

Grenzen und Gemeinsamkeiten: Die Beziehung zwischen der Computer- linguistik und den Digital Humanities

Gerhard Heyer
Universität Leipzig
heyer@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG

Institut für Informatik



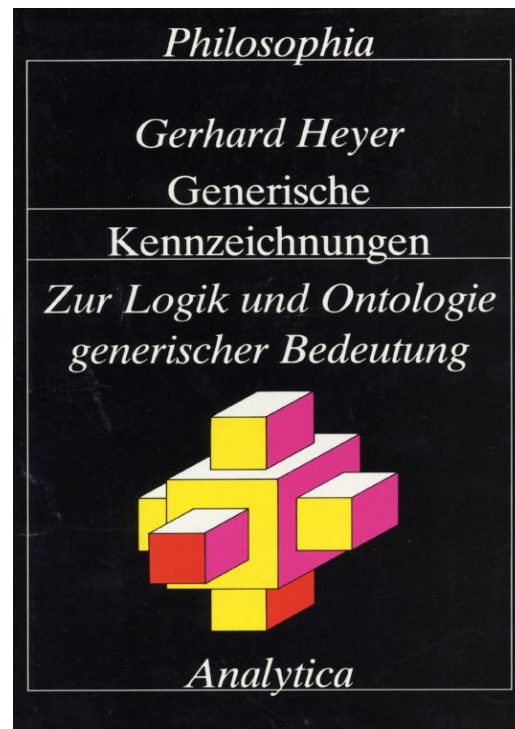
Kurzvorstellung ASV

- **Abteilung für Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig**
- **Teil der Angewandten Informatik**
- **Ziel ist nicht die Implementierung von linguistischen Theorien, sondern Anwendung von informatischen Modellen und Verfahren auf Sprache**
- **Schwerpunkt Text Datenbanken, Text Mining und Text basierte Forschungsinfrastrukturen**
- **Zahlreiche Projekt mit Industrie und in den *eHumanities***



Noch eine Bemerkung vorweg

- **Eigener Hintergrund auch in der Philosophie und Linguistik**
 - Philosophy Tripos, University of Cambridge
 - Promotion in formaler Semantik
(Mentor *Helmut Schnelle*)



Übersicht – Wie möchte ich argumentieren?

- 1. Die detaillierte Bereitstellung und Analyse von Text ist ein wesentliches Arbeitsmittel der (textorientierten) Geisteswissenschaften**
- 2. Die Berücksichtigung der linguistischen Dimension von Text erschließt diesen wichtige Interpretationsmöglichkeiten**

Zum Terminus „Geisteswissenschaften“ (*humanities*)

- **Wissenschaftshistorisch im frühen 19. Jh. als Oberbegriff der historisch-philologisch orientierten Wissenschaften entstanden**
- **Im Deutschen vor allem Abgrenzung gegenüber den Naturwissenschaften, „Verstehen“ in Abgrenzung zu „Erklären“**
- ***humanities* im Englischen bezeichnet „Wissenschaften vom Menschen“, insb. Handlungswissenschaften (*moral sciences*), und schließt damit an die traditionelle Unterscheidung zwischen *theoretischen* und *praktischen* Wissenschaften an**
- ***humanities* im Englischen schließt auch – anders als der Terminus „Geisteswissenschaften“ – die Sozialwissenschaften mit ein**

Zum Terminus „Geisteswissenschaften“ (*humanities*)

Geisteswissenschaften heute:

- „Lernen, was andere vor uns gesagt bzw. geschrieben haben“
- „Wissenschaft von der Interpretation tradierter Texte“

(S. Blasche, „Geisteswissenschaften“, Enzyklopädie Philosophie und Wissenschaftstheorie, S. 727, BI 1980)

- „Jene Disziplinen, die es mit dem Verstehen, Interpretieren, Erklären, Wiederherstellen, Prüfen und Deuten von „Texten“ im weitesten Sinne des Wortes zu tun haben.“

(Axel Horstmann, „Wozu Geisteswissenschaften? Die Antwort August Boeckhs“, Veröffentlichungen der Humboldt-Universität Berlin, H.93, S.26, 1998)

Aufgaben und Arbeitsweisen der Geisteswissenschaften

- (1) Sammeln und Bereitstellen von Texten/Textkollektionen**
- (2) Prüfen, Korrigieren, Wiederherstellen, Versionieren von Texten
(klassische Edition)**
- (3) Erklären, Interpretieren, Deuten und Verstehen von Texten**

Diese Aufgaben und Arbeitsweisen

- sind unabhängig vom Medium (analog oder digital)
- finden im *digitalen Medium* besonders gute Unterstützung

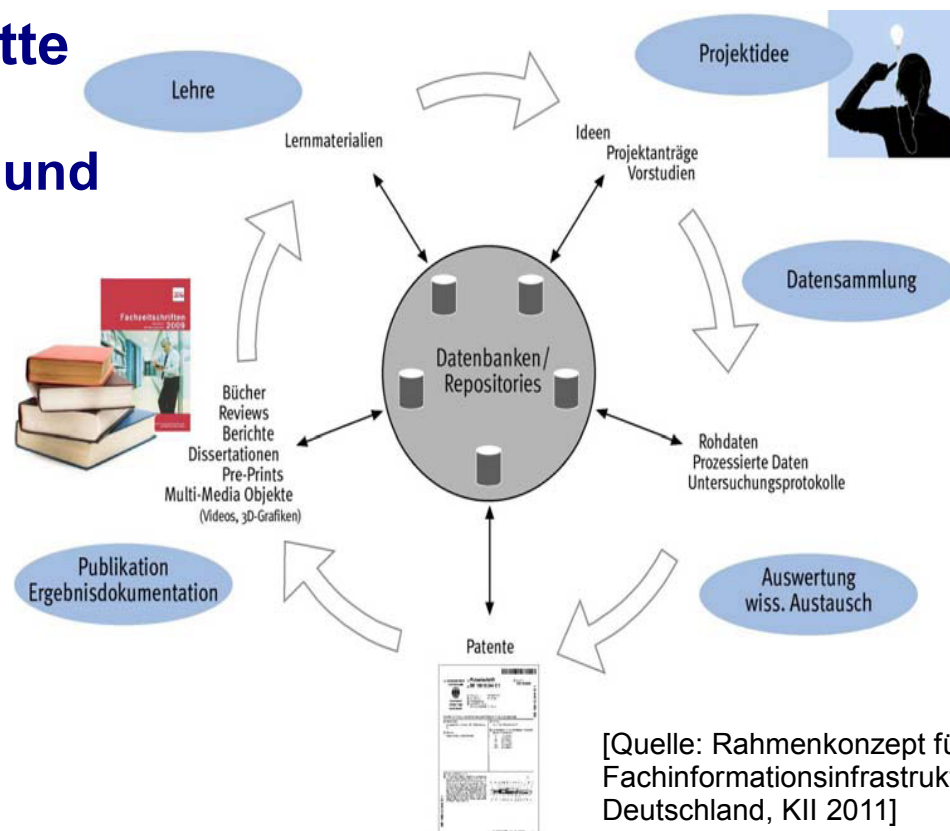
Methoden und Verfahren der *digital humanities* (Auswahl)

- **Aufbau von (verteilten) Textdatenbanken mit umfangreichen Metadaten**
- **Hierarchische und feingranulare Repräsentation von Text** (z. B. Functional Requirements for Bibliographic Records (*FRBR*), Canonical Text Services (*CTS*)) mit Rechten und Sichten
- **Für die Suche: IR-Verfahren, semantische u. explorative Suche, Metadaten basierte Suche** (z. B. CLARIN Component Metadata Infrastructure (*CMDI*)), ...
- **Für die Analyse: Text Mining, Information Extraction, Visual Analytics, ...**
- **Für die Annotation: Editoren, crowd sourcing, Versionierungen, ...**
- **Für die Verlinkung: linked open data, Netzwerkanalysen, ...**
- **Für die Archivierung: Referenzierungsstandards (z.B. DOI, PIDs), ...**

Forschungsinfrastrukturen

Forschungsinfrastrukturen unterstützen den gesamten Prozess der wissenschaftlichen Wertschöpfungskette

- **Bereitstellung, Austausch und Wiederverwendung von Daten und Verfahren**
- **Annotations-Plattformen**
- **eScience-Plattformen**



CLARIN-D

CLARIN-D – Eine web- und zentren-basierte Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Linguistische Daten, Werkzeuge und Dienste werden ...

- **in einer integrierten, interoperablen und skalierbaren Infrastruktur**
- **für die Fachdisziplinen der Geistes- und Sozialwissenschaften**

... bereitgestellt

- **gefördert durch das Bundesministerium für Bildung und Forschung**
- **Laufzeit: bis 31.05.2016**
- **im Web: <http://de.clarin.eu>**

MPI

Hamburg

BBAW

Leipzig



IDS

Saarbrücken

**Stuttgart
Tübingen**

München

GEFÖRDERT VOM



**Bundesministerium
für Bildung
und Forschung**

CLARIN-D

CLARIN-D – Eine web- und zentren-basierte Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Linguistische Daten, Werkzeuge und Dienste werden ...

- **in einer integrierten, interoperablen und skalierbaren Infrastruktur**
- **für die Fachdisziplinen der Geistes- und Sozialwissenschaften**

... bereitgestellt

- **gefördert durch das Bundesministerium für Bildung und Forschung**
- **Laufzeit: bis 31.05.2016**
- **im Web: <http://de.clarin.eu>**

MPI

Hamburg

BBAW

Leipzig



IDS

Saarbrücken

**Stuttgart
Tübingen**

München

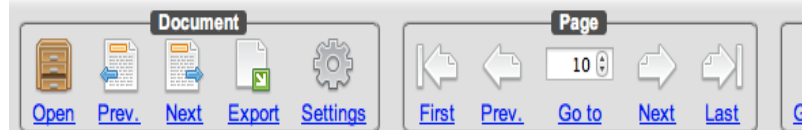
GEFÖRDERT VOM



**Bundesministerium
für Bildung
und Forschung**

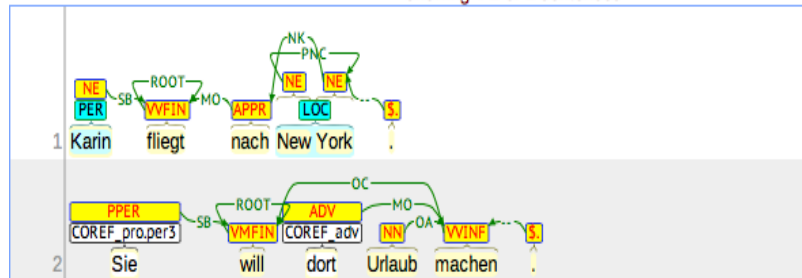
Linguistische Forschungsinfrastrukturen

Linguistisch annotierte Textdaten und Verfahren für deren (iterative) Erzeugung und Anwendung bei der Analyse von Texten



Test/tcf04-karin-wl.xml

showing 1-2 of 2 sentences



Technische Universität Darmstadt – Ubiquitous Knowledge Processing Lab – WebAnno | 0.3.0-SNAPSHOT

Open Science



HOME ABOUT USER HOME SEARCH CURRENT ARCHIVES

Home > No 1 (2013) > Kliegl

How Preview Space/Time Translates into Preview Cost/Benefit for Fixation Durations during Reading

Reinhold Kliegl, Sven Hohenstein, Ming Yan, Scott A. McDonald

Abstract

Eye-movement control during reading depends on foveal and parafoveal information. If the parafoveal preview of the next word is suppressed, reading is less efficient. A linear mixed model (LMM) re-analysis of McDonald (2006) confirmed his observation that preview benefit may be limited to parafoveal words that have been selected as the saccade target. Going beyond the original analyses, in the same LMM, we examined how the preview effect (i.e., the difference in single-fixation duration [SFD] between random-letter and identical preview) depends on the gaze duration on the pretarget word and on the amplitude of the saccade moving the eye onto the target word. The two key results were: (i) The shorter the saccade amplitude (i.e., the larger preview space), the shorter a subsequent SFD with an identical preview; this association was not observed with a random-letter preview. (ii) However, the longer the gaze duration on the pretarget word, the longer the subsequent SFD on the target, with the difference between random-letter string and identical previews increasing with preview time. A third pattern—increasing cost of a random-letter string in the parafovea associated with shorter saccade amplitudes—was observed for target gaze durations. Thus, LMMs revealed that preview effects, which are typically summarized under "preview benefit", are a complex mixture of preview cost and preview benefit and vary with preview space and preview time. The consequence for reading is that parafoveal preview may not only facilitate, but also interfere with lexical access.

Full Text: [PDF](#)

PID
11022/0000-0000-165F-E

PAPER PACKAGE
KlieglHohensteinYanMcDonald2012
(download)

USER
You are logged in as...
ojs
• [My Profile](#)
• [Log Out](#)

NOTIFICATIONS
• [View](#)
• [Manage](#)

JOURNAL CONTENT
Search

All

Browse
• [By Issue](#)
• [By Author](#)
• [By Title](#)

INFORMATION
• [For Readers](#)
• [For Authors](#)
• [For Librarians](#)

Wozu linguistische Forschungsinfrastrukturen?

In der Informatik ist Text vom Datentyp *string* – eine Aneinanderreihung (Konkatenation) von Elementen eines Alphabets (in einer Codierung). Die Reihenfolge ist meist nicht erheblich (*bag of words - Modell*).

Aber: In realen/historischen Texten ist die Konkatenation von Zeichen/Wortformen (über verschiedene Ebenen) *nicht* beliebig:

1. Jeder Text hat eine linguistische Struktur
(Syntax, Lexik, Morphologie)
2. Jeder Text steht in einem synchronen und diachronen Kontext
 - Verlinkung mit anderen Texten/Objekten
 - räumlich-zeitliche Evolution der Sprache

Nutzen linguistischer Forschungsinfrastrukturen

Die Berücksichtigung linguistischen Wissens verbessert

1. die Interpretation von Zusammenhängen
2. die Anwendung aufwändiger Text Mining Verfahren
3. die synchrone und diachrone Einordnung von Texten und einzelnen Textpassagen

Typische Fragestellungen:

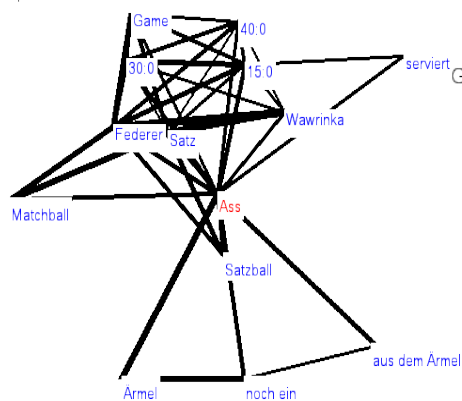
- diachrone/historische Semantik, Modellierung von Konzeptveränderungen, ...
- micro view Analysen, Nachweis seltener Belege, ...

Interpretation von Zusammenhängen

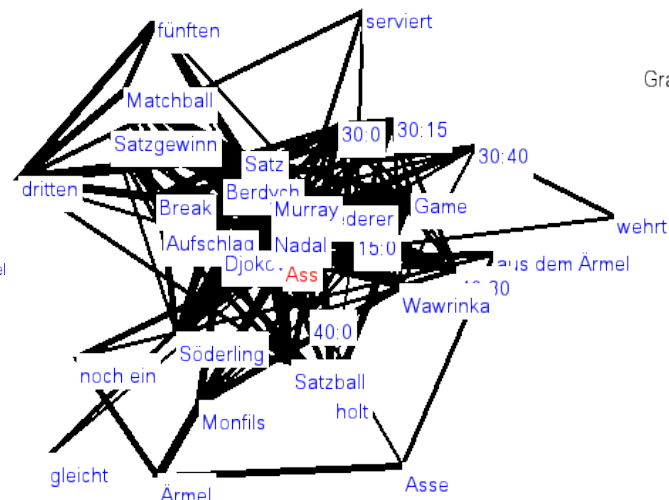
Auf Wortebene, z.B.

- (morphologische) Wortnormalisierung
- kookkurrenzähnliche Wortformen

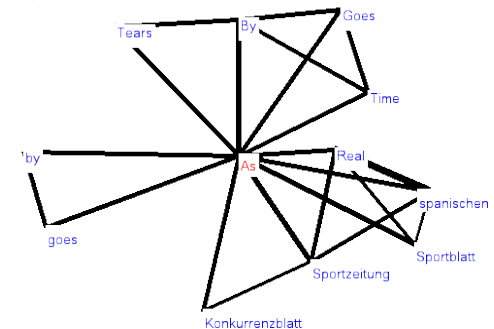
Graph v. 1.6 für Ass



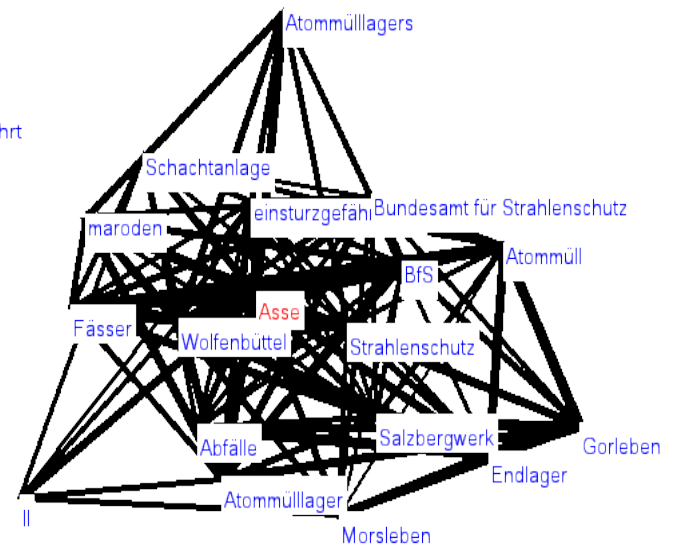
Graph v. 1.6 für Ass



Graph v. 1.6 für As



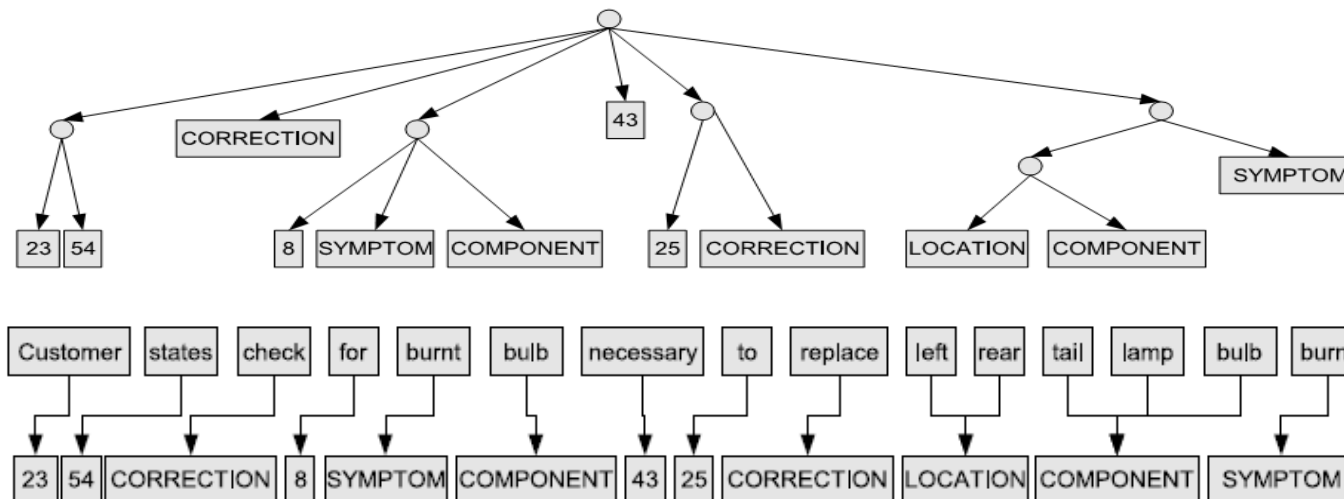
Graph v.1.6 für Asse



Text Mining Verfahren

Die Berücksichtigung linguistischen Wissens als *features* in einem Merkmalsraum für z.B.

- Autorenidentifikation
- *relation extraction* auf der Grundlage von POS-Tagging und (un-)überwachtem Parsing (ggf. auch *tree banks*)

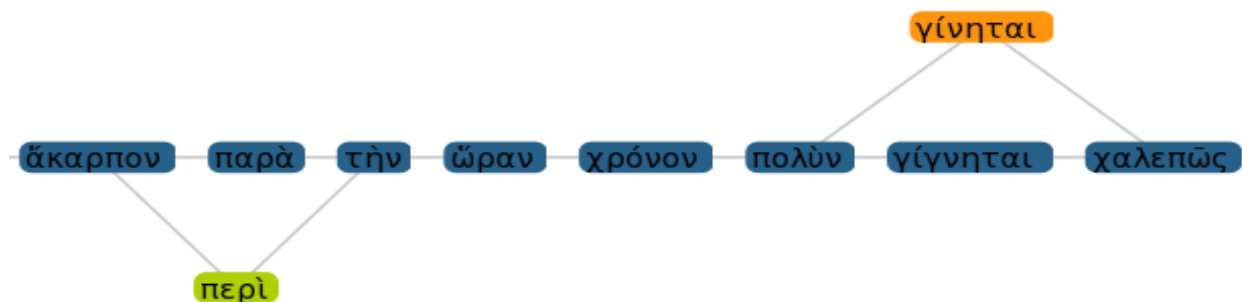


[Quelle: C. Hänig, Dissertation 2013]

Kontextualisierung

Synchrone und diachrone Einordnung von Texten und einzelnen Textpassagen wie z.B.

- Vokabularanalysen
- Zitationsanalysen



Quelle: Zitationsspuren von
Platons Timaios (Projekt eAQUA)

Zusammenfassung

- Linguistisches Wissen kann dazu beitragen, die *Qualität* von Text Mining Anwendungen und die *Interpretation* von einzelnen Texten und Textpassagen im synchronen und diachronen Kontext zu verbessern
- Die Bereitstellung von linguistischen *Daten* und *Analyseverfahren* sowie Umgebungen für deren Wiederverwendung, Annotation und den Austausch erfolgreicher Anwendungen ist eine zentrale Aufgabe linguistischer Forschungsinfrastrukturen

Zusammenfassung

„When the right kinds of annotations accumulate for particular texts, the realizable value of those texts can change radically. ... Corpus linguists have begun to transform our understanding of language by studying annotated corpora and these methods are — or must be — foundational to any modern philology: all students of historical languages are conducting corpus linguistics.“

(Crane et. al., Student researchers, citizen scholars and the trillion word library. In Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, S. 213-222)