32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, Postersession der Sektion Computerlinguistik 25. Februar 2011

Hagen Hirschmann, Marc Reznicek, Anke Lüdeling, Ines Rehbein

Das Ziel ist der Weg
Geparste Zielhypothesen im Lernerkorpus Falko

Arts & Humanities Research Council
WHiG?
Falko
DFG

Fragestellung: Welche syntaktischen Strukturen sind auch für fortgeschrittene Fremdsprachenlerner schwierig zu erwerben?
Vergleich von Lernertexten und Muttersprachlertexten (CIA-Untersuchung: Granger 2002) aus dem Falko Essay Korpus (Lüdeling et al. 2008)

Methoden:

Vergleich lokaler syntaktischer Muster

→ relative Frequenzen von Wortartenketten

Bigramme	L1	L2
Pronomen-Pronomen	0.0052	0.0079
Adverb-Adverb	**0.0128**	**0.0061**
Adverb-Präposition	0.0091	0.0053

Vergleich der Frequenzen von Wortartenabfolgen (Pronomen-Pronomen, Adverb-Adverb, Adverb-Präposition) zwischen Muttersprachlern (L1) und Lernern (L2). Rot bedeutet Overuse (Übergebrauch) der jeweiligen Kategorie, blau bedeutet Underuse (Mindergebrauch) (Zeldes 2009).

Vergleich komplexer syntaktischer Beziehungen

→ relative Frequenzen von Konstituenten, Satzfunktionen und Dependenzen

verlangt syntaktisch tief annotierte Lernerdaten

Aufgabe:

Parsing von Lernerdaten (Baumbank)

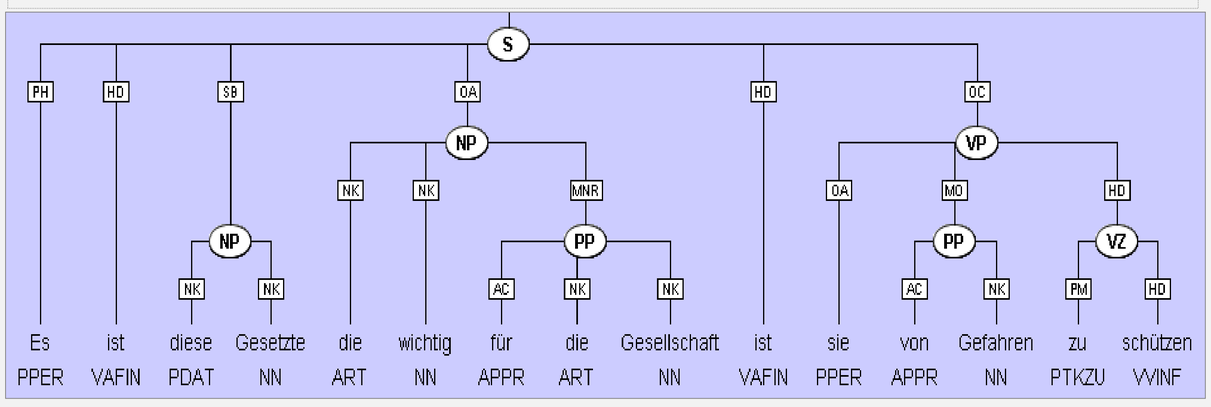
Berkeley Parser (Petrov et al. 2006)
- statistischer Parser
- auf Zeitungstexten trainiert
  - Konstituenten
  - Dependenzen
  - Satzfunktionen

Tiger-Schema (Hybrides Modell)(Albers et al. 2003)
- Erlaubt die gleichzeitige Darstellung von Konstituenten, Dependenzen, Satzfunktionen in einem Baum.

Problem:

Nicht-kanonische Strukturen (Hirschmann et al. 2007)
- Lerneräußerungen enthalten Strukturen, die weder durch das Tiger-Schema noch durch die "Berkeley-Parser-Grammatik" abgebildet werden können.

Exemplarischer Falko-Textauszug:
Um in der Gesellschaft akzeptiert zu werden , muss man bereit sein die Werte der Gesellschaft zu entsprechen . Es ist diese Gesetzte die wichtig für die Gesellschaft ist sie von Gefahren zu schützen . Gefahren bedeutet alles was die Gesellschaft drohen , auch alles was fremd ist .
(FalkoEssayv2:sa009_2006_09)



Originale Lerneräußerung geparst mit Berkeley Parser

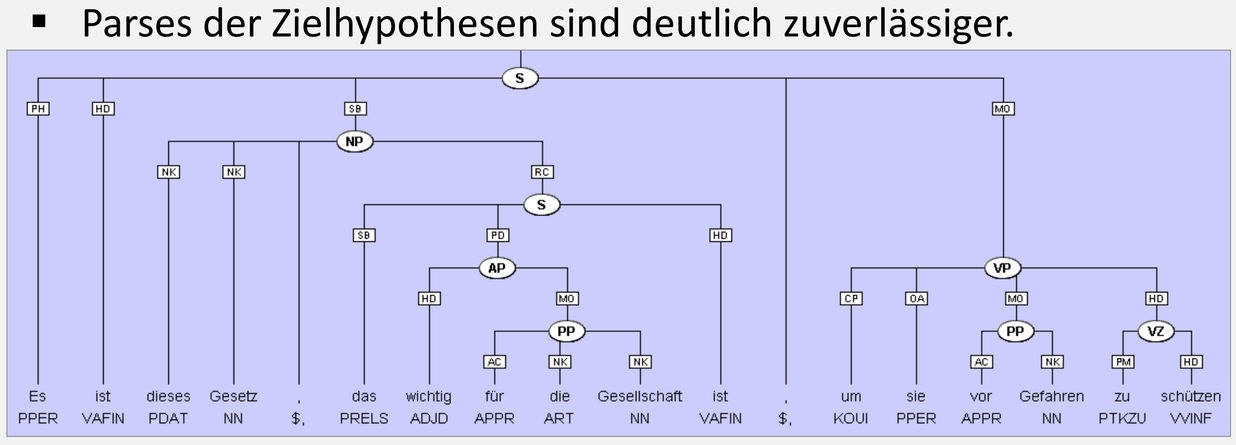
Für die Beschreibung des Beispielsatzes kann der Parser keine geeignete Lösung finden, vor allem weil die normgerechte Kommasetzung als Indikator für Satzgrenzen unabdingbar ist.

Lösung:

Zielhypothesen erstellen
- Jede nicht-kanonische Äußerung im Falko-Essay Korpus wird durch eine Zielstruktur ergänzt, die …
  - minimal vom Originaltext abweicht
  - eine kanonische Struktur erzeugt (Reznicek et al. 2010)

LT	Es	ist	diese	Gesetze		die	wichtig	für	die	Gesellschaft	ist			sie	von	Gefahren	zu	schützen	.
ZH	Es	ist	dieses	Gesetz	,	das	wichtig	für	die	Gesellschaft	ist	,	um	sie	vor	Gefahren	zu	schützen	.
Diff				CHA	INS	CHA						INS							

Zielhypothesen parsen
- Parses der Zielhypothesen sind deutlich zuverlässiger.



Kanonische Entsprechung in Zielhypothese I geparst mit Berkeley Parser

Ergebnis:

Extraktion von Satzfunktionsfrequenzen (Hirschmann et al. 2007).

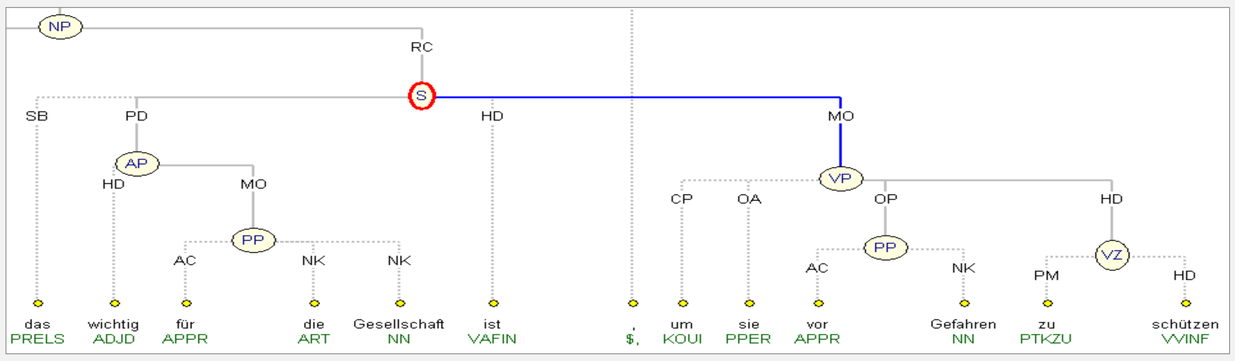
modifizierende Phrasen	L1	L2
Adverbenphrasen	**132,66**	97,417
Präpositionalphrasen	**57,974**	42,955
Adjektivphrasen	**21,231**	10,843

Vergleich der Frequenzen von modifizierenden Phrasen zwischen Muttersprachlern (L1) und Lernern (L2). Vortragsfolien Hirschmann 2011

→ Die Lernertexte im Falko-Essaykorpus weisen einen signifikanten Underuse von Modifikatoren auf, unabhängig von der konkreten syntaktischen Realisierung.

Ausblick:

Manuelle Korrektur der automatisch erstellten Syntaxbäume
- Tred (http://ufal.mff.cuni.cz/~pajas/tred/)



Einbindung der syntaktischen Bäume in RelAnnis

→ Metadaten, Spannenannotationen (z. B. Fehlerannotationen) und Syntaxbäume verbinden

→ Suche im Korpussuchtool ANNIS (Zeldes et al. 2009)

Literatur:

Albert, S.; Anderssen, J.; Bader, R.; Becker, S.; Bracht, T.; Brants, S.; Brants, T.; Demberg V.; Dipper, S.; Eisenberg, P.; Hansen, S.; Hirschmann, H.; Janitzek, J.; Kirstein, C.; Langner, R.; Michelbacher, L.; Plaehn, O.; Preis, C.; Pußel, M.; Rower, M.; Schrader, B.; Schwartz, A.; Smith, G.; Uszkoreit, H. (2003). TIGER-Annotationsschema. Technical Report. Universität Potsdam, Universität des Saarlandes, Universität Stuttgart. http://www.ifi.uzh.ch/cl/volk/treebank_course/tiger_annot.pdf. letzter Zugriff am 21.02.2011.

Granger, S. (2002). A bird's-eye view of learner corpus research. In: Granger;Hung/Petch-Tyson 2002, 3-33.

Hirschmann, H.; Doolittle, S. & Lüdeling, A. (2007) Syntactic annotation of non-canonical linguistic structures. In: Proceedings of Corpus Linguistics 2007, Birmingham.

Lüdeling, A.;Doolittle, S.;Hirschmann, H.;Schmidt, K.;Walter, M. (2008). Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 2/2008

Petrov S.; Barrett L.; Thibaux R.;Klein D. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, July 2006. Sydney, Australia

Reznicek, M.; Walter, M.; Schmid, K.; Lüdeling, A.; Hirschmann, H.; Krummes, C. (2010): Das Falko-Handbuch. Korpusaufbau und Annotationen Version 1.0

Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian (2009), "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". In: *Proceedings of Corpus Linguistics 2009*, July 20-23, Liverpool, UK.

Falko Projektseite: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/standardseite/