Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

herm>A

Dissemination digitaler Methoden

Erfahrungsbericht aus Schulungen mit der textverarbeitenden Pipeline WebLicht des Projekts CLARIN-D

Heike Zinsmeister

03.02.2018, Universität Hamburg

Workshop "Beratungskonzepte in den Digital Humanities"
Projekt forText

Wie können wir traditioneller arbeitenden Geisteswissenschaftlerinnen und Geisteswissenschaftlern die Nutzung digitaler Methoden nahebringen?

- Spielwiese zum Reinschnuppern/Ausprobieren anbieten

Foto: pixabay.com

Disziplinen-Shift: Von Literaturwissenschaft und DH zu Sprachwissenschaft und CL

Beispiel:

- Motto der Gesamttagung: Gottesteilchen der Sprache? Theorie, Empirie und die Zukunft sprachlicher Kategorien
- CL-Tutorium: Syntaktische Korpusaufbereitung mit WebLicht

  "Das Tutorium richtet sich an interessierte Linguisten ohne computerlinguistische Vorkenntnisse. Die Teilnehmenden lernen den Umgang mit korpuslinguistischen Methoden und Werkzeugen, darunter die Annotation und die Abfrage von Korpora mit verschiedenen Sprachwerkzeugen, sowie einfache statistische Analysen, Visualisierung und Interpretation der Ergebnisse."

  Durchführung: Hannah Kermes (Universität des Saarlandes) & Heike Zinsmeister (Universität Hamburg)

Wie können wir traditioneller arbeitenden Geisteswissenschaftlerinnen und Geisteswissenschaftlern die Nutzung digitaler Methoden nahebringen?

- Spielwiese zum Reinschnuppern/Ausprobieren anbieten
- Beispiel "Syntaktische Korpusaufbereitung mit WebLicht"
  - Halbtägiger Block im Rahmen der 1,5tägigen CLARIN-D Doktorandentage "Corpora" 2013
  - Eintägiges "Tutorium" bei der Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS) 2014
- Kontext:
  - Kostenfrei, aber mit Anmeldung
  - Teilnehmende haben Eigeninitiative gezeigt, d.h. sie habe grundsätzlich Interesse / sind neugierig
  - Eher Nachwuchswissenschaftler
  - Arbeit auf eigenen Notebooks

Warum syntaktische Analyse / Annotation?

- Segmentierung
  - Zerlegung in Sätze und Token
  - Basiseinheiten für weitere Analyse
  - Normalisierungsgröße für quantitative Untersuchungen

- Wortarten (parts of speech, pos)
  - Approximation der Syntax (der Funktion?)

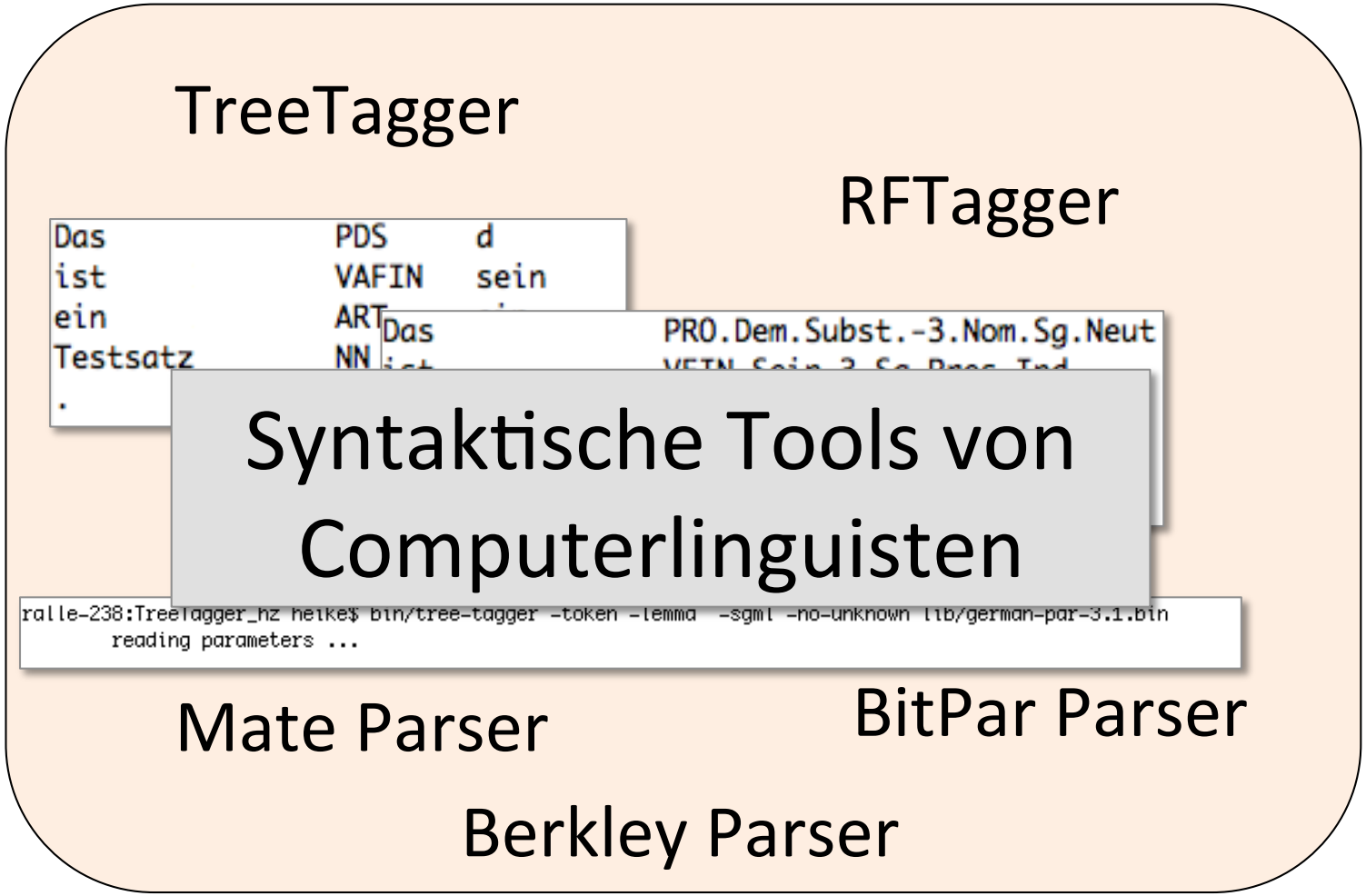
    ART ADJA NN

    VVFIN PPER ADV APPRART NN PTKVZ $.
  - Auffinden von Eigennamen

Warum syntaktische Analyse / Annotation?

- Dependenzen
  - Prädikat-Argument-Strukturen
  - Grammatische Rollen / Funktionen
- Konstituentenstruktur
  - Nominalphrasen als Approximation für Referenten oder für Handelnde
- Topologische Felder
  - Nicht-kanonische Konstituentenabfolgen

    $[Es]_{VF}$ $[meldeten]_{LK}$ $[sich\ drei\ Sprecher\ zu\ Wort]_{MF}$ (Vorfeld-*es*)


TreeTagger
RFTagger
Das PDS d
ist VAFIN sein
ein ART
Testsatz NN
.
Das PRO.Dem.Subst.-3.Nom.Sg.Neut
Syntaktische Tools von Computerlinguisten
ralle-238:TreeTagger_hz heike$ bin/tree-tagger -token -lemma -sgml -no-unknown lib/german-par-3.1.bin
reading parameters ...
Mate Parser
BitPar Parser
Berkley Parser

Syntaktische Tools

Probleme

- Download und lokale Installation der Software
- In- und Outputformate
- Nutzung per Kommandozeile

Lösung

Vorhandene Tools zur syntaktischen Verarbeitung als Webservices anbieten

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

herm>A

Webservice

Universität Konstanz

Wortarten-Analyse?

Karin fliegt nach New York. Sie will dort Urlaub machen.

Tools

Institut für **Maschinelle Sprachverarbeitung**

sfs seminar für sprachwissenschaft

ASV Automatische Sprachverarbeitung

(http://www.worknetncc.com/Uploads/computer.jpg, http://commons.wikimedia.org/wiki/File:Webservice.svg)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

herm>A

Webservice

Universität Konstanz

Wortarten-Analyse?

Karin fliegt nach New York. Sie will dort Urlaub machen.

(http://... , http://... vice.svg)

WebLicht: Ablauf

1. Login: Shibboleth
2. Auswahl / Upload eines Textes (Input Selection)
3. Aufbau einer Verarbeitungskette (Tool Chain)
   - Verschiedene Modi (easy und advanced)
4. Start der Verarbeitungskette
5. Sichten der Analyseresultate
6. Export der Analyseresultate (ggf. direkt in Suchtool Tündra)
7. Ende: Schließen des Browserfensters

Demo von WebLicht

(und Tündra)

https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

(https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra)

Erfahrungen / Lehren

- Digitales als Methode – nicht als Ersatz für die Interpretation

- Konzeptuell: Automatische linguistische Annotation
  - Zugänglichmachung/Strukturierung des "Textmeers"
  - "Bojen" für potenziell Interessantes
  - Annotationskategorien als Proxies für Analyseeinheiten
  - Grenzen der Annotationsqualität

Erfahrungen / Lehren

s1: `` Ross Perot w‰re vielleicht ein pr‰chtiger Diktator "

- Herausforderung der Formate / Toolhandhabung
  - Textkodierung (Sensibilisierung für Text als digitales Objekt, das mit Tools interagiert)
  - Eingabe- und Ausgabeformate für Tools
  - Konvertierung in durchsuchbare / manipulierbare Formate
  - Syntax von Suchanfragen (analytische Denkweise)

- Einstieg zur Eigenarbeit

Notwendige Horizontverschmelzung

- (Überhöhte) Erwartungen vs. Unkenntnis über tatsächliche Forschungsfragen und Bedarfe der Teilnehmenden
  - Wie man sprachliche Einzelhandlungen / Sequenzen von funktionalen Bausteinen extrahiert
  - Wie man argumentative / beschreibende / bewertende Strukturen extrahiert
  - Wie man fragmentarische Äußerungen / Ellipsen automatisch annotiert
  - Eigene Korpuserstellung
    - Für quantitative Untersuchungen
    - Für gesprochene Sprache
  - Computerlinguistische Anwendungen für kirchenslavisch-kyrillische Texte / valencianische Alltagssprache
  - Untersuchungen von Nominalisierungen

Zurück zu den weiteren Leitfragen des Workshops

Ermittlung von ‘digitalen Bedarfen’ in den geisteswissenschaftlichen Forschungsgemeinschaften

Persönliche Gespräche, vergleiche:

- Umfrage des Projekts gwin an der Universität Hamburg zu Bedarfen der Datenhaltung / Methodenunterstützung
- Experten-Interview im Rahmen des DFG-Projektes "Future Publications in den Humanities" (Universitätsbibliothek, Humboldt-Universität zu Berlin)

Umsetzung von Beratung und Dissemination in Forschung und Lehre

- **Beispielanalysen zeigen** (Neugier/Lust wecken!)
- Hemmschwellen abbauen
- Dateninput / Datenoutput nachvollziehbar machen
- Anknüpfen an die traditionelle Herangehensweisen und Formate (Look and Feel)
- Transparenz der Verfahren
- Rückführbarkeit auf die Datengrundlage

Sicherung der Nachhaltigkeit von Beratung und Dissemination im Bereich der Digital Humanities

DARIAH-DE
Digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften

CLARIN-D

- Nationale Infrastrukturen
- Dauerstellen – "Fakultäre MethodenberaterIn"
- Community Effort – Experten-Datenbank der Dhd?

Danke für Ihre Aufmerksamkeit!