folgenden vorgestellten Akkuratheitsbestimmung, dient der Vergleich in erster Linie dazu, durch Abweichungen in den Annotationen potenziell nicht-zielsprachliche Strukturen in den Lernertexten zu markieren und auffindbar zu machen.

4.1 Taggingergebnisse

Tabelle 4 fasst die Akkuratheit der Tagger in Bezug auf diesen manuell erstellten (Quasi-)Goldstandard zusammen. In einem Kontrollexperiment wurden in jeweils drei Texten pro Erstsprache alle Token manuell überprüft. Die Differenz zum eigentlichen Experiment deutet an, in welchem Umfang die Taggerleistung zu optimistisch eingeschätzt wird, wenn nur die nicht-übereinstimmenden Tags korrigiert, Taggerfehler, die sich hinter einer Taggerübereinstimmung verbergen, aber ignoriert werden. Für das Weißrussische (BEL) z.B. wird als durchschnittliche Akkuratheit des Tagger-Ensembles auf den Lernertexten 96,8 % gemessen. Auf den drei Texten des Kontrollexperiments, bei denen alle Tags

manuell korrigiert wurden, auch die, bei denen die drei Tagger sich einig waren, liegt die durchschnittliche Akkuratheit bei etwa 96,2 %, also 0,6 %-Punkte niedriger.

Die Taggingergebnisse auf der Zielhypothese sind erwartungsgemäß besser als auf den Lernertexten selbst und auch die Varianz zwischen den Texten wird geringer, wie die Standardabweichungen in den Klammern zeigen. Die Akkuratheit der ZH1 für das Weißrussische liegt z.B. bei 98,0 %, also um 1,2 %-Punkte besser als das Ergebnis auf der Lernertextebene (LT).[16]

	Experiment		Kontrolle	
	LT	ZH1	LT	ZH1
BEL	96,8 (±1,2)	98,0 (±0,8)	96,2 (±1,0)	97,2 (±1,0)
CMN	97,1 (±1,5)	98,3 (±0,8)	96,9 (±11,2)	97,8 (±0,6)
SWE	95,0 (±2,0)	97,7 (±0,9)	94,8 (±2,3)	97,0 (±0,7)
DEU	95,8 (±1,6)	97,8 (±0,9)	95,6 (±1,6)	96,5 (±0,8)

Tabelle 4: Durchschnittliche Tagging-Akkuratheit auf dem Kobalt-Korpus, welche manuell nur für Token, bei denen die Tagger nicht übereinstimmten, korrigiert wurde (links), und auf einer Kontrollgruppe von je drei Texten pro Sprache, die manuell für alle Token korrigiert wurde (rechts); Standardabweichung in Klammern.

Es fällt auf, dass die Tagging-Ergebnisse auf den muttersprachlichen DEU-Texten nicht die besten sind, sondern hinter den Ergebnissen auf den BEL- und CMN-Texten zurückbleiben. Dies lässt sich damit erklären, dass die deutschen Texte ebenfalls Tippfehler und andere für die Tagger unbekannte Wörter enthalten und potenziell komplexere Strukturen verwenden. Zudem besteht eine schwache negative Korrelation zwischen durchschnittlicher Satzlänge und Taggingakkuratheit (Spearmans Rangkorrelationskoeffizient, rho = -0.26, $p<0.05$):[17] Je länger die Satzlänge desto niedriger die Taggingakkuratheit.

Sprache	Min.	Median	Mean	Max.
BEL	10,80	14,00	15,23	23,75
CMN	11,57	15,08	15,00	18,00
SWE	12,16	17,64	17,30	26,58
DEU	12,80	18,83	19,97	32,35

Tabelle 5: Durchschnittliche Satzlänge pro Text im Kobalt-Korpus (in Token pro Satz; ergänzt um die minimale und maximale beobachtete Satzlänge).

[16] Uns ist bewusst, dass auch gerade die Fälle interessant sein können, in denen die Tagger sich zwar einig, der Goldstandard aber abweichend ist. Für diese Untersuchung war der erstellte Goldstandard allerdings zu klein.

[17] Wir verwendeten den nicht-parametrischen Rangkorrelationstest nach Spearman, da nach dem Shapiro-Wilk-Test weder die durchschnittlichen Satzlängen noch die Akkuratheitswerte annähernd normalverteilt sind.

In der folgenden Untersuchung vergleichen wir die automatischen Tags der LT-Ebene mit den Tags der ZH0-Ebene und kontrastieren die Lernerdaten mit den muttersprachlichen DEU-Daten.

4.2 Konfusionsklassen

Wenn Lernersprache Strukturen beinhaltet, in denen sich die unterschiedlichen Hinweise in Hinblick auf Distribution, morphologische Markierung und lexikalische Information widersprechen, sollten die Tagger bei der Vergabe dieser Tags besonders häufig falsch liegen. Der Abgleich der vergebenen Tags mit einem Goldstandard wird klassischerweise in einer Konfusionsmatrix (Abbildung 1) dargestellt. In einer solchen Tabelle wird der Taggeroutput (vertikal) zu den Tags im Goldstandard (horizontal) in Beziehung gesetzt. Bei einem perfekten Tagger würden alle Ergebnisse auf der Diagonalen liegen. In Abbildung 1 werden der Übersichtlichkeit halber nur Verwechslungen angezeigt und die Diagonale ausgespart.

Nicht alle Verwechslungen können auf Besonderheiten der Lernertexte zurückgeführt werden. So können auch andere Variablen (z.B. Textsorteneffekt) den auf Zeitungssprache optimierten Taggern Schwierigkeiten bereiten. Um diesen Einfluss möglichst auszuschließen, betrachten wir ähnlich wie in klassischen *Under*- und *Overuse*-Studien (Granger und Tyson, 1996; Hirschmann et al., 2013) nur die signifikanten Abweichungen zwischen L1- und L2-Texten. In Abbildung 1 werden daher nur die in Lernersprache häufiger auftretenden Wortartenkonfusionen aufgezeigt. Größe zeigt dabei die logarithmische Häufigkeit, Schwärze den Grad der Signifikanz an. So lässt sich aus der Graphik ablesen, dass die Verwechslungen von satzinterner (**$(**) mit satzexterner (**$.**) Interpunktion sowie von Partizipien Perfekt von Vollverben (**VVPP**) mit entsprechenden Infinitiven (**VVINF**) jene sind, die sich in den L2-Texten am signifikantesten häufiger finden als in den L1-Texten (dunkelste Kreise). Gleichzeit kann man sehen, dass insgesamt am häufigsten Infinitive von Modalverben (**VMINF**) mit finiten Modalverben (**VMFIN**) verwechselt werden (größter Kreis), dass diese aber nicht besonders lernerspezifisch ist (heller Kreis). Sortiert man die Konfusionsklassen zuerst nach Signifikanz und dann nach Häufigkeit, erhält man die Liste in Tabelle 6.

Anhand der rein quantitativen Analyse lässt sich allerdings noch nicht entscheiden, wie die Konfusionen zu interpretieren sind. Um diese Frage zu klären, wollen wir im nächsten Abschnitt einige Beispiele genauer betrachten.

5 Qualitative Analyse

In diesem Abschnitt analysieren wir beispielhaft Vertreter der in Abschnitt 4.2 eingeführten Konfusionsklassen, um festzustellen, ob sie tatsächlich auf Lücken im Tagset hinweisen oder von Experten hätten korrekt getaggt werden können.

In den Beispielen (9)-(16) werden für jede der gefundenen Konfusionsklassen zwei Beispiele präsentiert. Für jeden Fall wird explizit gemacht, ob Lexik, Morphologie und Distribution dem Standard entsprechen oder nicht. So hat das Tagger-Ensemble für

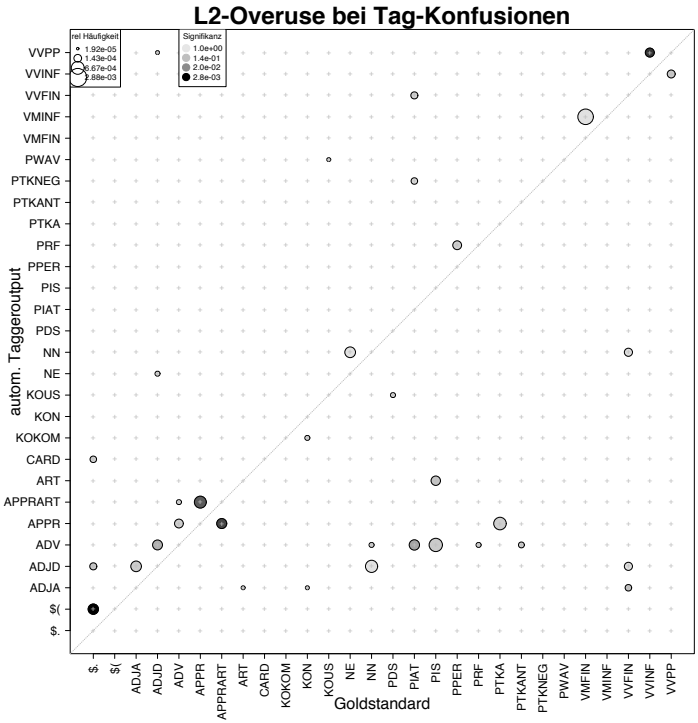

Abbildung 1: L2-Overuse bei Tag-Konfusionen auf dem Goldstandard (70 Texte: 43.285)
größer = häufiger in L1 und L2, dunkler = Kontrast L1 vs. L2 signifikanter

das Token *besprochen* im Beispiel (9) als Partizip Perfekt (**VVPP**) getaggt. Im Goldstandard wurde das in eckigen Klammern angegebene Goldtoken *besprechen* allerdings als Infinitiv des Vollverbs (**VVINF**) getaggt. Die Tatsache, dass *besprochen* an dieser Stelle nichtstandardsprachlich verwendet wird, wird erst im Kontext des vorangehenden Modalverbs erkennbar. Dieser relevante Kontext ist in den Beispielen durch Unterstreichung markiert. Am Goldtoken *besprechen* lässt sich ableiten, dass das Lemma und die distributionelle Position dem Standard entsprechen, während die morphologische Markierung nicht dem Standard folgt. Im Beispiel (13) weicht die Zielhypothese nicht für das fälschlicherweise als Artikel (ART) statt als attributives Indefinitpronomen (PIAT) getaggte Token *mehr* ab, sondern für das benachbarte *günstige*. Hier wurde das *mehr* also in einem nichtstandardsprachlichen Kontext verwendet, seine Distribution ist somit nichtstandardsprachlich.

Die Beispiele zeigen, dass eine Konfusionsklasse durch unterschiedliche Widersprüche

L2		GOLD	Signifikanz L2 vs. L1	$\sum$ L2	$\sum$ L1
$(	≠	$.	0.003	14	1
VVPP	≠	VVINF	0.004	11	0
APPR	≠	APPRART	0.004	15	0
APPRART	≠	APPR	0.006	2	2
ADV	≠	PIAT	0.048	14	1
ADV	≠	ADJD	0.075	12	1
ART	≠	PIS	0.190	10	1
ADJD	≠	ADJA	0.259	13	2

Tabelle 6: L2-spezifische Tagging-Fehler nach Signifikanz (L2 vs. L1) sortiert und Häufigkeiten

	Lexik	Morphologie	Distribution
9)	**VVPP**$_{LT}$ ≠ **VVINF**$_{Gold}$		
	Standard	**Nichtstandard**	**Standard**
	Wenn bei mir etwas passierte, kann ich dass mit meinen Eltern **besprochen**[⇒ **besprechen**]. (kobalt_BEL_018_2011_03)		
10)	**APPR**$_{LT}$ ≠ **APPRART**$_{Gold}$		
	Nichtstandard	**Standard**	**Standard**
	Junge Menschen sind oft **nach**[⇒ **zur**] Universität oder Arbeit gezogen. (kobalt_SWE_011_2012_03)		
11)	**APPRART**$_{LT}$ ≠ **APPR**$_{Gold}$		
	Nichtstandard	**Standard**	**Nichtstandard**
	früher hatte man einfach nicht die Moglichkeit herumzusitzen, fern zu schauen, **vom**[⇒ **vor** dem] Computer stundenlang zu sitzen. (kobalt_SWE_006_2011_12)		
12)	**VVINF**$_{LT}$ ≠ **VVPP**$_{Gold}$		
	Standard	**Nichtstandard**	**Standard**
	In den meisten Fällen machten sie nur das, was von ihnen **erwarten**[⇒ **erwartet**] wurde. (kobalt_BEL_012_2011_03)		

	Lexik	Morphologie	Distribution
13)		**ADV**$_{LT}$ ≠ **PIAT**$_{Gold}$	
	Standard	**Standard**	**Nichtstandard**
	Wir haben heute **mehr** günstigen[⇒ günstige] Möglichkeiten, um unseren richtigen Platz in der Welt zu finden (kobalt_BEL_015_2011_03)		
14)		**ADV**$_{LT}$ ≠ **ADJD**$_{Gold}$	
	Nichtstandard	**Standard**	**Standard**
	Wir machen es aber für uns einfach und nehmen Schweden, das kleine Land ganz **viel**[⇒ **weit**] oben auf dem Erdball. (kobalt_SWE_007_2011_12)		
15)		**ART**$_{LT}$ ≠ **PIS**$_{Gold}$	
	Standard	**Nichtstandard**	**Standard**
	Zuvor wurden das Familiensleben und das Lernen des Jugendes wegen Kriege und Reformen, die **ein** [⇒ **einer**] nach dem anderen kam, zerstört. (kobalt_CM_020_2011_03)		
16)		**ADJD**$_{LT}$ ≠ **ADJA**$_{Gold}$	
	Standard	**Nichtstandard**	**Standard**
	Andererseits, so argumentieren sie, hat die jüngere Generation zum Glück eine Welt mit **relativ**[⇒ **relativem**] Frieden und Freiheit. (kobalt_CM_008_2011_03)		

Lexik	Morphologie	Distribution	Beispiele
S	S	**NS**	13
S	**NS**	S	9, 12, 15, 16
NS	S	S	10, 14
NS	S	**NS**	11

Tabelle 7: Informationen auf linguistischen Ebenen
S: Standard, NS: Nichtstandard

der linguistischen Ebenen bedingt sein kann. Tabelle 7 fasst die einzelnen Kombinationen zusammen. Aufgrund dieser Beispiele lässt sich jetzt auch erkennen, dass die Verwechslungen tatsächlich Lernerstrukturen ermitteln, die durch das Tagset nicht beschrieben werden können, anstatt lediglich Schwächen der Tagger aufzuzeigen. Wäre dem nicht so, müsste man für die Beispiele Lesarten finden können, in denen sich die Ebenen nicht widersprechen. Dies ist aber nicht der Fall.

6 Diskussion

Wie im Abschnitt 5 deutlich geworden ist, können die Tags des STTS eine Reihe von Strukturen in Lernersprache nicht abdecken. Wir wollen drei Ansätze besprechen, mit diesem Problem umzugehen: Mehrdimensionale Tags (Abschnitt 6.1), Portemanteau-Tags (Abschnitt 6.2) und unterspezifizierte Tags (Abschnitt 6.3).

6.1 Mehrdimensionale Tags

Die konsequenteste Lösung wurde bereits von Díaz-Negrillo et al. (2010) für das Tagging von englischen Lernertexten vorgeschlagen. Um Lernersprache und Zielsprache einheitlich beschreiben zu können, wäre es sinnvoll, die Einzelinformationen auf den drei Ebenen (Lexik, Morphologie und Distribution) getrennt in einem mehrdimensionalen Tags (*tripartite POS*) anzugeben. Die Schwäche dieses Ansatzes liegt allerdings darin, dass die einzelnen Ebenen unabhängig voneinander beschrieben werden. Prinzipiell könnte man zwar sowohl die lexikalische als auch die morphologische Information aus einer Liste ziehen. Die Ambiguitäten jeder einzelnen Ebene werden normalerweise allerdings erst durch die Schnittmenge mit den anderen beiden Ebenen beherrschbar. Will man die Ebenen aber gerade unabhängig voneinander machen, so müsste das neue POS-Tags alle möglichen Kombinationen beinhalten. Dies lässt sich gut anhand von *erwarten* im Beispiel (12) zeigen, hier als (17) wiederholt.

(17) In den meisten Fällen machten sie nur das, was von ihnen erwarten [⇒ erwartet] wurde.

Rein lexikalisch handelt es sich um ein Vollverb (**VVFIN**, **VVINF**, **VVPP** oder **VVIMP** (letzteres mit abweichender Morphologie). Morphologisch im Sinne von Paradigmen können kleingeschriebene Wortformen mit Endung auf *en* attribuierende Adjektive (**ADJA**), finite oder nicht-finite Vollverben sein (**VVINF**, **VVIZU**, **VVFIN**, **VVPP**). Bezieht man geschlossene Wortformen mit ein, erhöht sich die Anzahl der möglichen Lesarten um Artikel (**ART**), Auxiliar- und Modalverben (**VAFIN**, **VAINF**, **VAPP**, **VMFIN**, **VMINF**), Pronomen (**PIAT**, **PPOSAT**, **PRELS**, **PDS**, **PIS**, **PRELAT**, **PDAT**, **PPER**, **PWAT**, **PWS**, **PPOSS**) und Kardinalzahlen (**CARD**). Operationalisiert man die morphologische Analyse einfach nur als Bedingung auf der reinen Buchstabenkette, erhöht sich die Anzahl der möglichen Tags erneut (z.B. **APPR**, **ADV**, **ADJD**, **PTKVZ**), da es insgesamt nur wenige Tags gibt, die keine kleingeschriebene Wortformen mit der Endung *en* beschreiben (z.B. **PTKNEG**, **KOKOM**,

KOUI). Insgesamt würde es sich anbieten, aus einem Korpus graduelle Erwartbarkeiten für die verschiedenen Tags zu ermitteln.[18]

Für die Distributionsebene stellt sich genauso wie bei der Morphologie die Frage nach der Operationalisierung, d.h. mit welchen Methoden die möglichen Tags ermittelt werden. Díaz-Negrillo et al. (2010) beziehen sich bei Beispielerklärungen auf den grammatischen Kontext (grammatical context) im Sinne von Phrasenstruktur und Selektionsbeschränkungen. Wenn man für das Beispiel (17) den umgebenden Teilsatz als distributionelle Suchmaske annimmt wie in (18a), dann sollte die Variable *X* mit **VVPP** gefüllt werden – gegeben, dass alle anderen Wörter korrekt formuliert sind. Wechselt man auf eine abstraktere Ebene und überlegt, welche Wortarten zwischen einer Präpositionalphrase mit Personalpronomen APPR PPER und einem finiten Auxiliarverb stehen können wie in (18b), wird die Liste der möglichen Tags natürlich länger z.B. **ADJD** wie in *...(dass der Vorsprung) vor ihm größer wurde.*

(18) a. ...was von ihnen X wurde.

b. ...APPR PPER X VAFIN

Aufgrund der oben genannten Schwierigkeiten, sehen wir in dieser Variante der Mehrebenenbeschreibung noch keinen gangbaren Weg. Weitere Forschung könnte hier einen Ausweg zeigen.

6.2 Portemanteau-Tags

In einer zweiten Variante werden Original- und Zielhypothesen-Information verbunden, indem das auf dem Originaltext vom Tagger-Ensemble zugewiesene Tag um das Tag im Goldstandard ergänzt wird. So trägt das Token *erwarten* in Beispiel (12) das Tag **VVINF_VVPP** vgl. (19).

(19) In den meisten Fällen machten sie nur das, was von ihnen **erwarten/VVFIN_VVPP** wurde.

Dieses Tag bringt zum Ausdruck, dass oberflächennähere Faktoren für einen Infinitiv sprechen, während unter Einbezug aller kontextuellen Informationen ein **VVPP** in der ZH1 stehen würde. Das Hauptproblem dieses Ansatzes liegt im fehlenden universellen Goldstandard für eine grammatische Zielhypothese. Wie bereits mehrfach erwähnt wurde, ist die Festlegung auf ein einziges Tag im Goldstandard nur möglich, indem man die Information auf einer der linguistischen Ebenen höher gewichtet als die einer anderen. Durch solch eine Hierarchisierung, geht die zugrundeliegende Information für die weitere Verarbeitung verloren. Die Kombination beider Ebenen kann dies jedoch zum Teil auffangen. Die Formulierung der grammatischen Zielhypothese folgt dabei stets den Anforderungen der jeweiligen Forschungsfrage (vgl. Reznicek et al., 2013).

[18] Zum Beispiel ergibt die Anfrage `#1:word=[a-zäöü.*en/]!` auf dem TiGer-Korpus (v2.1) 34 unterschiedliche STTS-Tags für die beschriebene Wortform, wobei sich die Erwartbarkeiten von 27% für ADJA bis hin zu unter einer Promille für z.B. NE, PWAT und APPO verteilen.

Diese Lösung ist in Mehrebenen-Korpora wie Kobalt oder Falko bereits umgesetzt. Hierbei werden die Tags allerdings nicht in einer Ebene verschmolzen, sondern sind parallel durchsuchbar, da neben dem Originaltext auch die Zielhypothese(n) und deren Annotationen ins Korpus integriert sind.

Dickinson und Ragheb (2013) verfolgen der Terminologie nach einen Mehrebenenansatz und beschreiben Lernersprache in einer morphologischen und einer distributionellen Ebene.[19] Die beiden Tags können sich dabei voneinander unterscheiden. Bei genauerer Betrachtung wird allerdings deutlich, dass es sich bei diesem Ansatz nicht um „echte" unabhängige Ebenen handelt, sondern dass die Tags auf der distributionellen Ebene eher der ZH1-Ebene entsprechen, auch wenn es sich nicht um explizite Zielhypothesen handelt. Damit gleicht die Annotation der beiden Ebenen eher der hier vorgestellten Portemanteau-Variante als den in Abschnitt 6.1 eingeführten mehrdimensionalen Tags.

> *We define a distributional slot as a position where a token with particular properties (e.g., singular noun) is predicted to occur, on the (syntactic) basis of its surrounding tokens.* (Dickinson und Ragheb, 2013, S. 27)

6.3 Unterspezifizierte Tags

Eine dritte Möglichkeit, Widersprüche zwischen den Ebenen in Lernersprache in der Beschreibung durch das STTS abzubilden, könnte darin bestehen, neue unterspezifizierte Tags zuzulassen. Innerhalb der Hauptwortarten des STTS ist dies durch den hierarchischen Aufbau der Tags schon heute leicht realisierbar. So könnten Überschneidungen aus **ADJD** und **ADJA** als **ADJ** getaggt werden. Das Tag **ADJ** ist dabei gegenüber beiden ursprünglichen Tag unterspezifiziert, da es ein Adjektiv beschreibt, aber keine Aussage darüber trifft, ob es attributiv oder prädikativ verwendet wird. Ein Blick in die relevanten Konfusionstypen in Tabelle 8 zeigt, dass sich ein Großteil der problematischen Fälle auf diese Weise abdecken lassen. Nur für Fälle, in denen Hauptwortarten überschritten werden (**ART** $\neq$ **PIS** & **ADV** $\neq$ **PIAT**), müssten neue Oberklassen im Sinne von unterspezifizierten Tags über zwei oder mehrere Hauptwortarten hinweg eingefügt werden.

Um die im Kobalt-Korpus untersuchten Strukturen adäquat abzubilden, würde es ausreichen, zwei solche Oberklassen einzuführen: **AD** als Vereinigung von Adjektiven und Adverben, sowie **ADPART**, die zusätzlich auch die Pronomen miteinschließt. Abbildung 2 zeigt die nötigen Unterspezifizierungen. Hierbei sind alle Hauptwortklassen bereits enthalten.

Inwieweit diese Oberklassen auch in anderen Varietäten genutzt werden können, muss untersucht werden. Gerade die Verschmelzung von Pronomen mit Adjektiven/Adverben scheint linguistisch unintuitiv, allerdings zeigen auch Clusteranalysen (vgl. Rapp, 2007) Ähnlichkeiten zwischen Wortformen, die nicht der linguistischen Intuition entsprechen.

[19] SALLE: `http://cl.indiana.edu/~salle/`

L2		GOLD		unterspez. Tag
VVPP	≠	VVINF	→	VV
VVINF	≠	VVPP		
APPR	≠	APPRART	→	AP
APPRART		APPR		
ADV	≠	ADJD	→	AD
ADJD	≠	ADJA		
ART	≠	PIS	→	PART
ADV	≠	PIAT	→	ADPART

Tabelle 8: Unterspezifikation für L2-relevante Konfusionsklassen

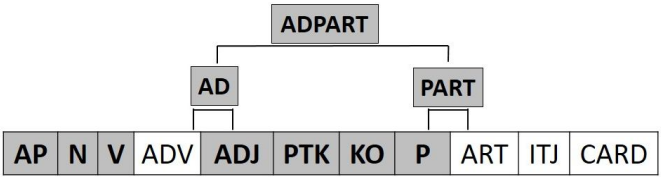

Abbildung 2: Vorschlag für Oberklassen (grau markiert) zu STTS-Hauptwortarten

6.4 Anwendung

Zwar sehen wir wie unter 6.1 für mehrdimensionale Tags noch keine Anwendungsmöglichkeiten, dagegen haben aber sowohl die Einbeziehung der ZH1 als auch die Unterspezifizierung von Tags anwendungsbezogene Stärken und Schwächen. Der Vorteil der Portemanteau-Tags liegt in der ausdifferenzierten Beschreibung der Form und der Funktion über die Tags der konkurrierenden Ebenen. Dies ist besonders für spezielle Suchen oder beispielsweise der Untersuchung von Frequenzkontrasten von POS-Ketten aussagekräftiger als die Reduktion der Tags auf Oberklassen. Für die weitere Verarbeitung (bspw. durch einen Parser) führt dieser Ansatz durch die Vielzahl möglicher Kombinationen der einzelnen Tags allerdings schnell zu *data sparseness* und verschlechtert die Ergebnisse des automatischen Trainings. Für diesen Fall scheinen die Oberklassen die bessere Lösung darzustellen, da durch die Reduktion auf weniger Tags die Anzahl der Instanzen in jeder Kategorie steigt. Darüber hinaus konnten Rehbein et al. (2012) zeigen, dass beim automatischen syntaktischen Parsing von Lernersprache nur bestimmte POS-Tag-Abweichungen die Ergebnisse verschlechtern, da sich eine Reihe von Tags in bestimmten Kontexten syntaktisch identisch verhalten.

> *This clearly shows that the overall accuracy is not enough to predict parsing scores, but that particular error types are more harmful for parser performance than others.* (Rehbein et al., 2012, S. 13)

In diesen Fällen könnte das Trainieren mit Oberklassen also sogar zu besseren Ergebnissen führen.

6.5 Einfügungen, Löschungen

Die hier vorgestellte Pilotstudie ignoriert bewusst einen wichtigen Bereich von Lernersprache: fehlende (1,6% aller Tokens) und überflüssige (0,78%) Wortformen, sowie falsche Auseinander-(0,27%) und Zusammenschreibungen (0,17%). Für all diese Fälle (insges. 2,82% aller Tokens) gibt es keine 1:1-Beziehung zwischen den Token im Lernertext und in der Zielhypothesenebene. So entspricht den beiden Tokens *Computer*/[**NN**] und *Spiele*/[**NN**] des Lernertextes im Beispiel (9) das Token *Computerspiele*/[**NN**] in der ZH1. Das Token *ein*/[**ART**] hat gar keine Entsprechung im Lernertext.

tok	**Computer**	**Spiele**	,	Online-Chatting		soziale	Netzwerk
pos	**NN**	**NN**	$(	NN		ADJA	NN
ZH1	**Computerspiele**		,	Online-Chatting	**ein**	soziales	Netzwerk
Npos	**NN**		$(	NN	**ART**	ADJA	NN
Diff	**MERGE**				**INS**		

Tabelle 9: (Kobalt_CMN_006_2011_03) tok: Lernertext, pos: automatische POS-Annotation des Lernertextes, ZH1: grammatische Zielhypothese, Npos: Goldtags auf der ZH1, Diff: Annotation der Abweichungen der ZH1 von tok. INS: im Lernertext nicht enthaltenes Token, MERGE: im Lernertext fehlerhaft zweigeteiltes Token

Während nicht vorhandene Tags (*ein*) für das Tagset kein Problem darstellen, sind Fälle von Zusammenschreibungen wie in Tabelle 9 deshalb problematisch, weil sie in der Kombination distributionell nicht an dieser **NN**-Positionen auftauchen können. Durch den Ausschluss dieser Fälle wurde sicherlich spannende Phänomene der Lernersprache für die Diskussion des STTS ignoriert, die hier vorgestellte Methode auf diese Fälle anzupassen, ist daher eine Aufgabe zukünftiger Forschung.

7 Zusammenfassung

In diesem Artikel haben wir gezeigt, dass das STTS in seiner derzeitigen Form und Verwendung (ein Token – ein Tag) eine Reihe sprachlicher Strukturen, die für Texte fortgeschrittener Deutschlerner typisch sind, nicht erfolgreich abbilden kann. Für die Aufdeckung problematischer Bereiche haben wir kontrastive Konfusionsmatrizen verwendet. Wir haben diskutiert, dass in bestimmten Anwendungskontexten sowohl Portemanteau als auch unterspezifierte Tags zu interessanten Verbesserungen führen

können. Einige lernersprachliche Phänomene lassen sich nicht über 1:1-Beziehungen zwischen Token im Lernertext und der Normalisierung darstellen. Diese Phänomene sollen aber im Zentrum zukünftiger Forschung stehen.

Danksagung

Wir danken den drei anonymen Gutachtern ganz herzlich für ihre konstruktiven Kommentare. Den anderen Mitgliedern des Netzwerks Kobalt-DaF möchten wir ebenfalls danken, da ohne sie unsere Datengrundlage, das Kobalt-Korpus, nicht existieren würde. Felix Golcher hat uns wiederholt bei der Erstellung von R-Skripts geholfen und für uns die Grafik programmiert. Ihm gilt unser ganz besonderer Dank.

Literatur

Breckle, M. und Zinsmeister, H. (2012). A corpus-based contrastive analysis of local coherence in L1 and L2 German. In: Karabalić, V., Varga, M. und Pon, L. (Hgg.), *Discourse and Dialogue / Diskurs- und Dialog*, Seiten 235–250. Peter Lang Verlag, Frankfurt am Main [u.a.].

Chrupala, G., Dinu, G. und van Genabith, J. (2008). Learning morphology with Morfette. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Marrakesch, Marokko.

Corder, S. P. (Hg.) (1986). *Error Analysis and Interlanguage*. Oxford University Press, Oxford, 4. Auflage.

Díaz-Negrillo, A., Meurers, W., Valera, S. und Wunsch, H. (2010). Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154. Special Issue on Corpus Linguistics for Teaching and Learning, in Honour of John Sinclair.

Dickinson, M. und Ragheb, M. (2009). Dependency Annotation for Learner Corpora. In: Passarotti, M., Przepiórkowski, A., Raynaud, S. und van Eynde, F. (Hgg.), *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories*, Seiten 59–70, Mailand, Italien.

Dickinson, M. und Ragheb, M. (2013). Annotation for Learner English Guidelines: v. 0.1. Technischer Bericht, Indiana University, Bloomington, IN.

Eisenberg, P. (2004). *Das Wort: Grundriß der deutschen Grammatik*. Metzler, Stuttgart [u.a.].

Giesbrecht, E. und Evert, S. (2009). Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In: Alegria, I., Leturia, I. und Sharoff, S. (Hgg.), *Proceedings of the 5th Web as Corpus Workshop (WAC5)*.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In: Aijmer, K. (Hg.), *Languages in contrast*, Band 88 von *Lund studies in English*, Seiten 37–51. Lund University Press [u.a.], Lund.

Granger, S. (Hg.) (1998). *Learner English on Computer*. Studies in language and linguistics. Longman Publishers, London [u.a.].

Granger, S. (2008). Learner Corpora. In: Lüdeling, A. und Kytö, M. (Hgg.), *Corpus linguistics*, Band 1 von *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science*, Seiten 259–275. Mouton de Gruyter, Berlin und New York.

Granger, S. und Tyson, S. (1996). Connector usage in the English Essay Writing of Native and Non-native EFL Speakers of English. *World Englishes*, 15(1):17–27.

Helbig, G. und Buscha, J. (2007). *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Langenscheidt, Berlin [u.a.].

Hennig, M. und Bücker, J. (2008). Grammatik der gesprochenen Sprache in Theorie und Praxis. *Deutsch als Fremdsprache*, 45(2):115–116.

Hirschmann, H. (2013). *Modifikatoren im Deutschen*. Doktorarbeit, Humboldt-Universität zu Berlin, Berlin.

Hirschmann, H., Lüdeling, A., Rehbein, I., Reznicek, M. und Zeldes, A. (2013). Underuse of Syntactic Categories in Falko: A Case Study on Modification. In: Granger, S., Gilquin, G. und Meunier, F. (Hgg.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*, Seiten 223–234, Louvain-la-Neuve. Presses universitaires de Louvain.

Jurafsky, D. S. und Martin, J. H. (2009). *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Pearson Education Internat., Upper Saddle River, NJ, 2. Auflage.

Lafferty, J., McCallum, A. und Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *International Conference on Machine Learning (ICML)*.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K. und Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 45(2):67–73.

Rapp, R. (2007). Part-of-Speech Discovery by Clustering Contextual Features. In: Decker, R. und Lenz, H.-J. (Hgg.), *Advances in Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, Seiten 627–634. Springer, Berlin und Heidelberg.

Rehbein, I., Hirschmann, H., Lüdeling, A. und Reznicek, M. (2012). Better Tags Give Better Trees or do they? In: *Proceedings of Treebanks and Linguistic Theory (TLT-10)*.

Reznicek, M., Lüdeling, A. und Hirschmann, H. (2013). Competing target hypotheses in the Falko Corpus: A flexible multi-layer corpus architecture. In: Ballier, N., Díaz Negrillo, A. und Thompson, P. (Hgg.), *Automatic Treatment and Analysis of Learner Corpus Data*, Band 59 von *Studies in Corpus Linguistics*, Seiten 101–124.

Schiller, A., Teufel, S., Stöckert, C. und Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technischer Bericht, Universitäten Stuttgart und Tübingen, `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf`.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, Großbritannien.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In: *Proceedings of the ACL SIGDAT-Workshop*.

Schmid, H., Fitschen, A. und Heid, U. (2004). SMOR:A German computational morphology covering derivation, composition and inflection. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*.

Schmid, H. und Laws, F. (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Seiten 777–784, Manchester, Großbritannien.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(3):209–231.

Toutanova, K. und Manning, C. D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In: *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Seiten 63–70, Hong Kong, China.

van Halteren, H., Daelemans, W. und Zavrel, J. (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics*, 27(2):199–229.

van Rooy, B. und Schäfer, L. (2002). The Effect of Learner Errors on POS Tag Errors during Automatic POS Tagging. *Southern African Linguistics & Applied Language Studies*, 20(4):325–335.