

Tagsets (allgemein und linguistisch)

Fröstl, Michael; frostlmichael@gmail.com / Eder, Elisabeth; elisabeth.eder@aau.at

Tagsets sind Annotationsinventare bzw. einfache Annotationsrichtlinien. Sie bilden die Grundlage digitaler Annotationsprozesse (linguistischer Natur) und machen diese transparent und nachvollziehbar. Tagsets stellen sich zumeist in Gestalt von Listen standardisierter Abkürzungen dar. Diese Abkürzungen bilden jene Tags (*Labels*), die im (linguistischen) Annotationsprozess verwendet werden. Sie beruhen auf Konvention. Jeweils ein Tag steht dabei eindeutig für exakt **ein** Phänomen eines Textes oder einer Quelle, respektive eines Gegenstandes, dessen inhärente Eigenschaften (digital) beschrieben, explizit und computerlesbar gemacht werden. Bei digitaler Kennzeichnung und Beschreibung von linguistischen Phänomenen beschreiben einzelne Tags jeweils **ein** linguistisches Phänomen eines Wortes bzw. Satzzeichens (eines Tokens; z. B. seine Wortart = *part of speech* = PoS oder eine morphologische Erscheinung wie etwa den Kasus etc.). Dabei können an das einzelne Wort mehrere Tags angelagert werden – in Abhängigkeit davon, welche linguistischen Phänomene beschrieben werden sollen. Linguistische Tags entsprechen in XML einem Attribut. Die grammatische Kategorie bildet dabei den Attributnamen links des Gleichheitszeichens. Als Attributwert rechts des Gleichheitszeichens (zwischen "...") fungiert der eigentliche linguistische Tag als Teil von Tagsets und als Repräsentant des sprachwissenschaftlichen Einzelphänomens, z. B.:

```
<w pos="noun" numerus="plural">characters</w>
```

Ihrer Funktion nach können linguistische Tagsets grob in zwei Gruppen eingeteilt werden: solche, die allein der morphologischen und/oder der Wortarten-(PoS)-Annotation (*Part-of-Speech-Tagging*) dienen, andererseits solche, die zur Erstellung von *Treebanks* (Baumbanken) vorgesehen sind, also zur syntaktischen Annotation geparster Texte. Je nach Sprache haben sich in der Corpuslinguistik verschiedene Tagsets de facto als Standard durchgesetzt, so etwa das Stuttgart-Tübingen-Tagset (STTS) (Schiller et al. 1999) im Falle der PoS-Annotation deutscher Texte, das beispielsweise beim TreeTagger (Tagger) eingesetzt wird, sowie das darauf aufbauende, aber leicht abgeänderte TIGER-Annotationsschema (Albert et al. 2003) (Verwendung bei *spaCy*). Daneben existieren Ansätze zu universell verwendbaren PoS-Tagsets, wie das *Universal Dependencies PoS-T*, ebenfalls bei *spaCy* im Gebrauch.

Je reduzierter und kürzer Tagsets aus linguistischer Sicht gestaltet sind, desto bessere und schnellere Tagging-Ergebnisse können bei automatischer Annotation mittels Tagger für gewöhnlich erzielt werden, allerdings auf Kosten sprachwissenschaftlicher Differenzierung und Genauigkeit.

Literatur:

- Albert, Stefanie; Anderssen, Jan; Bader, Regine; Becker, Stephanie; Bracht, Tobias; Brants, Sabine; Brants, Thorsten; Demberg, Vera; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; Hirschmann, Hagen; Janitzek, Juliane; Kirstein, Carolin; Langner, Robert; Michelbacher, Lukas; Plaehn, Oliver; Preis, Cordula; Pußel, Marcus; Schrader, Bettina; Schwartz, Anne; Smith, George; Uszkoreit, Hans: TIGER Annotationsschema: 2003. URL: https://www.linguistics.ruhr-uni-bochum.de/~dipper/pub/tiger_annot.pdf.
- Schiller, Anne; Stöckert, Christine; Teufel, Simone; Thielen, Christine: Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset): 1999. URL: <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>.

Software:

spacy , TreeTagger

Projekte:

Universal POS tags, Stuttgart Tübingen Tagset, Tagsets für das Deutsche

Verweise:

Annotation, Annotationsstandards, Lemmatisierung, Markup, Part-of-Speech-Tagging, Tagger, TEI, spaCy, NLP

Themen:

Natural Language Processing

Zitiervorschlag:

Tagsets (allgemein und linguistisch). In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: <https://gams.uni-graz.at/o:konde.177>