Deutsche Lernerwortarten im Falko Lernerkorpus

Was Mehrebenen-POS-tags leisten können

Marc Reznicek
Humboldt-Universität zu Berlin

STTS- Workshop

24.9.2012

Überblick

- STTS in Lernerkorpora
- Lernerfehler
- Zielhypothesen
- Evidenzbasierte Annotation
- POS auf mehreren Ebenen
- Vorteile
- Vorschläge für STTS-Tags

STTS in Lernerkorpora

Wortartenannotation

- heute Standard in Lernerkorpora

Lernerkorpora mit POS-Annotationen:

- Englisch: ICLE2, NOCE, …
- Deutsch: Falko, Kobalt, KanDel, Hamatac …

automatische Annotation:

- TreeTagger (Schmid 94) → STTS
- rfTagger (Schmid 2009) → TiGER-Tags (STTS gemappt)

Untersuchungen auf POS-Tag-Chains:

Borin&Prütz (2004), Zeldes et al.(2008), Hirschmann et al. (2009)

Lernerfehler

Viele ***Kriminal/NN*** *Aktivitäten* ***passiert/VVPP*** *jeden Tag in der* **Heutzutager/NN** *Gesellschaft.*

(FalkoEssayL2v2.3:kne19_2006_07)

"Wenn der Sinn erkennbar ist, wird die WF verbessert, und es wird so getaggt, wie die richtige Wortform ausgesehen hätte."

(Schiller et al. 1999:10)

Option1: Zielhypothesen taggen

*Viele **Kriminal/NN** Aktivitäten **passiert/VVPP** jeden Tag in der* **Heutzutager/NN** *Gesellschaft.*

(FalkoEssayL2v2.3:kne19_2006_07)

Minimale Zielhypothese (ZH1) (Reznicek et al. erscheint)

*Viel **kriminelle/ADJA** Aktivität **passiert/VVFIN** jeden Tag in der **heutigen/ADJA** Gesellschaft.*

Qualität:

POS-Tags für rfTagger 98.9% (Rehbein et al. 2012)

Option 2: evidenzbasierte Beschreibung

- Taggen von Lernersprache ist **kein reines *robustness*-Problem.** (Meurers 2010)
- Lernersprache sollte **evidenzbasiert annotiert** werden um Comparative Fallacy zu vermeiden . (Ragheb & Dickinson et al. 2011)
- Wortarten kombinieren Informationen über:
  - lexikalische Eigenschaften
  - Distribution im Satz
  - Morphologie
- In der Standardvarietät konvergieren diese meist.

Merkmale auf mehreren Ebenen

- **In Lernersprache können sich diese Ebenen widersprechen**

Viele Kriminal Aktivitäten passiert jeden Tag in der **Heutzutager** *Gesellschaft.*

.

Lexik: **ADV**

Distribution: **ADJA**

Morphologie: **NN/ADJA?**

POS-Konflikte in Lernersprache

- **Konflikte zwischen den Ebenen sind vielfältig**

Distribution = Morphologie ≠ Lexik

*Biologischen Verpflichtungen, die Realität der Schwangerheit schaffen nicht mehr ein **glase** Hemmung vor Frauen zu den Topjobs des Welts.*

(FalkoEssayL2v2.3:fkb033_2008_07)

Lexik: **NN**
Distribution: **ADJA**
Morphologie: **ADJA**

POS-Konflikte in Lernersprache

Lexik = Morphologie ≠ Distribution

*Es gibt doch freundliche Kriminale wie Robin Hood aber die meisten sind **Geldhunger** Männer, der noch mehr Geld und Mag haben wollen.*

(FalkoEssayL2v2.3:sa010_2006_09)

Lexik: **NN**
Distribution: **ADJA**
Morphologie: **NN**

POS-Konflikte in Lernersprache

Lexik = Distribution ≠ Morphologie

*Diese Entwicklung ist in den **skandinavien** Ländern schnell gegangen.*

(FalkoEssayL2v2.3:fkb010_2008_07)

Lexik: **ADJA**
Distribution: **ADJA**
Morphologie: **NN**

Lösungsvorschläge

1) POS-Tags auf mehreren Ebenen (Ragheb&Dickinson 2011)

	Tin	Toy	can	makes	different	music	sound
Morph:	NP1x	NP1x	VMo	VVZt	JJ	NN1u	NN1c
Dist:	NP1x	NP1x	VMo	VV0t	JJ	JJ	NN
	<>	<>	<>	<SUBJ,AUX, OBJ>	<>	<>	<>

2) Mehrebenen-POS-Tags

Heutzutager/ADVA.Sg.Fem.Dat *Gesellschaft*

→ *attributives Adverb mit Nominal-/Adjektivdeklination*

Lösungsvorschläge

kombinierte Tags

- *ein **glase** Hemmung vor Frauen* **NN-ADJA-ADJA**
- ***Geldhunger** Männer* **NN-ADJA-NN**
- er hat mehr geld **dann** die Frau **KON-KOKOM**

unterspezifizierte Tags

- ohne richtig zu wissen, wie es sich in der Wirklichkeit **abspiegelt** lässt **VV-VVFIN-VVINF →VV**
- und oft sind sie **verlasst** **VVIMP-VVPP → VV**
- In diesem Aufsatz werde ich diese Sichtweise **illustriert** **VV-VVPP-VVINF →VV**

Vorteile von Mehrebenen-POS-Tags

- Einheitliche Beschreibung von Standardvarietäten und Lernersprache (und anderen Nichtstandard-Varietäten)

- Fokus auf linguistische, evidenzbasierte Unterschiede zwischen Varietäten
- Erhöhte Unabhängigkeit von normbasierter Beschreibung (wie in Zielhypothesen)

(Díaz-Negrillo u. a. 2010)

Literatur

Borin, Lars; Prütz, Klas (2004): New wine in old skins? A corpus investigation of L1 syntactic t learner language. In: Guy Aston, Silvia Bernardini, Dominic Stewart (Eds.): Corpora and language learners. Amsterdam, Philadelphia: John Benjamins (Studies in corpus linguistics, 17), 67–87.

Díaz-Negrillo, Ana; Meurers, Walt Detmar; Valera, Salvador; Wunsch, Holger (2010): Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. In: *Language Forum*. http://purl.org/dm/papers/diaz-negrillo-et-al-09.html.

Hirschmann, Hagen; Zeldes, Amir; Lüdeling, Anke (2009): Interaction between Colligation, Register and Surface Variability in German Learners and Natives. Dgfs 2009, AG 6. Osnabrück, 4/03/2009.

Ragheb, Marwa; Dickinson, Markus (2011): Avoiding the Comparative Fallacy in the Annotation of Learner Corpora. In: Gisela Granena (Ed.): Selected Proceedings of the 2010 Second Language Research Forum. Somerville, MA:, 114–124.

Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999): Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical Report. University of Stuttgart; University of Tübingen.

Schmid, Helmut; Laws, Florian (2008): Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging. In: Donia Scott (Ed.): 22nd International Conference on Computational Linguistics. Coling 2008. Manchester,United Kingdom. COLING. Stroudsburg, Pa: Association for Computational Linguistics (ACL), 777–784. http://dl.acm.org/citation.cfm?id=1599081.1599179.

Zeldes, Amir; Lüdeling, Anke; Hirschmann, Hagen (2008): What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data. In: Antti Arppe, Kaius Sinnemäki, Urpo Nikanne (Eds.): Proceedings of Quantitative Investigations in Theoretical Linguistics 3 (QITL-3). Helsinki, Finnland, 74–77.