Modifikationsvorschläge zu STTS – Stand der bisherigen Diskussion

Heike Zinsmeister

SPONSORED BY THE

Federal Ministry of Education and Research

Eine kleine Chronik

- Vorläufer
  - ELWIS (Helmut Feldweg, Christine Thielen, Tübingen), Part-of-Speech-Tagset Stuttgart (Anne Schiller, Simone Teufel, Stuttgart)
- Stuttgart Tübingen Tagset STTS
  - Schiller, Teufel & Thielen 1995 (Großes Tagset 1999 mit C. Stöckert)
- 1. STTS-Workshop, 12/2004, Tübingen (Frank H. Müller)
  - Diskussion des Tagsets (Tagging von Standardsprache; Erwähnung von Dialekten); Protokoll von 03/2005.
  - 08/2010: Rundmail von Stefanie Dipper (Bochum) zur Umsetzung der Beschlüsse des 1. STTS-Workshops
    - Motivation: "Kategorisierungsprobleme bei Wortarten-Annotation von Textkorpora" (Keßelmeier & von Könemann 2010)
- 2. STTS-Workshop, 09/2012, Stuttgart (CLARIN-D)
  - Dokumentation (und Umsetzung) der Beschlüsse von 2004
  - Diskussion um Erweiterung des Tagsets für andere Sprachdomänen

CLARIN-D

Teilnehmende

- Jena: Petra Steiner (ehemals Münster)
- Prag/Wien: Karel Oliva
- Potsdam: Stefanie Dipper,
- Saarbrücken: Sandra Hauser, Annette Klinger
- Stockholm: Martin Volk, Yvonne Samuelsson
- Stuttgart: Stefan Klatt (auch Wien), Heike Zinsmeister
- Tübingen: Sandra Kübler, Lothar Lemnitzer, Frank Henrik Müller, Heike Telljohann, Tylman Ule und studentische Hilfskräfte
- Zürich: Simon Clematide

CLARIN-D

Allgemeine Beschlüsse

- Abwärtskompatibilität sicherstellen
  - Erweitungen bzw. zusätzliche Untergliederungen als Unterstrich-Ergänzung zum Tagnamen

    (1) Anna hat/VAFIN_V Zweifel. (Vollverb)

    (2) Max hat/VAFIN_P getroffen. (Perfektauxiliar)

  - Neu: Kleines / Mittleres / Großes Tagset
  - Keine Unterspezifikation im Mittleren Tagset, d.h. wenn Unterstrich-Kategorie, dann obligatorisch
- Kein eigenes Tagset für Dialekte

CLARIN-D

Streichung von PRELAT

- STTS 1999
  - (1) die Dinge, deren/PRELAT Nutzen wir erkennen
  - (2) er weiß, wessen/PWAT Auto er gerammt hat
- "AT" Linguistisch nicht motiviert, weil NP-Substitution
- Beschluss
  - (3) das Instrument, dessen/PRELS Nutzen wir erkennen
  - (4) er weiß, wessen/PWS Auto er gerammt hat
- Beachte
  - (5) das Instrument, dessen/PRELS wir uns bedienen
    - → neue Ambiguität (vgl. 3, gleicher Tag vs. Distribution)
  - (6) die Frage, welche/PWAT Aufgabe gestellt wurde
    - → PWAT bleibt erhalten

CLARIN-D

Appellativa (NE) vs. Eigennamen (NN)

- Beschluss
  - Listenbasierte Einteilung
  - Grenzenfälle, die nicht NE sind:

    Immer NN sind: Museen (auch "das Städel/NN"), Ämter (auch Universitäten), Ereignisse (Kriege, …), …
- Einzeldiskussionen
  - NE: Unter den "Firmenbegriff" fallen auch Vereine, Parteien
  - NN:
    - Bestätigung: Produktnamen ("ein Tempo/NN")
    - Eigennamen, die als Gattung verwendet werden

      (1) Jede Nation hat einen Goethe/NN

K&vK 2010: 38 schlagen weitere Tests vor.

Fremdspachliches Material (FM)

- Ist opak (für die Grammatik / den Annotator/ das Tagging-Tool)
- Diskussion

  (1) University/FM of/FM Michigan/NE

  - Michigan als FM oder NE?
  - NE, da auch im Deutschen verwendet

(siehe STTS 1999, Kap 3.12.4)

- Hinweis: Für text-to-speech-Anwendungen ist es praktisch, wenn mehr FM als NE getaggt wird.

CLARIN-D

Prädikativ oder adverbial gebrauchtes Adjektiv (ADJD)

- Feinere Klassifikation gewünscht
- Beschluss:
  - ADJD_P (prädikativ) nur mit *sein, bleiben, werden, scheinen*
  - sonst: ADJD_V (adverbial)

    auch bei *sich fühlen, schmecken, riechen, essen* usw.

    (1) sie ist groß/ADJD_P

    (2) er läuft schneller/ADJD_V

CLARIN-D

Adverbial gebrauchtes Adjektiv (ADJD) oder Adverb (ADV)

- Diskussion
  - Zusammenführung von adverbialem ADJD und ADV?
- Motivation
  - Ungleiche Koordination

    (1) er isst langsam/ADJD und oft/ADV
  - Projektion von ADJD zur AVP vs. ADJP (Tiger, nicht in TüBA 7)

    (2) [$_{AVP}$ früher/ADJD einmal]
- Beschluss: Beibehaltung der alten Trennung

K&vK 2010: 26 Experiment brachte keine Verbesserung

CLARIN-D

VVPP vs. ADJD

- Bessere Klassifikationskriterien gewünscht

  (1) Er hat die Haare kurz geschnitten/(VVPP|ADJD)

(vgl. STTS 1999, Kap. 3.2.3)

- Beschluss

  1. Nachschlagewerk: Lexikalisierung?
     - nein: VVPP
     - ja (aber möglichweise ambig): ADJD?
       2. Kann der Satz ins Aktiv gesetzt werden mit gleicher Semantik? Ja: VVPP

          Von-PP oder ähnliche PP (=Verbsemantik)? Ja: VVPP

          Semantisch ähnliches Adjektiv möglich? Ja: ADJD
       3. Sonst: ADJD

Duden Deutsches Universalwörterbuch

K&vK 2010: 31 schlagen weitere Tests vor.

Verben

- Feinere Klassifizierung von *haben, sein* und Modalverben gewünscht
- Beschluss
  - Unterstrichtags
    - V Vollverb *haben* mit NP
    - K Kopula *sein, werden*
    - P Perfekt- / Passivhilfsverb *haben, sein, werden*
    - X mit abgetrennter Verbpartikel *haben, sein*
    - I mit einfachem Infinitiv *haben, werden, wollen*
    - Z mit zu Infinitiv *haben, sein*

CLARIN-D

Übersicht der verbalen Unterstrichtags

V{A,M}	Verwendung	Unterstrichtag	Beispiel
haben	mit NP	V	er hat ein Auto
	mit zu-Infi nitiv	Z	er hat zu gehorchen
	mit Partizip II	P	
	mit einf. Inf.	I	er hat kommen wollen
	mit abgetr. VZ	X	
werden	mit e. Inf.	I	er wird ihn suchen
	mit Partizip II	P	er wird gesucht
	Kopula	K	
sein	mit zu-Infi nitiv	Z	
	mit Partizip II	P	
	Kopula	K	er ist groß
	mit abgetr. VZ	X	
wollen	mit NP	V	er will ein Eis
	mit einf. Inf.	I	

(Quelle: Protokoll, S. 3)

CLARIN-D

Artikel

- Eine feinere Klassifizierung erwünscht
- Motivation
  - schnellerer Zugriff auf Definitheitsinformation
- Beschluss
  - Unterstrichkategorien

    (1) das/ART_D

    (2) ein/ART_I
- Neue Unterscheidung (ART vs. CARD)
  - Wenn "ein " durch andere Zahlen ersetzbar → CARD

    (3) Der Euro fällt unter einen/CARD Dollar

CLARIN-D

Pronomina

- PIDAT
  - Problem: notorisch unklar
  - Diskussion: Prä- und Post-Determiner?
    (2) all/PIDAT die / die beiden/PIDAT
  - Beschluss: PIDAT wird zugunsten von PIAT gestrichen
- PIS
  - Problem: Pronomina mit Artikel
    (1) Die eine/PIS die andere/PIS
  - Beschluss: PIS hier beibehalten

CLARIN-D

Pronominaladverbien (PAV, PWAV)

1. Relativisch genutzte PAVs

   (1) eine Meinung, derzufolge …

- Beschluss
  - Neues Tag PRELAV

2. Subklassen von PWAV gewünscht

- Beschluss
  - Unterstrichkategorien

    (3) wann/PWAV_A versus wonach/PWAV_P vgl.

    (4) dann/ADV versus danach/PAV
  - Motivation: nur PWAV_Ps können Präpositionalobjekte sein

Mehrteilige Konjunktionen

- Erwünscht
  - Feinere Unterscheidung von einleitenden und nicht-einleitenden Konjunktionen
- Beschluss
  - Unterstrichkategorien

(1) weder/KON_V … noch/KON_K

(2) entweder/KON_V … oder /KON_K

CLARIN-D

Mehrwortausdrücke

- Infragestellung der Maxime Tokenbezug (vgl. STTS 1999, Kap. 2.2)
- Beschluss
  - Tokenbezug soll beibehalten werden
- Wunsch
  - Beispielsammlung für die Annotation gängiger Mehrwortausdrücke

    gang und gebe, klipp und klar, sage und schreibe, ab und zu, unter anderem, *samt und sonders*, usw.

CLARIN-D

Weitere Themen

- Tagging von "Ex-Partikelverb-Präfixen", die nach neuer Rechtschreibung getrennt geschrieben werden

---

- Wortverschmelzungen (*wie gehts*?)
- Nicht flektierende Adjektive
  - Bsp.: *In ganz Bayern* TüBa: ADV; Tiger: ADJD (sonst ADV)
- Pronominaladverbtag mit Angabe der Präposition
- Semantische Unterkategorien von NE, z.B. NE_O
- Morphologische Erweiterungen (nicht besprochen)

ab hier nicht mehr im Protokoll dokumentiert

Vorschlag zur Umsetzung

CLARIN-D

nach Martin Volk (E-Mail, 06.08.2010)

1. Die neuen Vorschläge ausführlich und mit vielen Beispielen beschreiben
2. Das Dokument im Paket mit den ursprünglichen STTS-Richtlinien anbieten
3. Eine Webseite einrichten, wo das alles gut beschrieben ist
4. (Optional aber sehr wünschenswert) ein Tool anbieten, das mit STTS annotierte Texte durchgeht und auf Präzisionsmöglichkeiten hinweist