Proseminar Linguistische Annotation

Ines Rehbein und Josef Ruppenhofer

SS 2010

Seminarplan

- I. Linguistische Annotation - Überblick
  - Welche Arten von linguistischer Annotation gibt es?
  - Wozu sind sie gut?
- II. Der Annotationsprozess
  - Wie wird annotiert?
  - Welche Probleme treten dabei auf?
  - Welche Faktoren können die Annotation beeinflussen?
- III. Wie gut sind unsere Annotationen?
- IV. Wie bekomme ich größere Mengen an annotierten Daten?

Linguistische Annotation

- Hinzufügen von linguistischer Information zu einem Korpus
  - phonetische Annotation (SAM-PA, BITS Sprachsynthesekorpora)
  - Intonation / prosodische Annotation (ToBI/GToBI)
  - Wortarten-Annotation (POS-Tagging)
  - Morpho-Syntax
  - Syntax (Baumbanken)
  - Word Senses (WordNet)
  - Semantische Rollen (Propbank, Framenet, SALSA)
  - Named Entities (Person, Organisation, Datum, ...)
  - Temporale Annotation (TimeBank)
  - Anaphor/Coreference Annotation (TüBa-D/Z, PoCos)
  - Diskurs (Penn Discourse Treebank, Chinese Discourse Treebank)
  - Sentiment-Annotation
  - Meta-Information (Alter, Herkunft, Geschlecht, ...)
  - ...

Linguistische Annotationen - Beispiele

- Text

Er tritt in die GM-Verwaltung ein und wird Großaktionär des Autokonzerns .

Linguistische Annotationen - Beispiele

- Text + Lemmatisierung

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	

Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging)

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	
PPER	VVFIN	APPR	ART	NN	PTKVZ	KON	VAFIN	NN	ART	NN	$.

Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging) +
- morphologische Information

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	
PPER	VVFIN	APPR	ART	NN	PTKVZ	KON	VAFIN	NN	ART	NN	$.
3.Nom.Sg.Masc	3.Sg.Pres.Ind		Acc.Sg.Fem	Acc.Sg.Fem			3.Sg.Pres.Ind	Nom.Sg.Masc	Gen.Sg.Masc	Gen.Sg.Masc	

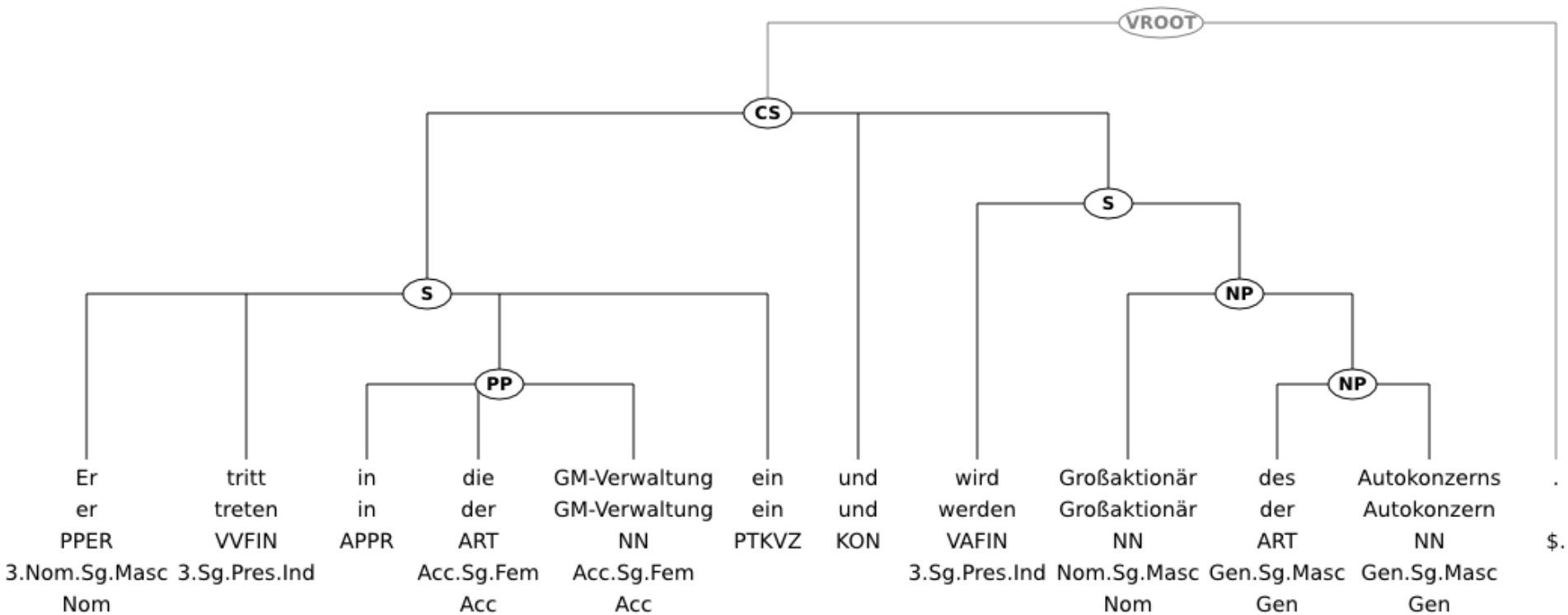
Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging) +
- morphologische Information + Kasus

Er	tritt	in	die	GM-Verwaltung	ein	und	wird	Großaktionär	des	Autokonzerns	.
er	treten	in	der	GM-Verwaltung	ein	und	werden	Großaktionär	der	Autokonzern	
PPER	VVFIN	APPR	ART	NN	PTKVZ	KON	VAFIN	NN	ART	NN	$.
3.Nom.Sg.Masc	3.Sg.Pres.Ind		Acc.Sg.Fem	Acc.Sg.Fem			3.Sg.Pres.Ind	Nom.Sg.Masc	Gen.Sg.Masc	Gen.Sg.Masc	
Nom			Acc	Acc				Nom	Gen	Gen	

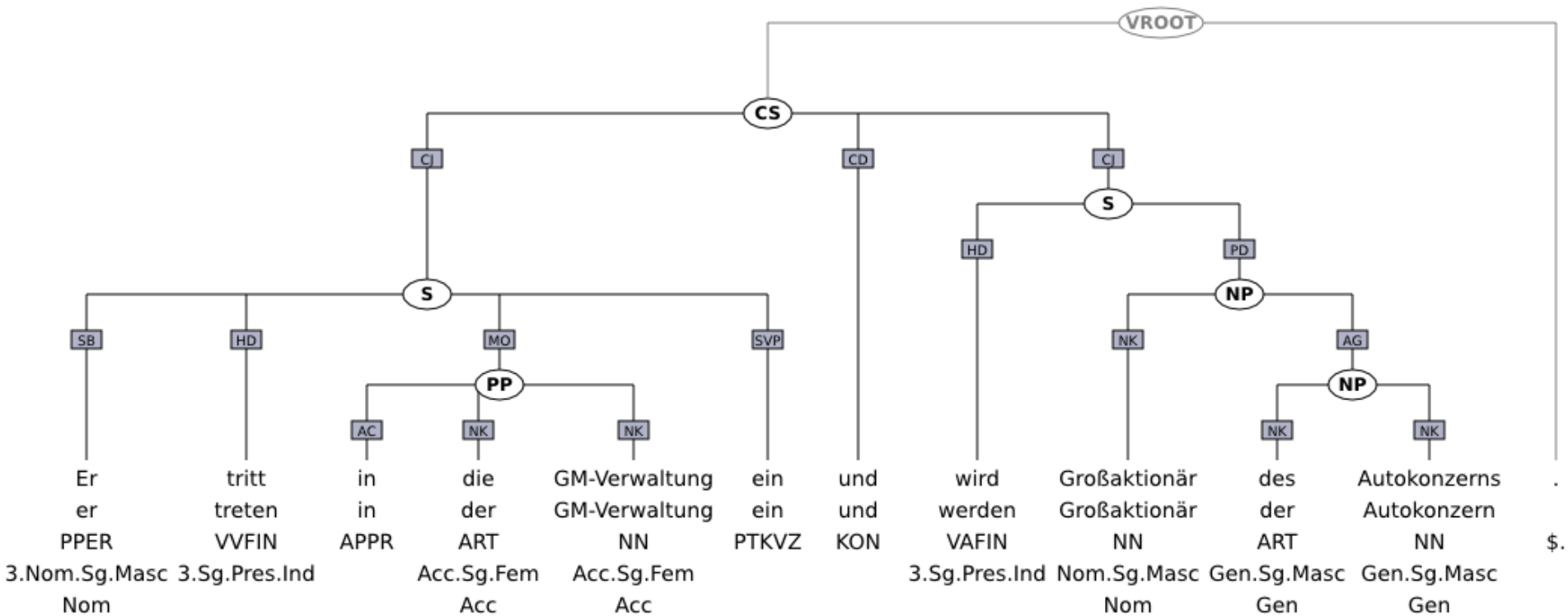
Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging) +
- morphologische Information + Kasus + Syntax



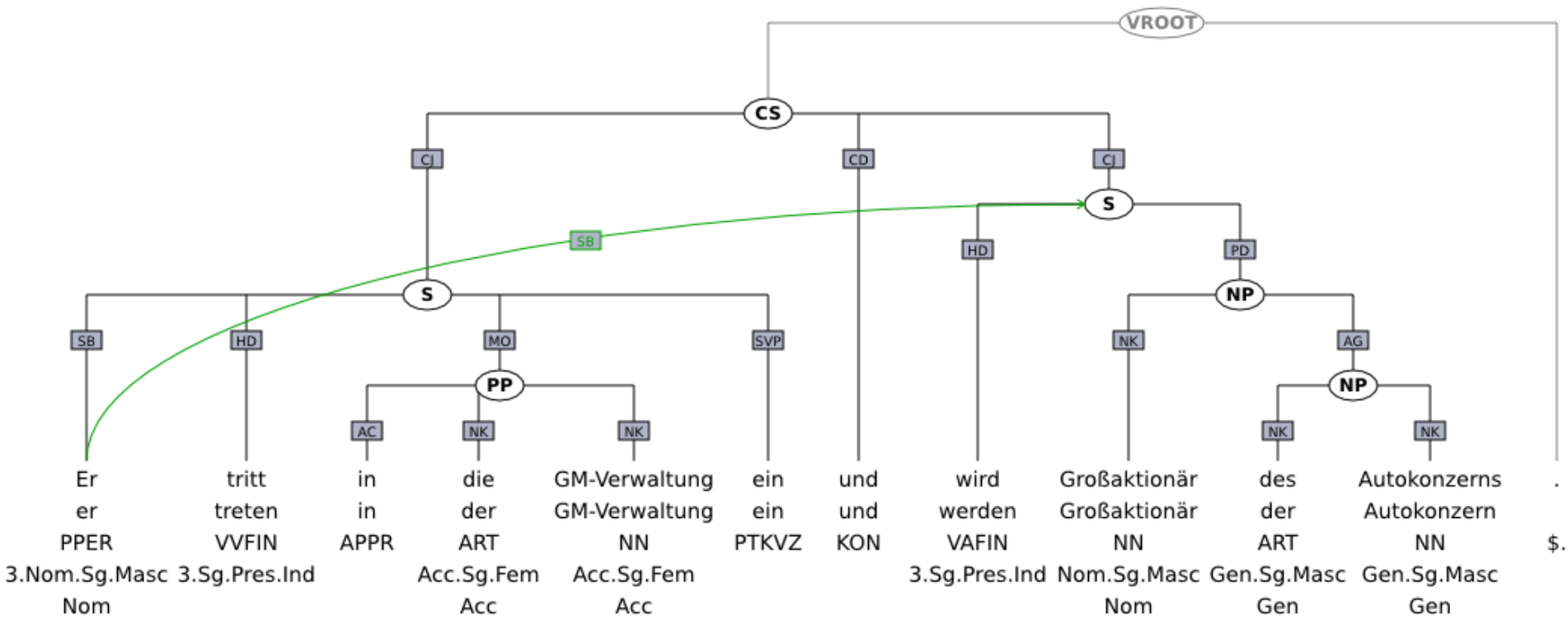
Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging) +
- morphologische Information + Kasus + Syntax +
- Grammatikalische Funktionen (GF)



Linguistische Annotationen - Beispiele

- Text + Lemmatisierung +
- Part-of-speech (POS) (Wortarten-Tagging) +
- morphologische Information + Kasus + Syntax +
- Grammatikalische Funktionen (GF) + sekundäre Kanten



Linguistische Annotationen - Beispiele

- Wozu das Ganze?
  - Lemmatisierung
    - ★
  - Part-of-speech (POS) (Wortarten-Tagging)
    - ★
  - morphologische Information
    - ★
  - Syntax
    - ★
  - Grammatikalische Funktionen (GF)
    - ★
  - sekundäre Kanten
    - ★

Linguistische Annotationen - Beispiele

- Lemmatisierung
  - Zurückführung von flektierten Wortformen auf ihre Grundform - ermöglicht Nachschlagen im Lexikon
  - Problem: Mehrdeutigkeiten (Rohrohrzucker - Roh rohr zucker - Rohr ohr zucker)
- Part-of-speech (POS) (Wortarten-Tagging)
  - Voraussetzung für syntaktische Analyse
  - hilft bei Information Extraction (und bei vielen anderen NLP tasks)
- morphologische Information
  - Bedeutungsdisambiguierung:

    (1) Die Vortragenden geben der Studentin das Buch.
    NOM PL PL DAT SG ACC

    (2) Den Vortragenden gibt die Studentin das Buch.
    DAT PL SG NOM SG ACC

- Syntax
  - Suche nach bestimmten syntaktischen Strukturen
  - Trainingsdaten für statistische Parser
- Grammatikalische Funktionen (GF)
  - Disambiguierung: Mann/SB beißt Hund/OA vs. Mann/OA beißt Hund/SB
- sekundäre Kanten
  - vollständige semantische Interpretation einer Äußerung, Identifikation fehlender Subjekte etc.

Linguistische Annotationen - Beispiele

- Lemmatisierung
  - Zurückführung von flektierten Wortformen auf ihre Grundform - ermöglicht Nachschlagen im Lexikon
  - Problem: Mehrdeutigkeiten (Rohrohrzucker - Roh rohr zucker - Rohr ohr zucker)
- Part-of-speech (POS) (Wortarten-Tagging)
  - Voraussetzung für syntaktische Analyse
  - hilft bei Information Extraction (und bei vielen anderen NLP tasks)
- morphologische Information
  - Bedeutungsdisambiguierung:

    (3) Die Vortragenden geben der Studentin das Buch.
    NOM PL PL DAT SG ACC

    (4) Den Vortragenden gibt die Studentin das Buch.
    DAT PL SG NOM SG ACC
- Syntax
  - Suche nach bestimmten syntaktischen Strukturen
  - Trainingsdaten für statistische Parser
- Grammatikalische Funktionen (GF)
  - Disambiguierung: Mann/SB beißt Hund/OA vs. Mann/OA beißt Hund/SB
- sekundäre Kanten
  - vollständige semantische Interpretation einer Äußerung, Identifikation fehlender Subjekte etc.

Linguistische Annotationen - Beispiele

- Lemmatisierung
  - Zurückführung von flektierten Wortformen auf ihre Grundform - ermöglicht Nachschlagen im Lexikon
  - Problem: Mehrdeutigkeiten (Rohrohrzucker - Roh rohr zucker - Rohr ohr zucker)
- Part-of-speech (POS) (Wortarten-Tagging)
  - Voraussetzung für syntaktische Analyse
  - hilft bei Information Extraction (und bei vielen anderen NLP tasks)
- morphologische Information
  - Bedeutungsdisambiguierung:

(5) Die Vortragenden geben der Studentin das Buch.
NOM PL PL DAT SG ACC

(6) Den Vortragenden gibt die Studentin das Buch.
DAT PL SG NOM SG ACC

- Syntax
  - Suche nach bestimmten syntaktischen Strukturen
  - Trainingsdaten für statistische Parser
- Grammatikalische Funktionen (GF)
  - Disambiguierung: Mann/SB beißt Hund/OA vs. Mann/OA beißt Hund/SB
- sekundäre Kanten
  - vollständige semantische Interpretation einer Äußerung, Identifikation fehlender Subjekte etc.

Linguistische Annotationen - Beispiele

- Lemmatisierung
  - Zurückführung von flektierten Wortformen auf ihre Grundform - ermöglicht Nachschlagen im Lexikon
  - Problem: Mehrdeutigkeiten (Rohrohrzucker - Roh rohr zucker - Rohr ohr zucker)
- Part-of-speech (POS) (Wortarten-Tagging)
  - Voraussetzung für syntaktische Analyse
  - hilft bei Information Extraction (und bei vielen anderen NLP tasks)
- morphologische Information
  - Bedeutungsdisambiguierung:

    (7) Die Vortragenden geben der Studentin das Buch.
    NOM PL PL DAT SG ACC

    (8) Den Vortragenden gibt die Studentin das Buch.
    DAT PL SG NOM SG ACC

- Syntax
  - Suche nach bestimmten syntaktischen Strukturen
  - Trainingsdaten für statistische Parser
- Grammatikalische Funktionen (GF)
  - Disambiguierung: Mann/SB beißt Hund/OA vs. Mann/OA beißt Hund/SB
- sekundäre Kanten
  - vollständige semantische Interpretation einer Äußerung, Identifikation fehlender Subjekte etc.

Linguistische Annotationen - Beispiele

- Lemmatisierung
  - Zurückführung von flektierten Wortformen auf ihre Grundform - ermöglicht Nachschlagen im Lexikon
  - Problem: Mehrdeutigkeiten (Rohrohrzucker - Roh rohr zucker - Rohr ohr zucker)
- Part-of-speech (POS) (Wortarten-Tagging)
  - Voraussetzung für syntaktische Analyse
  - hilft bei Information Extraction (und bei vielen anderen NLP tasks)
- morphologische Information
  - Bedeutungsdisambiguierung:

    (9) Die Vortragenden geben der Studentin das Buch.
    NOM PL PL DAT SG ACC

    (10) Den Vortragenden gibt die Studentin das Buch.
    DAT PL SG NOM SG ACC

- Syntax
  - Suche nach bestimmten syntaktischen Strukturen
  - Trainingsdaten für statistische Parser
- Grammatikalische Funktionen (GF)
  - Disambiguierung: Mann/SB beißt Hund/OA vs. Mann/OA beißt Hund/SB
- sekundäre Kanten
  - vollständige semantische Interpretation einer Äußerung, Identifikation fehlender Subjekte etc.

Linguistische Annotationen - Beispiele

- Lemmatisierung
  - Zurückführung von flektierten Wortformen auf ihre Grundform - ermöglicht Nachschlagen im Lexikon
  - Problem: Mehrdeutigkeiten (Rohrohrzucker - Roh rohr zucker - Rohr ohr zucker)
- Part-of-speech (POS) (Wortarten-Tagging)
  - Voraussetzung für syntaktische Analyse
  - hilft bei Information Extraction (und bei vielen anderen NLP tasks)
- morphologische Information
  - Bedeutungsdisambiguierung:

    (11) Die Vortragenden geben der Studentin das Buch.
    NOM PL PL DAT SG ACC

    (12) Den Vortragenden gibt die Studentin das Buch.
    DAT PL SG NOM SG ACC

- Syntax
  - Suche nach bestimmten syntaktischen Strukturen
  - Trainingsdaten für statistische Parser
- Grammatikalische Funktionen (GF)
  - Disambiguierung: Mann/SB beißt Hund/OA vs. Mann/OA beißt Hund/SB
- sekundäre Kanten
  - vollständige semantische Interpretation einer Äußerung, Identifikation fehlender Subjekte etc.

Seminarplan

- I. Linguistische Annotation - Wozu?
  - mehr Information (erhöht die Interpretierbarkeit eines Korpus)
  - Untersuchung linguistischer Phänomene
  - Überprüfung linguistischer Theorien
    - ★ viele linguistische Theorien entstehen aufgrund von Introspektion
      $\longrightarrow$ Armchair linguistics
    - ★ aber manchmal übersieht man was...
    - ★ Überprüfung linguistischer Theorien mit Hilfe von realistischen Daten
      Läßt sich meine Theorie anhand der Daten widerlegen?

Beispiel I: Partikelverben (Müller & Meurers, 2006)

- *Theorie:* Verbpartikeln können nicht vorangestellt werden
  (Ausnahme: prädikative Partikeln wie *auf* in *aufmachen*)

- *Korpusevidenz:*
  Los$_{PART}$ **ging** es schon in dieser Woche. (taz, 11.10.1995)
  Vor$_{PART}$ **hat** er das jedenfalls. (taz, 15.07.1999)

Beispiel II: Idiome (Geyken et al., 2004)

- *Theorie:* klassische Ansätze betonen die Invariabilität von Idiomen (Katz, 1973; Chomsky, 1980)
- *Korpusevidenz:* **ein Blatt vor den Mund nehmen**
  - Pluralisierung:
    - ohne Blätter vor den Mund zu nehmen
  - Quantifizierung:
    - Hier nahm er manches Blatt vor den Mund
    - der sich 100 Blätter vor den Mund nimmt
  - Adjektivische Modifikation eines oder beider Nomen:
    - mit einem postmodernen Blatt vor dem Munde
    - kein Blatt vor seinen republikfeindlichen Mund
  - Nomen-Modifikation:
    - ohne das geringste (Klee-)Blatt vor den vorlauten Mund zu nehmen

Seminarplan

- I. Linguistische Annotation - Wozu?
  - mehr Information (erhöht die Interpretierbarkeit eines Korpus)
  - Untersuchung linguistischer Phänomene
  - Überprüfung linguistischer Theorien
  - Ressourcen zum Training von statistischen NLP-Systemen:
    - Wortarten-Tagger
    - Syntaktische Parser
    - Semantische Parser / Labelling von Semantischen Rollen
    - Systeme zur Lesarten-Disambiguierung
    - Anaphern-Auflösung
    - Maschinelles Übersetzen
    - Automatische Spracherkennung
    - ...
  - Linguistisch annotierte Daten zur Evaluation von NLP-Systemen (Goldstandard)

Linguistische Annotation

- erhöht die Interpretierbarkeit eines Korpus
- zeitaufwändig!

1. Standardisierung
   - Standards erhöhen Konsistenz und Verwendungsbreite
   - ermöglichen den Austausch von Daten
     - EAGLES (Expert Advisory Group on Language Engineering Standards)
     - TEI (Text Encoding Initiative)
     - GrAF (A Graph-based Format for Linguistic Annotations)
     - ...
2. Interoperabilität
   - z.B. die Übertragung von vorhandenen Annotationsschemata auf neue Sprachen ▸ Penn Chinese Treebank, ▸ Penn Arabic Treebank
   - oder die Kombination verschiedener Annotationsebenen in eine vereinte Repräsentation (z.B. Propbank + Nombank + TimeBank + Penn Discourse treebank + Coreference) ▸ XBank Browser

- Aber: bevor man Annotationsschemata vereint oder überträgt
  - Was sind die Vor- und Nachteile verschiedener Annotationsschemata?
  - Wie vergleicht man Annotationsschemata?

Linguistische Annotation

- erhöht die Interpretierbarkeit eines Korpus
- zeitaufwändig!

1. Standardisierung
   - Standards erhöhen Konsistenz und Verwendungsbreite
   - ermöglichen den Austausch von Daten
     - ★ EAGLES (Expert Advisory Group on Language Engineering Standards)
     - ★ TEI (Text Encoding Initiative)
     - ★ GrAF (A Graph-based Format for Linguistic Annotations)
     - ★ ...
2. Interoperabilität
   - z.B. die Übertragung von vorhandenen Annotationsschemata auf neue Sprachen ▸ Penn Chinese Treebank, ▸ Penn Arabic Treebank
   - oder die Kombination verschiedener Annotationsebenen in eine vereinte Repräsentation (z.B. Propbank + Nombank + TimeBank + Penn Discourse treebank + Coreference) ▸ XBank Browser

- Aber: bevor man Annotationsschemata vereint oder überträgt
  - Was sind die Vor- und Nachteile verschiedener Annotationsschemata?
  - Wie vergleicht man Annotationsschemata?

Linguistische Annotation

- erhöht die Interpretierbarkeit eines Korpus
- zeitaufwändig!

1. Standardisierung
   - Standards erhöhen Konsistenz und Verwendungsbreite
   - ermöglichen den Austausch von Daten
     - ★ EAGLES (Expert Advisory Group on Language Engineering Standards)
     - ★ TEI (Text Encoding Initiative)
     - ★ GrAF (A Graph-based Format for Linguistic Annotations)
     - ★ ...
2. Interoperabilität
   - z.B. die Übertragung von vorhandenen Annotationsschemata auf neue Sprachen ▸ Penn Chinese Treebank, ▸ Penn Arabic Treebank
   - oder die Kombination verschiedener Annotationsebenen in eine vereinte Repräsentation (z.B. Propbank + Nombank + TimeBank + Penn Discourse treebank + Coreference) ▸ XBank Browser

- Aber: bevor man Annotationsschemata vereint oder überträgt
  - Was sind die Vor- und Nachteile verschiedener Annotationsschemata?
  - Wie vergleicht man Annotationsschemata?

Linguistische Annotation

- erhöht die Interpretierbarkeit eines Korpus
- zeitaufwändig!

1. Standardisierung
   - Standards erhöhen Konsistenz und Verwendungsbreite
   - ermöglichen den Austausch von Daten
     - ★ EAGLES (Expert Advisory Group on Language Engineering Standards)
     - ★ TEI (Text Encoding Initiative)
     - ★ GrAF (A Graph-based Format for Linguistic Annotations)
     - ★ ...
2. Interoperabilität
   - z.B. die Übertragung von vorhandenen Annotationsschemata auf neue Sprachen ▸ Penn Chinese Treebank, ▸ Penn Arabic Treebank
   - oder die Kombination verschiedener Annotationsebenen in eine vereinte Repräsentation (z.B. Propbank + Nombank + TimeBank + Penn Discourse treebank + Coreference) ▸ XBank Browser

- Aber: bevor man Annotationsschemata vereint oder überträgt
  - Was sind die Vor- und Nachteile verschiedener Annotationsschemata?
  - Wie vergleicht man Annotationsschemata?

Seminarplan

- II. Der Annotationsprozess
  - Wie wird annotiert?
    - ★ Erstellung von Annotationsrichtlinien
    - ★ Training
    - ★ Annotationsprozess
    - ★ Qualitätssicherung
  - Welche Probleme treten dabei auf?
  - Welche Faktoren können die Annotation beeinflussen?
    - ★ Annotations-Tools
    - ★ Richtlinien
    - ★ persönliche Eignung und Neigung der Annotator/innen

Seminarplan

- III. Evaluation - Wie gut sind unsere Annotationen?
  - Evaluation gegen einen manuell annotierten Goldstandard
  - Inter-Annotator Agreement
  - Einsatz der Daten als Trainingsset für Systeme der automatischen Sprachverarbeitung (Task-based evaluation)

Seminarplan

- IV. Wie bekomme ich große Mengen an annotierten Daten?
  - Halb-automatische Annotation
  - Bootstrapping
  - Active Learning
  - Games with a Purpose (z.B. ESP-Game)
  - kollaborativ erstellte Ressourcen wie Wikipedia
  - ...

Seminarplan

- I. Linguistische Annotation - Überblick
  - Welche Arten von linguistischer Annotation gibt es?
  - Wozu sind sie gut?
- II. Der Annotationsprozess
  - Wie wird annotiert?
  - Welche Probleme treten dabei auf?
  - Welche Faktoren können die Annotation beeinflussen?
- III. Wie gut sind unsere Annotationen?
- IV. Wie bekomme ich größere Mengen an annotierten Daten?

- Leistungsnachweis:
  - 5 Leistungspunkte
  - Schein für Hausarbeit + Vortrag
  - Beteiligung an kleinen praktischen Übungen