

CLARIN-D M12 Workshop
27-28 Juni, 2012; Leipzig

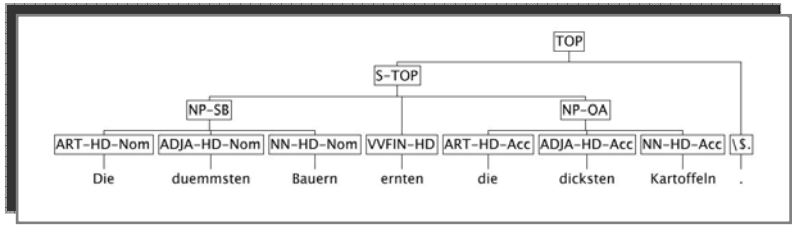
COMMON LANGUAGE RESOURCES AND TECHNOLOGY INFRASTRUCTURE

CLARIN-D

Corpus-based language analysis in research and teaching:

Experiences, requirements and perspectives

Angelika Storrer
technische universität dortmund



Corpus-based research projects

„Bericht zur Lage der deutschen Sprache“
[report on the state of the German language]

Joint project of the *Union der deutschen Akademien der Wissenschaften* (Union of the German Academies of Sciences and Humanities) and the *Deutsche Akademie für Sprache und Dichtung* (German Academy of Language and Literature).

Overall goal: Investigate selected aspects of the German language on the basis of digital text corpora and document the results for the general public.

Focus of first report: Development of German vocabulary in the 20th century.

Project corpus:
Three subcorpora with data from three time spans:

- 1905-1914
- 1948 -1957
- 1995-2004

Four genre types:

- fiction
- press
- science
- other non-fictional text types

(corpus data from the BBAW DWDS project and from the *Institut für deutsche Sprache* in Mannheim)

Corpus-based research projects

German Linguistics in Dortmund:
Subproject on support verb constructions

Support verb constructions:

make* a *decision*, *set* into *motion

eine Entscheidung treffen*, *in Bewegung setzen

Comparison of frequency and productivity of support verb constructions…

… between the three time spans of the project corpus

… between the four genre types of the project corpus

… between the project corpus and a corpus with legal texts
⇨ cooperation with Ulrich Heid / IMS Stuttgart

… between article pages and talk pages of a German Wikipedia Corpus
⇨ cooperation with Torsten Zesch / Ubiquitous Knowledge Lab (UKP), TU Darmstadt

Corpus-based research projects

Project context: Scientific network (DFG):
Empirische Erforschung internetbasierter Kommunikation
[Empirical research on internet-based communication]

empirikom
DFG Deutsche Forschungsgemeinschaft
www.empirikom.net

⇨ Corpus-based investigation of linguistic and interactional phenomena in computer-mediated communication (cmc)

Research on:

- **Style and register variation in the Dortmund Chat-Korpus**
  ⓘ www.chatkorpus.tu-dortmund.de
  🕮 Beißwenger & Storrer (2011)
- **Stylistic and structural features in *Wikipedia* articles and talk pages**
  ⇨ cooperation with Dr. Torsten Zesch / UKP, TU Darmstadt
  🕮 Storrer (2012)
- **Building a reference corpus of German cmc (*DeRiK*)**
  ⇨ cooperation with Dr. Alexander Geyken and Dr. Lothar Lemnitzer from the DWDS project at the BBAW)
  🕮 Beißwenger et al. (2012a)
- **TEI-compatible annotation schema for cmc genres**
  ⇨ cooperation with Dr. Alexander Geyken and Dr. Lothar Lemnitzer from the DWDS project at the BBAW)
  🕮 Beißwenger et al. (2012b)

Corpora and teaching

Courses related to research projects, e.g.:

-> Empirische Erforschung internetbasierter Kommunikation
(Beißwenger, SoSe 2007 & SoSe 2011)

-> Nominalisierungsverbgefüge des Deutschen
(Radtke, WiSe 2010/11)

Courses on (computational) lexicography

Hands-on sessions for creating *Wiktionary*-style dictionary entries (DWDS word profile; IDS cooccurrence profile):

-> Lexikographie im Internet (Storrer, SoSe 2011)

Courses on Corpus-based Language Analysis:

Basic methods and technical skills; tutored hands-on sessions with simple research questions (DWDS corpora, COSMAS, WebLicht)

-> Korpusgestützte Sprachanalyse (Storrer, SoSe 2008/2009, WiSe 2011/12)

Requirements

- ✓ One common interface with a German language version and German online tutorials
- ✓ Tools to further work with the results of search queries (clean-up and search again; manually annotate and search again; interface to statistic tools)
- ✓ Word sense disambiguation / semantic clustering or filter mechanisms
- ✓ Orthographic variation: example: *Stress / Streß* important issue when dealing with historical corpora or with computer-mediated communication
- ✓ Find interesting examples (e.g. metaphor and metonymy) in long KWIC lists

Example *Festplatte*

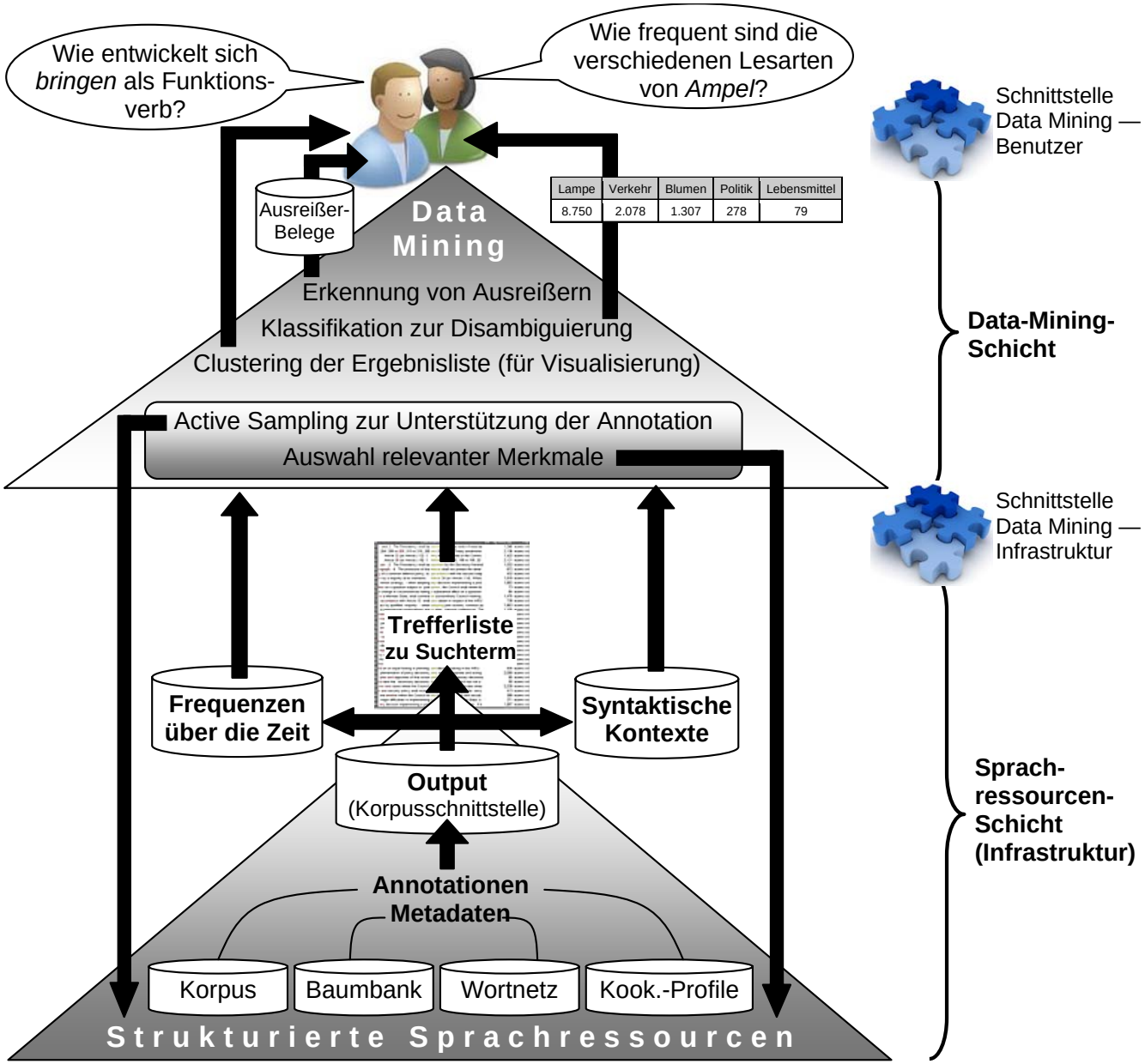
78 2008	.... Wenn ich in die USA reise, können die Behörden die	Festplatte	meines Computers kopieren. Wenn Finanzmärkte zusamme...
79 2008	... Vergangenheit heraufbeschwören und diese wie auf einer	Festplatte	abspeichern konnte (der Autor als früher Vorläufer des ...
80 2008	...iner Bank auszuspähen oder den kompletten Inhalt seiner	Festplatte	zu kopieren und an Dritte zu verkaufen. Otto N. wäre...
81 2008	...ich ist, dass er ihn in die Tasche stecken kann. Die	Festplatte	, auf der rund 6000 Lieder von 158 Künstlern gespeicher...
82 2008	...en noch von Diskette geladen. Erst das Aufkommen von	Festplatten	auch für den Hausgebrauch brachte grafischen Oberfläche...
83 2008	...d abgeben sollte «, sagt Lüngen. Also wertete er die	Festplatte	vor Ort aus und ließ den Laptop da. Christian kümmer...
84 2008	...m Mainzer Institut öffnet, » die Daten füllen jetzt die	Festplatten	des Zentralcomputers. Der bärtige Chemieingenieur is...
85 2008	...agesrhythmus Nachrichten öffentlich werden, die von der	Festplatte	eines Laptops stammen, der dem im März getöteten FARC-K...
86 2008	...ammlung von 4000 Pornofotos aus dem Internet auf seiner	Festplatte	archiviert und begibt sich immer wieder auf die Suche n...
87 2008	...en der Literatursuche im Netz, des Archivierens auf der	Festplatte	, des direkten Kontakts zu Kollegen in aller Welt. A...
88 2008	...inmal nachdenken" hat der Besserwisser nicht auf seiner	Festplatte	. Wenn du zu dem Besserwisser sagst: "Schönes Wetter...

DIE ZEIT, 21.05.2008, Nr. 22

Martenstein Keine Hechte in der Havel Besserwissertum und Sensibilität für Gesprächssituationen schließen einander aus, weiß unser Kolumnist aus eigener, leidvoller Erfahrung Von Harald Martenstein Den Besserwisser erkennt man unter anderem daran, dass er ungefragt zu allem sofort eine Meinung hat. Den Satz "Darüber müsste ich erst einmal nachdenken" hat der Besserwisser nicht auf seiner Festplatte . Wenn du zu dem Besserwisser sagst: "Schönes Wetter heute", dann antwortet der Besserwisser: "Letztes Jahr um diese Zeit war es aber drei Grad wärmer.

89 2008	... «, und das eigene Gehirn wird gern als » fragmentierte	Festplatte	« beschrieben. Auch die Macht von Führungskräften wi...
90 2008	..., was nach jedem Columbine oder Erfurt von der mentalen	Festplatte	purzelt: Virtuell gehts los, violent gehts weiter. R...
91 2008	...Informationen zu unterbinden, können alle Daten auf der	Festplatte	verschlüsselt werden. Auch das kopieren direkt am Ge...
92 2008	...leich sicherer ist es aber, wenn der Computer gar keine	Festplatte	hat, sondern wenn sie ihn jedes Mal von einer Dvd-Rom h...
93 2008	...u finden sind. Zumindest sollten sie alles auf ihren	Festplatten	verschlüsseln. Ungleich sicherer ist es aber, wenn d...
94 2008	...sschutz erlauben sollte, heimlich Computer auszuspähen,	Festplatten	zu verwanzen und vertraulichen E-Mail-Verkehr mitzulese...
95 2008	...richtet. Sie soll verhindern, dass der Staat auf den	Festplatten	seiner Bürger herumschnüffelt, auf welchem technischen ...
96 2008	...omputer Nackte Datensammelwut Wir speichern auf unseren	Festplatten	, was wir können. Nur: Wie lange noch?

Outlook


Wie entwickelt sich *bringen* als Funktions-verb?
Wie frequent sind die verschiedenen Lesarten von *Ampel*?
Schnittstelle Data Mining — Benutzer
Ausreißer-Belege
Lampe | Verkehr | Blumen | Politik | Lebensmittel
8.750 | 2.078 | 1.307 | 278 | 79
Data Mining
Erkennung von Ausreißern
Klassifikation zur Disambiguierung
Clustering der Ergebnisliste (für Visualisierung)
Active Sampling zur Unterstützung der Annotation
Auswahl relevanter Merkmale
Data-Mining-Schicht
Schnittstelle Data Mining — Infrastruktur
Trefferliste zu Suchterm
Frequenzen über die Zeit
Syntaktische Kontexte
Output (Korpusschnittstelle)
Annotationen
Metadaten
Korpus
Baumbank
Wortnetz
Kook.-Profile
Strukturierte Sprachressourcen
Sprach-ressourcen-Schicht (Infrastruktur)

Outlook

Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining (*KobRA*)

Prof. Dr. Angelika Storrer
(German Linguistics)

technische universität dortmund

Prof. Dr. Katharina Morik
(Computer Science)

Dr. Alexander Geyken
(BBAW)

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN
DWDS

INSTITUT FÜR DEUTSCHE SPRACHE
Mitglied der Leibniz-Gemeinschaft

Dr. Andreas Witt
Dr. Marc Kupietz
(IDS)

CLARIN-D | WEBLICHT
WEB-BASED LINGUISTIC CHAINING TOOL

Prof. Dr. Erhard Hinrichs
(SfS, U Tübingen)

References

Beißwenger, Michael (2012, in press): **Space in computer-mediated communication: Corpus-based investigations on the use of local deictics in chats**. In: Peter Auer & Anja Stukenbrock (Eds.): Space in language and linguistics: geographical, interactional and cognitive perspectives. Berlin. New York: de Gruyter.

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012a, in press): ***DeRiK:* A German Reference Corpus of Computer-Mediated Communication**. In: Proceedings of *Digital Humanities 2012*.

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012b, in press): **A TEI Schema for the Representation of Computer-mediated Communication.** In: Journal of the Text Encoding Initiative (TEI).

Beißwenger, Michael; Storrer, Angelika (2011): **Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen - Erfahrungen - Desiderate.** In: Journal for Language Technology and Computational Linguistics (Special issue „Language Resources and Technologies in E-Learning and Teaching“), 119-139.

Storrer, Angelika (2011): **Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie.** In: Knapp, Karlfried u.a. (Hrsg.): Angewandte Linguistik. Ein Lehrbuch. 3. Auflage. Tübingen: Francke Verlag, S. 216-239.

Storrer, Angelika (2012, in press): **Sprachstil und Sprachvariation in sozialen Netzwerken**. In: Barbara Frank-Job, Alexander Mehler & Tilmann Sutter (Hrsg.): Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW. Wiesbaden: VS Verlag für Sozialwissenschaften.

Preprints: http://www.studiger.tu-dortmund.de/index.php?title=Publikationen_von_Angelika_Storrer