Wortartannotation für die digitalen Geisteswissenschaften

Ulrich Heid

Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie –
Bereich Computerlinguistik – Sprachtechnologie
Universitätsplatz 1, D 31141 Hildesheim

Wien, DH-Galerie, 13. Oktober 2015

Überblick

- Textanalyse für die digitalen Geisteswissenschaften:
  Der Kontext der Wortartannotation ("Tagging")
- Motivation: Wozu überhaupt Texte mit Wortarten annotieren?
- Was wird annotiert?
  - Typische Ausgabe eines Taggers
  - Wortartenklassifikation und Inventar von Annotations-Etiketten
- Das Werkzeug *TreeTagger*
  - Zweck und Prinzipien
  - Wissensquellen und Verfahren
  - Ergebnisse, Vorteile und Probleme
- Nutzung von *TreeTagger* bzw. Tagging im allgemeinen
  in DH-Aktivitäten:
  Beispiele – Probleme – Lösungsansätze
- Zusammenfassung

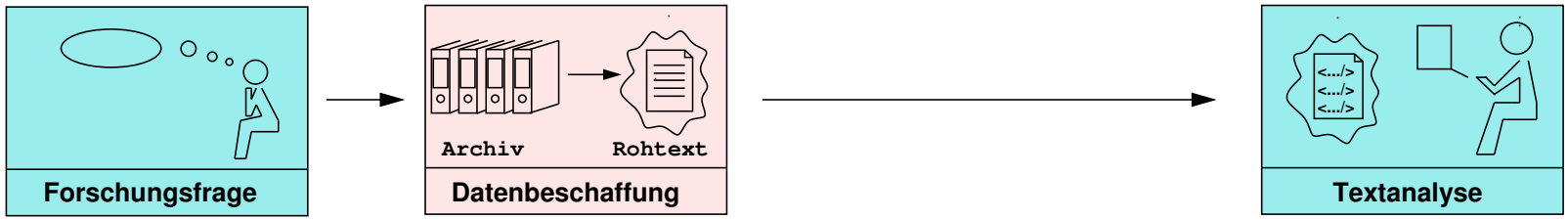
Kontext

- Projekte und Initiativen – Kollegen und Mitarbeitende:
  - 1994–97: Tagset- und Taggerentwicklung:
    Projekt "Textkorpora und Werkzeuge zur Erschließung" (Land BW):
    Helmut Schmid – Anne Schiller, Simone Teufel (Stuttgart)
  - 2010–: Tagging-Korrektur bei Fachtexten:
    EU-Projekt TTC (2010–12) und Industriekooperationen:
    Anita Ramm, Johannes Schäfer, Simon Tannert (Stuttgart)
  - 2012–: Initiative zur Erweiterung von STTS:
    "Grass roots"-Initiative und CLARIN-D: F-AG 7:
    Heike Zinsmeister (Hamburg), Katrin Beck (Tübingen) u.a.
  - 2012-15: Einbindung in eine Arbeitsumgebung für die DH:
    BMBF-Projekt e-Identity:
    Fritz Kliche (Hildesheim), André Blessing (Stuttgart)
- Beispiele aus Arbeiten zur Terminologie des Car-Sharing:

  Kooperation mit Wirtschaftsinformatik U. Hildesheim (R. Knackstedt)

Textanalyse für die digitalen Geisteswissenschaften

Zielsetzungen

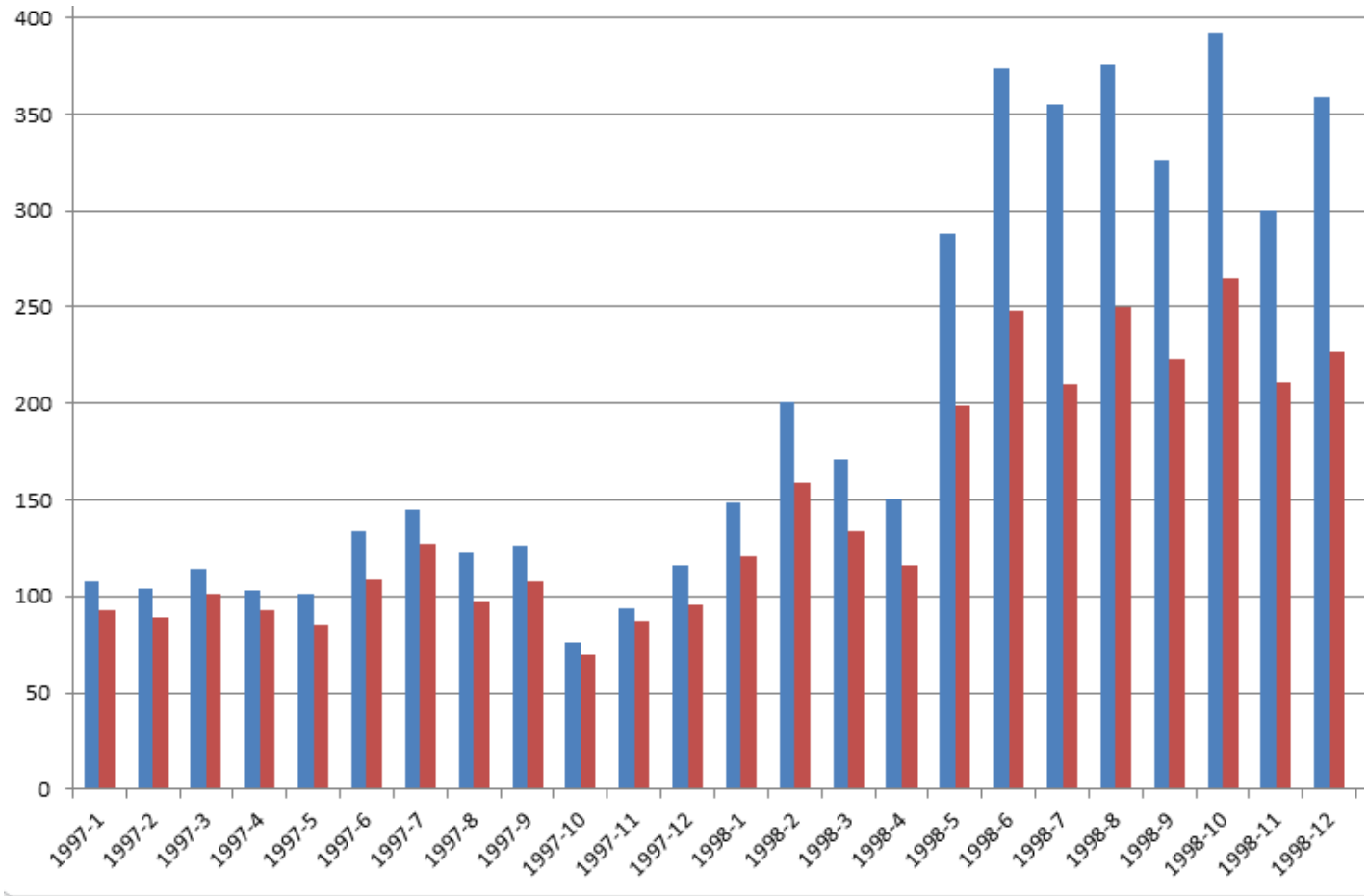


- Ausgangspunkt: Fachwissenschaftliche Forschungsfrage – analysierbar anhand größerer Mengen von Texten
- Beispiele für mögliche Zielsetzungen:
  - Suche nach relevanten Passagen
    - Meinungen von Politikern in Zeitungstexten
    - Bastelanweisungen in online-Heimwerker-Foren
  - Beschreibung von Entwicklungen in der Zeit
    - Medienaufmerksamkeit für ein Thema in einem Zeitraum
    - Entwicklung von Diskursen zu einem Thema
  - Identifikation von Sprechweisen bestimmter Akteure
    - Angebote für Car-Sharing von entsprechenden Firmen
    - Naturschützer vs. Agrarlobby zu nachhaltiger Landwirtschaft

Textanalyse für die digitalen Geisteswissenschaften

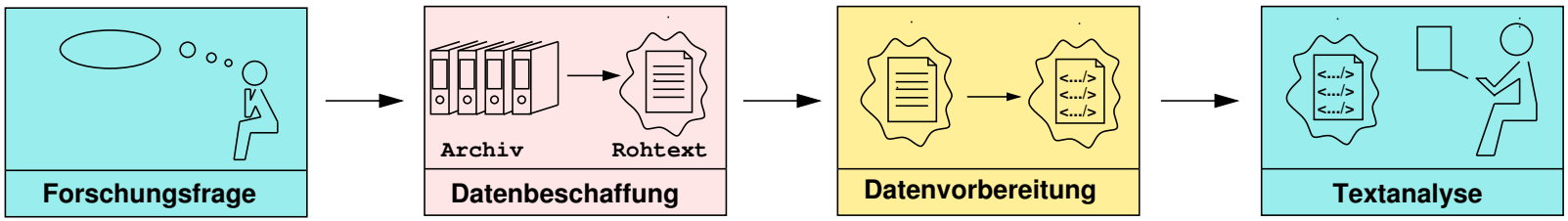
Ein Beispiel für typische Ergebnisse aus der Politikwissenschaft

Medienaufmerksamkeitsanalyse:



Textanalyse für die digitalen Geisteswissenschaften

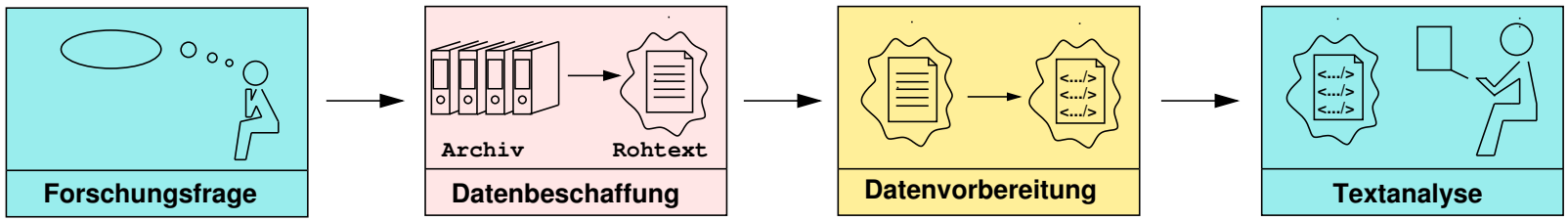
Korpuslinguistische Sicht auf die Arbeitsschritte (1/2)



1. Beschaffung digital(isiert)er Texte
2. Formatkonvertierung und -homogenisierung $\longrightarrow$ UTF-8
3. Auszeichnung mit Metadaten Provenienz, Texteigenschaften, etc.
4. Linguistische Annotation
5. Abfrage der Texte: interaktiv oder automatisiert

Textanalyse für die digitalen Geisteswissenschaften

Korpuslinguistische Sicht auf die Arbeitsschritte (1/2)



1. Beschaffung digital(isiert)er Texte
2. Formatkonvertierung und -homogenisierung $\longrightarrow$ UTF-8
3. Auszeichnung mit Metadaten Provenienz, Texteigenschaften, etc.
4. Linguistische Annotation
5. Abfrage der Texte: interaktiv oder automatisiert

Textanalyse für die digitalen Geisteswissenschaften

Korpuslinguistische Sicht auf die Arbeitsschritte (2/2)

- Linguistische Annotation:
  Anreicherung von Texten mit Sprachwissen
  - Anfang und Ende von Sätzen — Satz-Tokenisierung
  - Einteilung in Wörter — Wort-Tokenisierung
  - Klassifikation der Wortformen nach der Wortart — POS-Tagging
  - Zuordnung von Wortformen zu Lemmata — Lemmatisierung
  - Identifikation syntaktischer Strukturen — Parsing
  - Semantische Annotation der Lesarten — Semant. Tagging
  - und andere...

Wieso Texte linguistisch annotieren?

Motivation: Generalisierung und Spezifizierung bei der Suche

Generalisierung (1/2)

- Lemmatisierung:
  Statt Suche nach einzelnen Wortformen $\longrightarrow$ Lemmata
- Wortartannotation (POS-Tagging):
  Statt Suche nach einzelnen Wörtern $\longrightarrow$ Wortklassen

Beispiel: Suche nach einer Sequenz:

Artikel – Adjektiv – Nomen – Verb (im Infinitiv)

– *eine gültige E-Mail-Adresse angeben*
– *das absolute Rauchverbot beachten*
– *den gewünschten Ort anklicken*
– *Die nächstgelegene Vertriebsstelle heraussuchen*
– *die folgenden Bedingungen erfüllen*

Wieso Texte linguistisch annotieren?

Motivation: Generalisierung und Spezifizierung bei der Suche

Generalisierung (2/2)

- Beispiel: Suche auf lemmatisiertem Text:
  - *Was kostet es, car2go zu fahren?*
  - *Fahren Sie mit stadtmobil, dem CarSharing-Marktführer.*
  - *Wie fährt sich ein bestimmtes Automodell?*
  - *ist mehr Kilometer gefahren als ursprünglich vereinbart*
  - *in welcher Wagenklasse Ihr Wunschfahrzeug fährt*

$\Rightarrow$ In allen Fällen:
Eine generalisierte Anfrage, mehrere verschiedene Ergebnisse

Wieso Texte linguistisch annotieren?

Motivation: Generalisierung und Spezifizierung bei der Suche

- Spezifizierung: Auflösung von Mehrdeutigkeiten:
  Ein(e) Wort(form) $\longleftrightarrow$ zwei (oder mehr) Analysen
- Beispielfall *sichere Sachwerte*:
  - Flucht aus Aktien in$_{PRAEPOS}$. sichere$_{ADJEKTIV}$ Sachwerte$_{NOMEN}$
  - Er sagte, sein Unternehmen sichere$_{VERB}$ Sachwerte$_{NOMEN}$
- Relevant für präzise Suche und adäquate Ergebnisse:

2424	sichere	Adjektiv	1984
		Verb	440

Korpus HGC, 204 M Wörter

Wortartannotation: Ausgabe des Werkzeugs *TreeTagger*

Ergebnis-Ausgaben von POS-Taggern

- Annotation von Wortart (und Lemma) an je einzelne Wortformen
  - Tabellarisch, wortweise — one word per line

das	ART	d-
absolute	ADJA	absolut
Rauchverbot	NN	Rauchverbot
beachten	VVINF	beachten

  - An Wortformen angehängt, zur Visualisierung in Korpora: das/ART absolute/ADJA Rauchverbot/NN beachten/VVINF
- Repräsentation:
  – In der Regel als separate XML-Daten, — stand-off XML
  – mit Verweisen — Links

Wortartannotation: Welche Art von Annotationen?

Wortarten-Tagsets – am Beispiel von STTS

- Wortartklassifikation setzt folgendes voraus:
  - Ein Modell der Wortarten, d.h. Kriterien zu ihrer Unterscheidung — Annotationsrichtlinien für manuelle Annotation
  - Ein Inventar der zu unterscheidenden Wortarten — Tagset
- Unterscheidungskriterien für Wortarten:
  - Linguistisch motiviert — nach linguistischer Theorie
  - Durch *automatische* Unterscheidbarkeit motiviert — für Tagger
- Aktuelle(s) Tagset(s) für Deutsch: konsensbasiert
  - Stuttgart-Tübingen TagSet — STTS: Schiller et al. 1994 u.ö
  - Alternativ: Münsteraner Tagset — Steiner et al.
  - Varianten von STTS: Universität Zürich u.a.
  - Ergänzungen zu STTS: z.B. für Nicht-Standard-Texte

Wortartannotation: Welche Art von Annotationen?

Das STTS-Tagset (1/3)

- Umfang: Standardversion hat 54 Tags, davon

für Verben	12
für Nomina	2
für Adjektive	2
für Konjunktionen	4
für Partikeln	6
für Pronomina	14

- Dokumentation:
  - Annotationsrichlinien — Guidelines
    - Kriterien pro Wortart
    - Tests zur Abgrenzung gegen andere Wortarten
  - Tabellen als Kurz-Erläuterung

Wortartannotation: Welche Art von Annotationen?

Das STTS-Tagset (2/3)

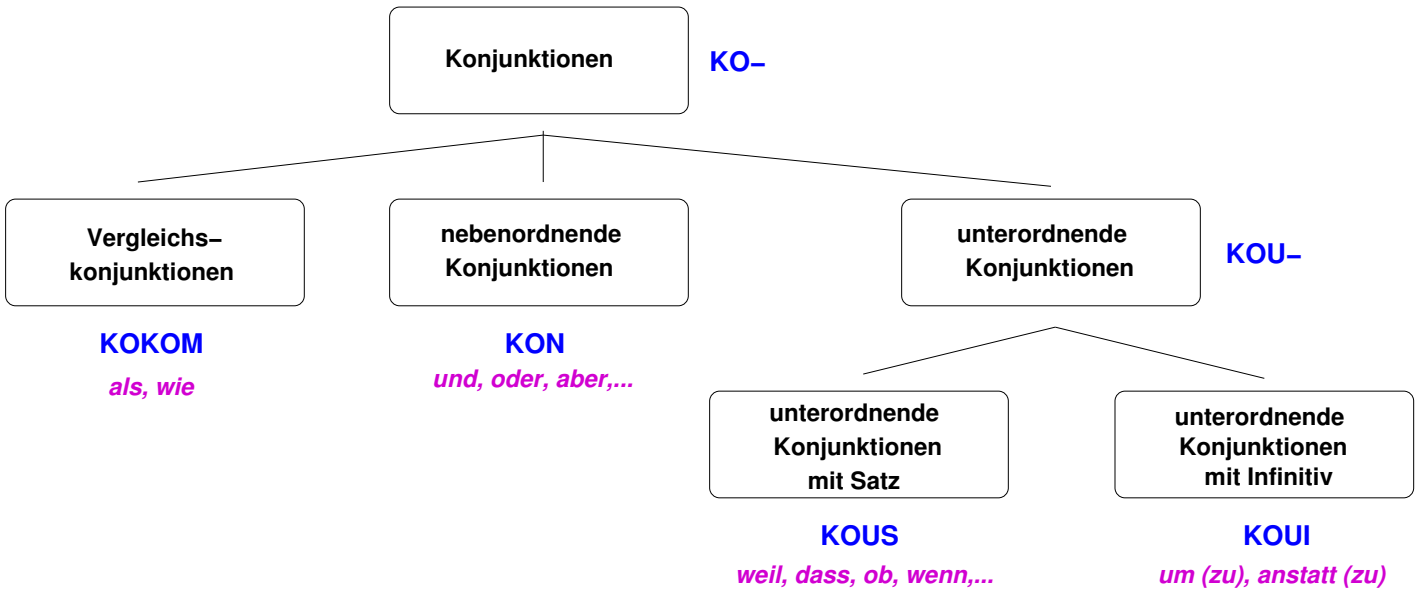
- Klassifikationskriterien
  - Lexikalisch, z.B.:
    - Vollverb *Was kostet es, car2go zu fahren$_{VVINF}$?*
    - Modalverb *cambio-Kunden können$_{VMFIN}$ ... Fahrzeuge nutzen*
    - Hilfsverb *Dafür wurde$_{VAFIN}$ cambio ... ausgezeichnet*
  - Formbezogen, z.B.:
    - Attributives Adjektiv *eine gültige$_{ADJA}$ Fahrerlaubnis*
    - Prädikatives Adjektiv *sind 6 Monate ab Einreise gültig$_{ADJA}$*
  - Distributionell, z.B.:
    - Attribuierendes Indefinitpronomen *alle$_{PIAT}$ Fahrten ...*
    - Substituierendes Indefinitpronomen *alle$_{PIS}$ ... sind versichert*
- Hierarchische Struktur:
  unterstützt unterspezifizierte Suchanfragen

Wortartannotation: Welche Art von Annotationen?

Das STTS-Tagset (3/3)

- Hierarchische Strukturierung

Logical Tagset: Leech 1997: 33



- Partielle Generalisierungen möglich in den Teilhierarchien:
  z.B. Suche nach allen unterordnenden Konjunktionen ("KOU–")

Das Annotationswerkzeug *TreeTagger*

Allgemeines

- Entwickelt von Helmut Schmid (CIS, München) Schmid 1994
- Einer von mehreren Taggern,
  die zu einem Standardwerkzeug geworden sind
- Information und Parameterdateien:

  URL:http://cistern.cis.lmu.de/

  URL:www.linguistics.ruhr-uni-bochum.de/stts
- Parameterdateien für verschiedene Sprachen
- Verfügbar auch über WebLicht

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

Das Annotationswerkzeug *TreeTagger*

Zweck und Prinzipien

- Tagger dienen dazu,
  - Wörter (= Wortformen) in laufendem Text
    mit POS-Annotationen zu versehen
  - Hypothesen für unbekannte Wörter zu machen,
    d.h. solche, die nicht im Lexikon des Werkzeugs sind,
    deren POS aber aus dem Kontext erschlossen werden kann
  - Entscheidungen zu treffen, wenn
    einer Wortform mehrere Beschreibungen zugeordnet werden können,
    wiederum auf Basis von Wissen aus dem Kontext
- Grundprinzip:
  - Kombination
    aus Lexikonwissen
    und Wissen aus manuell annotierten Texten
  - Anwendung des Wissens auf neue (ungesehene) Texte

Training

Das Annotationswerkzeug *TreeTagger*

Wissensquellen: Lexikon

- Lexikon für Wortformen:
  - Enthält Form, Lemma, Tag und Wahrscheinlichkeiten für die Tags:

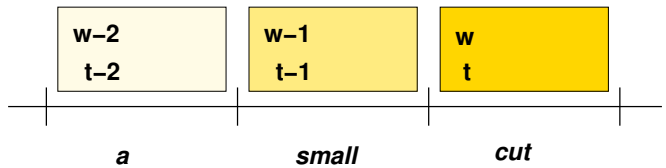
fährt	fahren	VVFIN	1.0
fährst	fahren	VVFIN	1.0
fahren	fahren	VVFIN	0.46
fahren	fahren	VVINF	0.54
gefahren	fahren	VVPP	1.0
...			

- Lexikon für "Wortenden" suffix lexicon
  - Enthält Wahrscheinlichkeiten für POS-Tags, gegeben Buchstabenfolgen am Wortende
  - Beispiele aus dem Deutschen:
    - *-ung* → NN
    - *-iges* → ADJA
    - *-st* → VVFIN
  - Organisiert als Buchstabenbaum, zum schnellen Nachschlagen

Das Annotationswerkzeug *TreeTagger*

Wissensquellen: Kontextwahrscheinlichkeiten aus manuell annotierten Texten

- Beispiel: Tag für EN *cut* in:
  In fact ... does not preclude a small cut
  in interest rates soon,
  because sterling has lately been strong.

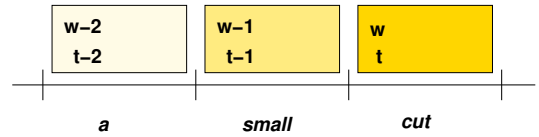


BNC

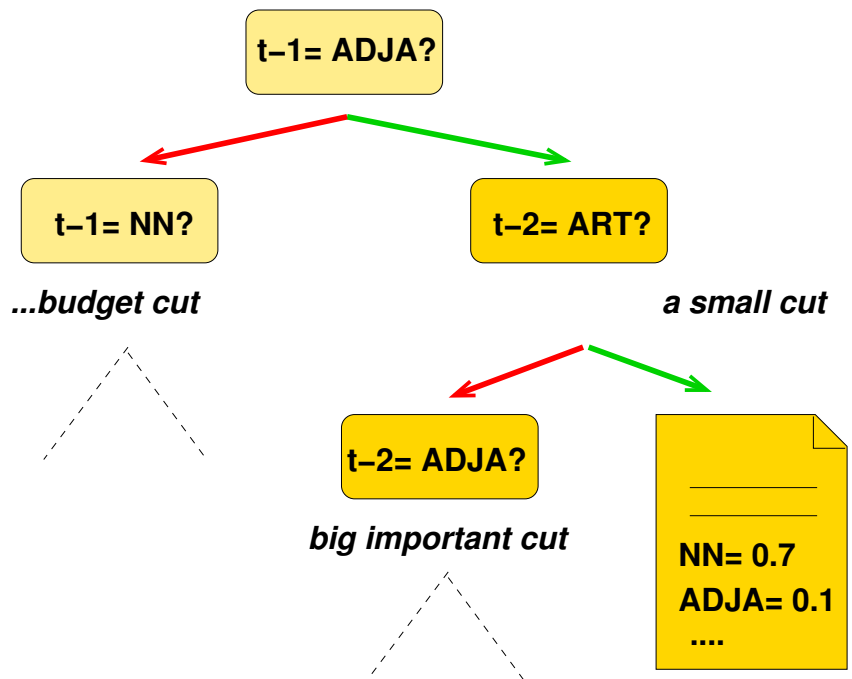
- Berücksichtigung des Kontexts:
  - *cut* könnte Nomen oder Verb sein
  - Aus handannotierten Korpora ist berechnet worden:
    - Wahrscheinlichkeit von *cut*/`NN`, gegeben vorausgehendes `ADJA`
    - Wahrscheinlichkeit von *cut*/`VVFIN`, gegeben vorausgehendes `ADJA`
    - Wahrscheinlichkeit von *cut*/`VVINF`, gegeben vorausgehendes `ADJA`
    - Wahrscheinlichkeit von *cut*/`NN`,
      gegeben vorausgehende Sequenz aus `ART` und `ADJA`
    - usw.
  - Benutzt wird der beste Wert:
    maximal entscheidungsrelevanter Test für die beiden Vorgängerwörter

Das Annotationswerkzeug *TreeTagger*

Entscheidungsbäume als Wissensquelle



- Die namengebenden Entscheidungsbäume
  werden automatisch aus Tripel-Wahrscheinlichkeiten errechnet:
  - Die Sequenzmuster werden als Tests kodiert
  - Entscheidungsziel ist, die informationsreichsten Tests anzuwenden:
    mit welchen Tests wird am meisten Information über das dritte Tag gewonnen?



Das Annotationswerkzeug *TreeTagger*

Ergebnisqualität

- Qualität: Genauigkeit (Akkuratheit) der Annotation tagging accuracy
  - Gemessen als Anteil korrekter POS-Tags an allen vergebenen Tags
  - Messung gegen manuell annotiertes Material Gold Standard
  - Übliche Werte für deutschen Zeitungstext,
    wobei Training auf Zeitungstext erfolgt: 96-98%
- Einflussgrößen für die Genauigkeit:
  - Wie "zeitungs-ähnlich" ist der Text? – Stil, Grammatik, etc.
  - Welcher Anteil unbekannter Wörter ? – Fachtexte, historische Texte
  - Wie gut passt das Lexikon zum zu bearbeitenden Texttyp?
    Hinweis: TreeTagger kann Items markieren, die nicht im Lexikon sind

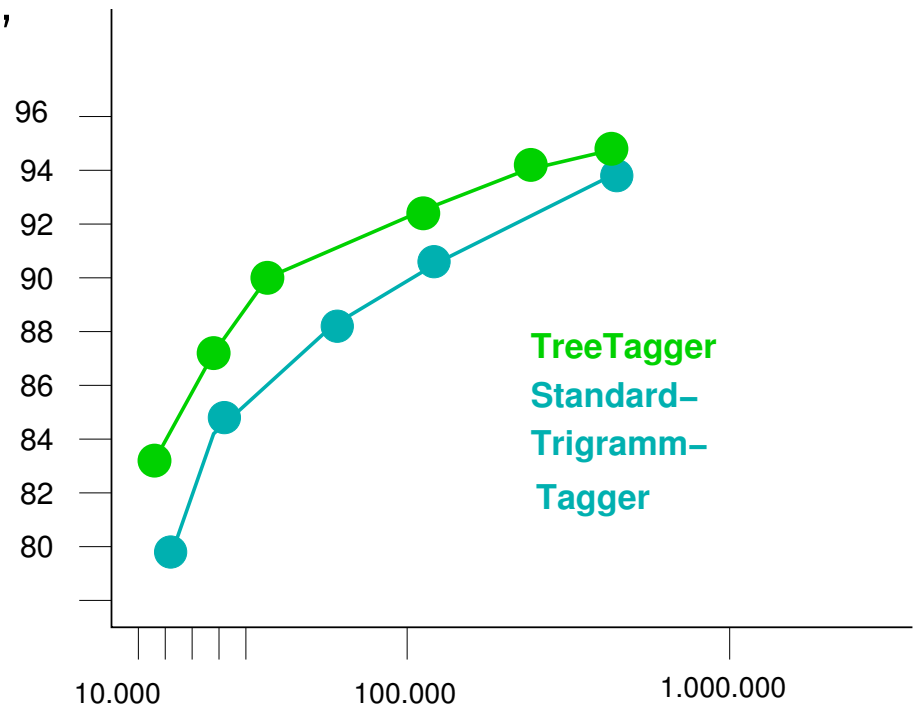
Das Annotationswerkzeug *TreeTagger*

Vorteil: geringe Abhängigkeit von der Größe der Trainingstexte

- Bei nahezu allen statistischen Werkzeugen gilt:
  je mehr (Trainings-)Daten, desto besser
- Trainingsdaten für TreeTagger
  müssen von Hand erstellt werden: Aufwand und Kosten
- Daher macht es einen Unterschied,
  wieviel Trainingsdaten nötig sind
- Empfohlene Mindestmenge:
  abhängig von
  - Anzahl der Tags im Tagset
  - Unterscheidbarkeit der Tags
  - "Homogenität" der Texte

  Bei DE Zeitungstext und STTS:
  40.000 Wörter

Schmid 1994



Das Annotationswerkzeug *TreeTagger*

Probleme für die Annotationsqualität deutscher Texte (1/2)

- Schwer unterscheidbare Wortklassen (bzw. Tags):
  Beispielfall `NN` $\leftrightarrow$ `NE`:
  - *CarSharing mit cambio$_{NE}$ funktioniert nach einem einfachen Prinzip*
  - *Laden Sie sich nun die kostenlose car2go$_{NE}$ App herunter*
  - *einen bequemen Van$_{NE}$ für den Urlaub*
- Konstruktionen von geringer Wahrscheinlichkeit:
  - *So geht's: Anmelden, finden, fahren, abstellen.*
    Anmelden/VVFIN, finden/VVFIN, fahren/VVFIN, abstellen/VVFIN.
- Verbkomplexe:
  - *Sobald ..., wird Ihr Konto innerhalb einer Woche wieder entsperrt$_{VVFIN}$*
  - *Die Kosten ... sind bereits im Nutzungstarif inkludiert$_{VVFIN}$*

  Probleme:
  - Tagger kann nicht weit genug nach links "schauen"
  - Rechter Kontext (z.B. Satzende) kann nicht berücksichtigt werden

Das Annotationswerkzeug *TreeTagger*

Probleme für die Annotationsqualität deutscher Texte (2/2)

- Fenster nach rechts als mögliche Verbesserung Sandra Kübler, Ohio
- Beispiel: Abgetrennte Partikelverben
  - *Steigen Sie einfach bei cambio ein !*
  - *geben Sie Ihre Führerschein- und -Bankdaten ein .*
  - *Laden Sie sich nun die kostenlose car2go App herunter*
  - *Statt eines privaten PKW stehen dort ein oder mehrere cambioAutos*
  - *Sie erreichen den BuchungsService unter : 0421 - 7946643*
- Probleme:
  - Rechter Kontext für Entscheidung nicht nutzbar: *ein oder mehrere*
  - "Unvollständige" Präpositionalphrase: *unter : 0421 - 7946643*
- Große Mehrheit der Fälle wird aber korrekt analysiert

Nutzung von *TreeTagger* in DH-Aktivitäten

Allgemeines

- Grundlage für die Suche nach Textbelegen
- Lemmatisierung als Abstraktionsebene, statt Wortformen
- Mitunter auch als Grundlage für tiefere linguistische Annotation: syntaktische Analyse – Koreferenz – Sentiment – ....
- Beispielfall: Projekt *e-Identity*:
  Textanalyse für die Politikwissenschaft Blessing et al. 2014
  - Suche nach Termini: *zivilgesellschaftlich + Organisation*
  - Suche nach Sätzen, die die Meinung eines politischen Akteurs zitieren:
    – *Kardinal ... bestätigte die Haltung der katholischen Kirche*
    – *Merkel sagte, Europa müsse gemeinsam handeln...*

Nutzung von *TreeTagger* in DH-Aktivitäten

Extraktion von Metadaten aus Textarchiv-Dokumenten

Kliche et al. 2014, 2015

- Metadaten zu
  Zeitungsname – Autor – Erscheinungsdatum – Texttyp – ...
- Halbstrukturierte Dokumente:
  Innerhalb eines Archivs meist gleichartig
  z.B. LexisNexis, Factiva
- Oft gibt es Indikatoren für Metadaten-Typen:
  Wörter/Lemmata – Wortsequenzen – Zahlenangaben – ...
- Diese Indikatoren kann man flexibel im Text"strom" suchen
  und als "Aufhänger" für eine Metadatensuche nehmen
- Typähnliches Beispiel:

Protests break out in Kabul, page 32	$\longrightarrow$ *Titel,Verweis*
172 words	$\longrightarrow$ *Wortzählung*
2 December 2008	$\longrightarrow$ *Datum*
English	$\longrightarrow$ *Sprache*
(c) 2008	$\longrightarrow$ *Copyright-Jahr*
document 009762456D456	$\longrightarrow$ *Dokument-ID*

Nutzung von *TreeTagger* in DH-Aktivitäten

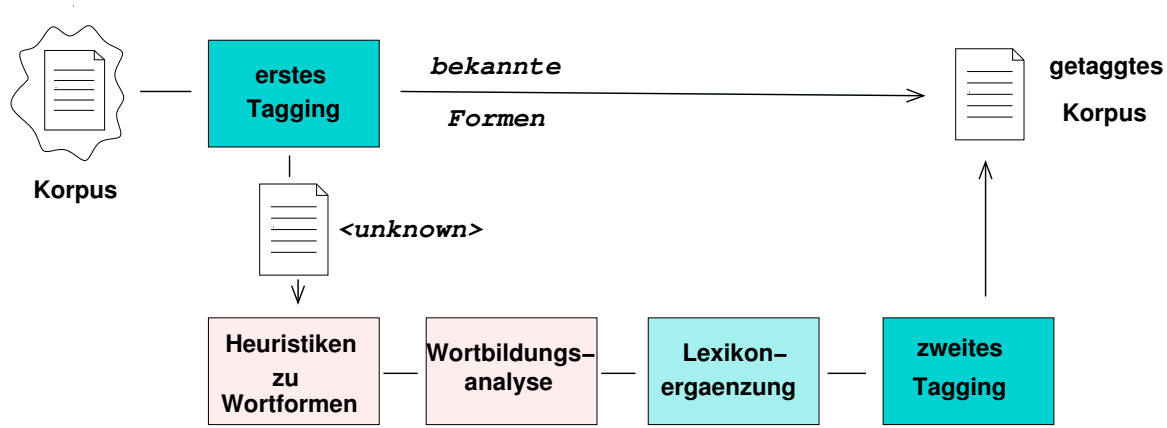
Extraktion von Ontologie-Bausteinen aus Texten von online-Foren

- Analyse von Texten zur Extraktion von Definitions-Sätzen:
  Beispiele:
  - *Der Deltaschleifer ist ein Schleifgerät mit dreieckiger Schleifplatte*
  - *Der Tacker ist ein Elektrowerkzeug, mit dem man Nägel...*
- Suche im Korpus anhand von
  POS-basierten Mustern und Wortformen:
  ART – NN – "ist" – "ein." – NN – (APPR)? – PRELS – ...
- Problem: Unbekannte Wörter
  - Nicht im Taggerlexikon:
    Deltaschleifer, Schleifgerät, Schleifplatte, Tacker
  - Tagger gibt aus: [lemma = unknown]
  - Tagger kann die Wortart "unbekannter" Wörter raten,
    aber wenn zuviele unnbekannte Wörter im Text sind,
    sinkt die Qualität der Annotation

Nutzung von *TreeTagger* in DH-Aktivitäten

Verbesserung der Tagging-Qualität

- Erster Tagging-Durchgang:
  Ausgabe aller Wortformen, die nicht im Taggerlexikon sind
- Sammlung ähnlicher Formen in einer Menge, die gemeinsam behandelt wird
- Linguistische Analyseverfahren:
  - Heuristiken auf Wortformbasis: *–lichem* → ADJA
  - Zerlegung von Komposita, Suche nach den Köpfen im Taggerlexikon



- Ergänzung des Taggerlexikons, ggf. mit manueller Intervention
- Zweiter Tagging-Durchgang

Gojun et al. 2012, Arbeiten von S. Tannert

Nutzung von *TreeTagger* in DH-Aktivitäten

Tagging von Nicht-Standard-Texten: erste Schritte (1/3)

- *TreeTagger* und alle anderen Tagger für Deutsch
  sind zunächst an Zeitungstexten trainiert
  und geben daher für diese Textsorte optimale Ergebnisse
- Anderen Textsorten/Genres: Frühe Experimente — Giesbrecht/Evert 2009
  - Einfach zu taggen: "expository prose"
    Medizininformation, politische Reden, CeBIT-Nachrichten
  - Problematisch: Texte zu einer Fernsehserie, online-Foren-Texte
- Neuere Untersuchungen zur konzeptionell gesprochenen Texten:
  - Kiezdeutsch-Korpus — Rehbein et al. 2014, Wiese
  - Sprechsprachtexte im IdS-Korpus FOLK — Westpfahl & Schmidt: IdS
  - Internet-basierte Kommunikation — Beisswenger, Storrer u.a.
- Lernerkorpora — Lüdeling, Reznicek: Berlin, DFG-Netzwerk KOBALT
- NoSta-D-Korpus — Lüdeling und andere

Nutzung von *TreeTagger* in DH-Aktivitäten

Tagging von Nicht-Standard-Texten: Vorschläge (2/3)

- Arbeitsgruppe zur Annotation von Nicht-Standard-Texten
  Initiiert von Prof. Zinsmeister (Hamburg), 2012,
  Workshops 2012 und 2013 — Sonderband von JLCL, 2014
- HiTS (Historisches TagSet) für Frühneuhochdeutsch:
  Inspiriert von und abbildbar auf STTS — Dipper, Univ. Bochum
- Annotation von Texten der Barockzeit — Resch, Krautgartner: OeAW

Nutzung von *TreeTagger* in DH-Aktivitäten

Tagging von Nicht-Standard-Texten: Beispiele für Vorschläge (3/3)

- Vorschläge für Ergänzungen von STTS, z.B. für folgende Phn̈omene:
  - Kontraktionsformen: *biste ("bist du"), hamses ("haben Sie es")*, etc.
  - Selbstständige Interaktive Einheiten: TU Dortmund, IdS
    - Interjektionen: *ach; ok; oops; ...*
    - Responsiva: *ja; nein; ...*
    - Onomatopoetica: *krach; boing, miau, ...*
    - Emoticons: *:-) – ;-) – :-(– ...*
    - Aktionswörter: *lach; grins; lol; ...*
- Reklassifikationsvorschläge für Partikeln – Adverbien – Interjektionen

Alt	VVPP ADJD ADV	Partizip adverbiales Adj. nicht-flekt. Adv.	bedingt wahrscheinlich vermutlich	Neu	ADV	Satzgliedköpfe
	PTK.+ ITJ	gramm. Partikeln Interjektionen	ja, nein, nicht ach, mmhm, ...		PTK	Satzgliedteile

Zusammenfassung

Es wurde gezeigt, ...

- wofür
  POS-Tagging in den Digitalen Geisteswissenschaften einsetzbar ist
- welche Prinzipien und Wissensquellen dem *TreeTagger* zugrundeliegen
- wo Probleme liegen und ggf. Ergänzungsbedarf besteht

- dass POS-Tagging
  die interaktive und halbautomatische Suche in Texten
  generalisieren hilft
- dass *TreeTagger* eine übliche "Standardsoftware" ist

Weitere Arbeiten in Deutschland

- Weiterführung der Anpassung von STTS an Nicht-Standard-Texte
- Tests mit ergänzten bzw. modifizierten Tagsets
- Generellere Werkzeuge für die Ergänzung von Taggerlexika,
  Mehr zur Interaktion
  zwischen Tokenizing, Tagging und Named Entity Recognition
- Ausgehend von (i) Arbeiten zu Taggern mit einer Ausgabe von morphosyntaktischen Merkmalen
  und von (ii) syntaktischem Parsing:
  Arbeiten zur Nutzung von syntaktischer Analyse in der Wortartenannotation:
  syntaktisches Wissen hilft Wortart-Mehrdeutigkeiten auflösen

  RF-Tagger: Schmid/Laws 2008

  z.B. mate-Parser: Bohnet 2010

  Seeker/Kuhn 2014 etc.

$\Rightarrow$ POS-Tagging ist noch nicht zu den Akten gelegt,
kann aber mit ordentlicher Qualität benutzt werden