

Sprachtechnologie als Grundlage für die maschinelle Auswertung von Texten

Dr.-Ing. Michael Piotrowski
Leibniz-Institut für Europäische Geschichte
<piotrowski@ieg-mainz.de>  @true_mxp

Bamberg, 20. November 2015



Quantitative Analysen von Texten

... liefern Ausgangspunkte für qualitative Analysen.

- ▶ Wie oft kommt das Wort **backen** vor?
- ▶ Wie oft kommen **backen** und **kochen** gemeinsam vor?
- ▶ Wie oft kommen Verben vor, die mit der Zubereitung von Essen zu tun haben?
- ▶ Wie oft kommt **kochen** in der Bedeutung »Essen zubereiten« vor? Und was ist hier mit **Gericht** gemeint?
- ▶ Welche Wörter kommen im Kontext von **kochen** und **backen vor**? Sind diese mehrheitlich positiv konnotiert?
- ▶ Usw. usf.

Schon um diese **einfachen Fragen** zu beantworten, ist linguistisches Wissen nötig – z. B., dass **Wörter** in verschiedenen **Wortformen** auftreten können.

Was ist Sprachtechnologie (NLP)?

- ▶ Angewandte Computerlinguistik
- ▶ Textverarbeitung unter Nutzung linguistischen Wissens
- ▶ Grundlegende Werkzeuge:
 - ▶ Tokenisierung
 - ▶ Morphologische Analyse und Lemmatisierung
 - ▶ Wortartenerkennung (part-of-speech [POS] tagging)
 - ▶ Syntaktische Analyse
- ▶ Höhere Analysen und Anwendungen bauen darauf auf
- ▶ Viele wichtige Anwendungen: Volltextsuche, Text mining, akustische Spracherkennung, OCR, Rechtschreibprüfung, Textzusammenfassung, Fragebeantwortung, Eigennamenerkennung, maschinelle Übersetzung, ...

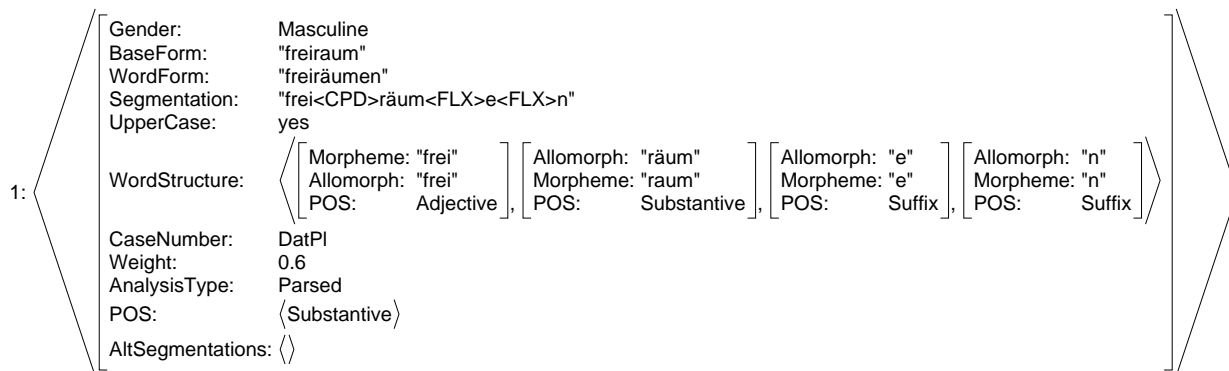
Part-of-speech tagging

- ▶ POS-Tagger annotieren die Wortformen eines Texts mit morphosyntaktischen Informationen (teils auch mit Grundform)

| | | |
|-----------------|-----|-----------------|
| The | DT | the |
| business | NN | business |
| unceremoniously | RB | unceremoniously |
| collapsed | VBD | collapse |
| back | RB | back |
| in | IN | in |
| 2009 | CD | @card@ |
| after | IN | after |
| suffering | VBG | suffer |
| ... | ... | ... |

Beispielhafte morphologische Analyse

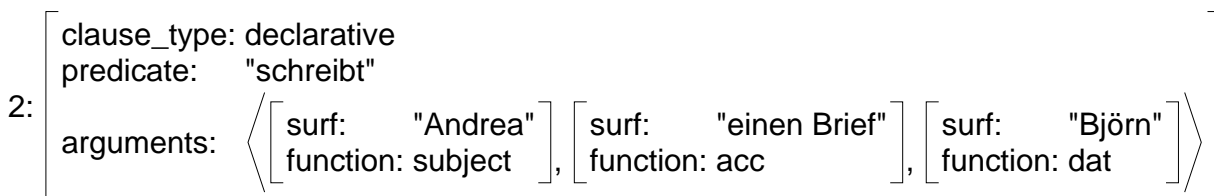
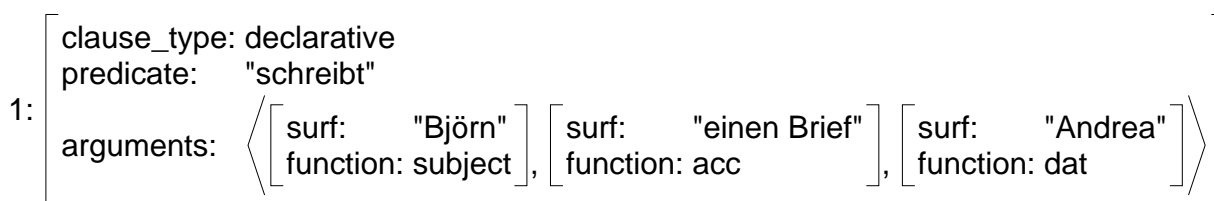
"Freiräumen"



IEG
Leibniz-Institut für
Europäische Geschichte

Beispielhafte syntaktische Analyse

"Björn schreibt Andrea einen Brief."



IEG
Leibniz-Institut für
Europäische Geschichte

NLP-Werkzeuge

Heute sind eine Vielzahl von NLP-Ressourcen, -Werkzeugen und -Werkzeugkästen verfügbar, z. B.

- ▶ RFTagger <http://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/>
- ▶ Mate Tools <http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.html>
- ▶ SMOR <http://www.cis.uni-muenchen.de/~schmid/tools/SMOR/>
- ▶ Stanford NLP tools <http://nlp.stanford.edu>
- ▶ DKPro
<https://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/>
- ▶ GATE <https://gate.ac.uk>
- ▶ NLTK <http://www.nltk.org>
- ▶ WordNet <http://wordnet.princeton.edu>
- ▶ GermaNet <http://www.sfs.uni-tuebingen.de/GermaNet/>



Aber ...

- ▶ Für Werkzeuge angegebene Performance gilt für **homogene** Textsammlungen aus der **gleichen Domäne** wie die Trainingsdaten (üblicherweise Zeitungstexte)
 - ▶ Performance auf Texten aus **anderen Domänen** ist typischerweise **niedriger**
 - ▶ **Englisch** ist eine Ausnahme, für andere Sprachen sind sehr viel weniger Ressourcen verfügbar
 - ▶ **Historische Sprachen und Texte** stellen zusätzliche Probleme
- ↪ NLP-Werkzeuge nicht als Black box anwenden



Historische Texte

Einige typische Merkmale:

- ▶ **Medium:** Original gedruckt auf Papier, geschrieben auf Pergament, gemeißelt in Marmor oder eingedrückt in Ton – muss zuerst ins digitale Medium transferiert werden.
- ▶ **Sprache:** Ausgestorben und anders als heutige Form; keine Native speakers verfügbar, keine oder nur wenige NLP-Ressourcen.
- ▶ **Schrift:** Historisches Schriftsystem oder benutzt historische Zeichen und Formen, Schriften, Abkürzungen, Ligaturen usw.
- ▶ **Schreibung:** Orthografie unterscheidet sich von der heutigen und/oder gar keine normierte Orthografie.
- ▶ **Integrität:** Text möglicherweise beschädigt oder unklar; Transkription (inkl. Digitalisierung) hat u. U. neue Fehler hinzugefügt.

→ **Auswirkungen auf NLP**



Historische Texte

- ▶ Breite Vielfalt von Ansätzen und Ergebnissen zeigen, dass NLP-Werkzeuge für historische Sprachen meistens spezifisch für bestimmte Textsammlungen sind.
- ▶ NLP für historische Texte oft im Kontext diachroner Studien; diachrone Korpora sind per definitionem heterogen → kein einzelnes Werkzeug kann optimale Performance über das ganze Korpus liefern.
- ▶ Einzelne Teile von diachronen Korpora sind oft zu klein, um statische Werkzeuge damit zu trainieren.
- ▶ Auch Texte aus der gleichen Periode weisen hochgradig variable Schreibung auf, wodurch die verfügbare Datenmenge weiter sinkt.
- ▶ **Generell schlechtere Performance auf historischen Texten als auf »modernen« Texten**



Zusammenfassung

- ▶ Sprachtechnologie wird für jede Auswertung von Texten benötigt, die über das Zählen von Oberflächen hinausgeht.
- ▶ Das gilt besonders für Sprachen mit reicher Morphologie (z. B. Deutsch).
- ▶ Inzwischen sind eine Vielzahl von Werkzeugen und Ressourcen für »gängige« Sprachen und Anwendungen verfügbar.
- ▶ Aber: Geisteswissenschaftliche Vorschung beschäftigt sich oft mit »speziellen« Texten. Werkzeuge verhalten sich anders und müssen sorgfältig evaluiert werden.
- ▶ Historische Texte sind besonders komplex; zusätzlich Interferenzen durch Transkription. Hier gibt es keine einfachen Lösungen.