Stand und Perspektiven des Wortarttagsets STTS: Einführung in den Workshop

Ulrich Heid, Heike Zinsmeister, Kathrin Beck

Stuttgart, 24. September 2012

Überblick

- Kontext des Workshops
- STTS: Ist-Zustand - Varianten - Soll-Zustand
  - Quellen
  - Arten von Varianten in STTS
  - Soll-Zustand: Bedarfsvermittlung
  - Beispielfall: Taggingprobleme bei Web-Daten Giesbrecht/Evert 2009
- Vorschläge zur STTS-Ergänzung: allgemeine Überlegungen
- Ziele für die Dokumentation von STTS

Kontext des Workshops

Arbeiten zur Nachhaltigkeit von Sprachressourcen

- CLARIN-D-Ziele
  schließen nachhaltige Bereitstellung von Ressourcen mit ein:
  Dokumentation - Verfügbarkeit - Integrierbarkeit von Ressourcen – Überarbeitung und Verbesserung bestehender Ressourcen
- STTS (Stuttgart-Tübingen TagSet) Schiller/Teufel/Thielen/Stöckert
  ist eine der Standard-Ressourcen für das Deutsche
- Nachhaltigkeitsziel *Dokumentation* impliziert:
  - Dokumentation des Ist-Zustands samt "Varianten"
  - Dokumentation des Soll-Zustands, z.B. möglicher Ergänzungen
  - Bereitstellung der Dokumentation,
    sowie ggf. von Beispieltexten und geeigneten Tools
    ⇒ Die CLARIN-D-Standorte Tübingen und Stuttgart
    sind in solche Nachhaltigkeitsarbeiten involviert

STTS – Ist-Zustand

Überblick – Quellen

- STTS ist ein Wortart-Tagset für das Deutsche
  - Entwickelt als logisches Tagset: Spezialisierungshierachie:
    V.∗ → VA.∗ | VM.∗ | VV.∗
    VV.∗ → VVFIN | VVINF | VVPP | VVIZU
  - Kriterien für Tag-Definitionen sind
    * distributionell: z.B. attributive vs. substituierende Pronomina
    * lexikalisch: z.B. Modalverben vs. Vollverben
    * formbezogen: z.B. attributive vs. prädikative Adjektive
- STTS - Quellen:
  - Guidelines für die Annotation
  - Tag-Liste
  - Beschreibung

STTS - "Varianten"

Überblick und Quellen – Beispiele Details: Vortrag Zinsmeister et al.

- Tiger - Annotationsrichtlinien
  - Unterschied PIDAT - PIAT entfällt: immer PIAT
  - Umbenennung von PAV: in Tiger: immer PROAV
  - Zusatzrichtlinie für ADV-Gebrauch von bestimmten Präpositionen: *es kamen* **an/um**$_{ADV}$ *die 50 Leute*
- TüBa-D/Z - Richtlinien
  - Umbenennung von PAV: in TüBa-D/Z: immer PROP
- UIS (Universitäts-Informations-System) Uni Zürich: Beispiele (unvollständige Liste):
  - Unterschied PIAT / PIDAT entfällt
  - Verfeinerungen: ART $\rightarrow$ ARTDEF | ARTINDEF
  - Neue Tags: KONS (für *usw.*, *etc.*: satzwertige Abkürzungen)
  - Zusatzrichtlinien für die Annotation von *haben/sein*

$\Rightarrow$ Aufgabe bzw. Ergänzung von Unterscheidungen, neue Tags

STTS-Varianten, ist-Zustand

Ziele für den Workshop und die kommenden Monate

- Zusammenführen möglichst aller Informationen über Varianten
  - aus Tagsets, Tools (Taggern...), annotierten Texten, ...
  - mit Kriterienbeschreibung (z.B. Richtlinie für *es sind* an$_{ADV}$ *die 50 Leute* vs. *er schreibt* an$_{APPR}$ *die 50 Leute*)
  - mit Beispielen, ggf. mit Ur-STTS vs. Variante
- Berücksichtigung früherer Versuche zum "Aufräumen" von STTS
  - Tübinger Arbeitstreffen von Dezember 2004 Vortrag Zinsmeister
  - Analysen und Fehlerberichte bzw. Taggingfehler-Beschreibungen
- Dokumentation: Anforderungen
  - systematisch: nach Tags oder nach Variationstypen, etc.
  - ausführlich: vorher/nachher, Beispiele ...
  - mit linguistischen Tests, Listen von klaren Fällen und Grenzfällen
  - online: erst Wiki, dann Website – ggf. auch als Journal-Artikel
  - ggf. mit Tool-Unterstützung zum Abbilden: Alt $\leftrightarrow$ Neu

STTS-Varianten, soll-Zustand

Nutzung der Gelegenheit zum Überblick über Wünsche

- Diskussion des Tagsets unter zwei Gesichtspunkten
  - Linguistische Fragestellungen:
    * (Z.T. notorische) Klassifikationsprobleme,
      z.B. *verrückt*: VVPP↔ADJA — Klatt 2002
    * Ergänzungsbedarf aus theoretischer oder deskriptiv-angewandter Sicht
  - Probleme der automatischen Annotierbarkeit
- Ergänzungsbedarf nach Genres, Registern, Texttypen ...
  - Gesprochene Sprache — Schmidt, Rehbein
  - User-generated content — Faaß, Storrer
  - Lernertexte — Reznicek
  - Historische Texte — Dipper
  - evtl. andere

⇒ Ziel für den Workshop:
Beginn einer Sammlung von möglichen Ergänzungen

Studien zur STTS-Nutzung für spezielle Genres

Beispiel Web-Texte – Sicht der automatischen Annotierbarkeit (1/2)

- Beispielfall DeWaC
  Vergleich: Zeitungstext ↔ Web
  Baroni/Kilgarriff 2006
  Giesbrecht/Evert 2009
  - Tagging mit STTS: verschiedene Tagger
  - Analyse (u.a.) der Tagging-Probleme
- Genres:
  - einfach zu taggen: "expository prose":
    Medizininformation, politische Reden, CeBIT-Nachrichten, ...
  - problematisch:
    Online-Forum, Konferenzinformationen zur Psychologie,
    Texte zu einer Fernsehserie
- Taggingprobleme:
  - Distributionelle Probleme (generelle Schwierigkeit)
  - Sonderzeichen, Sonderformen, Fehler
  - Probleme im Zusammenhang mit Tokenizing

Studien zur STTS-Nutzung für spezielle Genres

Beispiel Web-Texte – Sicht der automatischen Annotierbarkeit (2/2)

- "Standard"-Probleme beim Tagging: Giesbrecht/Evert 2009
  Verwechslung distributionell ähnlicher Tags
  - NN ↔ NE und Klassifizierungsprobleme (z.B.: Tag von *Bahnhofstrasse*?)
  - VVFIN ↔ VVINF und Fragen Klatt 2006
    des Kontexts beim stat. Tagging: links/rechts? Gurevich/Kübler 2008
  - ADJD ↔ ADV: hier u.U. lexikalische Lösung
  - *ein*: PIS↔ CARD
- Spezielle Probleme des Genres 'Web-Daten':
  - Interpunktion innerhalb des Satzes (vgl. lexikal. Lösung U. Zürich)
  - Schwierigkeiten bei Annotation von FW, XY:
    durch "Raten" verwechselt mit NN, NE, ADJ: Distribution oft gleich

⇒ Kombinierbar mit Sicht der Spezifika von bestimmten Texttypen, dabei auch Vorschläge für vereinfachte Tagsets

Vorschläge zur Ergänzung von STTS

Allgemeine Überlegungen (1/2)

- Einfacher Fall: Aufgabe von Unterscheidungen (Typ: `PIAT` $\leftrightarrow$ `PIDAT`)
  $\Rightarrow$ Tilgung ist einfach: Skript für Ersetzung von Tags
- Verfeinerung von Unterscheidungen
  Typ: `ART` $\rightarrow$ `ARTDEF` | `ARTINDEF` U. Zürich
  $\Rightarrow$ Optionale Ergänzung im Sinne des logischen Tagsets,
  oder: `ART`$_{indef}$, wie `NN`$_{sing.dat.mask}$, cf. RF-Tagger Schmid/Laws 2008
- Achtung: Manche Probleme hängen nicht mit dem Tagset zusammen:
  Beispiel: Nicht-lemmatisierbare Items sind oft falsch annotiert:
  für Analyse von Fachtexten: Lemmatisierungskorrektur-Tool
  $\Rightarrow$ keine Auswirkungen aufs Tagset

Vorschläge zur Ergänzung von STTS

Allgemeine Überlegungen (2/2)

- Genre-typische Ergänzungen u.U. als "zweite Ebene" des Tagsets, wie `NN_sing.dat.mask`
- Probleme mit mehrteiligen Items u.U. ausklammern:
  z.B. als Ganzes annotiert als *ADV* U. Zürich
  $\Rightarrow$ Interaktion mit Tokenizing
- Ergänzungen zu Guidelines (Typ: an_{ADV} *die 50 Leute*) sollten sehr gut dokumentiert und ggf. durch Abbildungstools unterstützt werden:
  - $\rightarrow$ auch Korpora brauchen "Prozess-Metadaten": Welcher Tagger, welche (Variante eines) Tagsets, welche zusätzlichen Tools wurden benutzt?
  - $\rightarrow$ Stand von CMDI hierfür?

Angestrebte Dokumentation zu STTS

Allgemeine Zielsetzungen

- Rückwärtskompatibilität:
  Ein "neues" STTS sollte durch wohldefinierte Schritte
  mit dem "alten" STTS in Relation gesetzt werden können
- Dokumentation der Varianten
- Wo möglich Skripte o.ä. für
  - Abbildung
  - Prüfung, wo eventuell manuelle Änderungen nötig sind
  - An-/Ausschalten von Verfeinerungen ("zweite Ebene")

$\Rightarrow$ Alle entsprechenden Beiträge sind sehr willkommen!