# Geisteswissenschaftliche Anwendungen der Computerlinguistik

Seminar im Modul B-GSW-12 SoSe 2018

# Prof. Dr. Udo Hahn

Lehrstuhl für Angewandte Germanistische Sprachwissenschaft / Computerlinguistik

Institut für Germanistische Sprachwissenschaft

Friedrich-Schiller-Universität Jena

http://www.julielab.de

# Allgemeine Hinweise

- Termin:
  Mi, 10-12h (CZ3, SR 318)
- Materialien im Netz
  - http://www.julielab.de
     "Students"
- Sprechstunde: Mi, 12-13h (FG 30, R 004)
- © Email: udo.hahn@uni-jena.de
- Fachliteratur: überwiegend in Englisch

# Kontext: Digital Humanities (ohne Computerlinguistik)

- Primäre DH-Nutzer besitzen keine/wenig ausgeprägte Informatikexpertise
  - Installationen: BS-Abhängigkeiten
  - Programmierung bestenfalls in engem Rahmen, aber eher nicht
  - Out-of-the-box-Lösungen
- DH-Lösungen zielen im wesentlichen auf
  - Datenresourcen
    - Korpora, Lexika, Enzyklopädien
  - Portale
  - Annotationswerkzeuge
  - Vorkonfigurierte bzw konfigurierbare Systeme

# Datenresourcen: Korpora

- DeReKo (Deutsches Referenzkorpus)
  - Synchrones Referenzkorpus der geschriebenen neuhochdeutschen Sprache
  - Zeitungen, Belletristik, Handbücher, Parlamentsprotokolle (seit 1956)
  - Umfang: ca. 40 Mrd. Tokens
  - http://wwwl.ids-mannheim.de/kl/projekte/korpora/

# Datenresourcen: Korpora – DeReKo





#### Montageanleitung Seite 2 Rücksitzbezua



#### Montage des Rücksitzes

Beginnen Sie mit dem Sitzteil. Ermitteln Sie die Breite Ihres Rücksitz-Bezuges durch einfaches Auflegen der vollständigen Grundelemente auf den Originalrücksitz. So können Sie leicht feststellen, welche Mittelstreifen Sie entfernen müssen, um den Sitzbezug optimal dem Sitz anzupassen. Entfernen Sie die nicht benötigten Mittelstreifen durch Öffnen der Reißverschlüsse.

Der Rücksitzbezug ist werksseitig auf die maximale Breite eines Pkw-Rücksitzes vorgearbeitet.

#### Kombinationsmöglichkeiten

Durch Entfernen der Mittelstreifen ① oder ② und bei Kleinwagen ① + ② wird der Rücksitzbezug optimal der Sitzbreite angepasst.

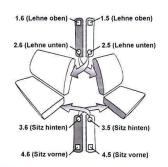
Meine Damen und Herren! Die letzte Frage, nämlich Frage 15, betrifft meine persönliche Homepage. - Ich hoffe, sie hat bei Ihnen Gefallen gefunden! Offensichtlich sind Sie eine der Kundschaften und schauen sie sich öfters an. Meine Damen und Herren! Das ist gut, denn sie hat wichtige Inhalte, die Sie aufnehmen können und unter Umständen auch transportieren sollten. (Beifall bei Abgeordneten der ÖVP und der Freiheitlichen.)

Allerdings möchte ich darauf verweisen, dass diese Frage nicht Gegenstand des Fragerechts gemäß § 90 Geschäftsordnungsgesetz ist. (Zwischenruf bei der SPÖ.) - Danke vielmals für das Signal! - Ich möchte aber festhalten, damit es hier keine Gerüchtebildung gibt: Selbstverständlich wird kein einziger Euro und kein einziger Cent meiner privaten Homepage mit Steuergeld finanziert. Das ist selbstverständlich nicht der Fall! Es wäre sehr plump, wenn ich Ihnen auf eine solche Frage etwas anderes sagen müsste. Natürlich ist diese Homepage privat und über Sponsoren finanziert.



#### Montage der Seitenteile

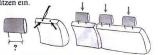
Trennen Sie die mitgelieferten Seitenteile durch vollständiges Öffnen der Reißverschlüsse und verbinden Sie jeweils ein Seitenteil mit dem Grundelement. (s. Skizze)



#### G Montage der Kopfstützenbezüge

Ziehen Sie die Kopfstützen-Bezüge über die Kopfstützen. (Wegen der Vielfalt der hinteren Kopfstützsysteme ist hier herstellerseitig keine Vormontage möglich.)

Markieren Sie die Öffnungen für die Kopfstütz-Halterungen auf der Oberseite des Lehnenbezuges. Schneiden Sie dort mit einem spitzen Messer kleine Öffnungen in den Bezug und setzen Sie die Kopf-



#### Hinweis

Sie haben ein Naturprodukt erworben. Kleine Struktur- und Farbunterschiede sind keine Fehler, sondern Beweise für die Echtheit

Damit Sie lange Freude an Ihrem Rücksitzbezug haben. verwenden Sie zur Reinigung bitte geeignete Pflegemittel

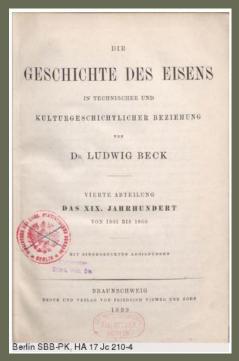
35 82 80 · E-Mail: Henckell-AG@T-Online.de

# Datenresourcen: Korpora

- DTA (Deutsches TextArchiv)
  - Historisches (diachrones) Referenzkorpus der neuhochdeutschen Sprache
  - Disziplinen- und gattungsübergreifende Bücher (1600-1900)
  - Umfang: ca. 112 Mio. Tokens
  - http://www.deutschestextarchiv.de/

# Datenresourcen: Korpora – DTA









# Datenresourcen: Enzyklopädien

- Wikipedia (deutsch)
  - Deutschsprachige Online-Enzyklopädie (Erklärungstexte, tabellarische Zusammenfassungen und Abbildungen (visuelle Daten)
  - Umfang: ca. 2.1 Mio Artikel
  - https://de.wikipedia.org/

# Datenresourcen: Enzyklopädien - Wikipedia



Hauptseite Themenportale Von A bis Z Zufälliger Artikel

Mitmachen

Artikel verbessern Neuen Artikel anlegen Autorenportal Hilfe Letzte Änderungen

Werkzeuge

Kontakt

Spenden

Links auf diese Seite Änderungen an Artikel Diskussion

Lesen

Quelltext anzeigen

Versionsgeschichte

Wikipedia durchsuchen

Nicht angemeldet Diskussionsseite Beiträge Benutzerkonto erstellen Anmelden

Q

#### Friedrich Schiller

(Weitergeleitet von Schiller)



Schiller ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind unter Friedrich Schiller (Begriffsklärung) und Schiller (Begriffsklärung) aufgeführt.

Johann Christoph Friedrich von Schiller (\* 10. November 1759 in Marbach am Neckar; † 9. Mai 1805 in Weimar), 1802 geadelt, war ein Arzt, Dichter, Philosoph und Historiker. Er gilt als einer der bedeutendsten deutschen Dramatiker, Lyriker und Essayisten.

Schiller wurde als einziger Sohn eines württembergischen Militärarztes, der später den Rang eines Hauptmanns erreichte, und der Tochter eines Bäckers geboren. Mit seinen fünf Schwestern wuchs er in Schwäbisch Gmünd, Lorch, später in Ludwigsburg auf. Dort besuchte er die Lateinschule und begann nach viermaligem Bestehen des Evangelischen Landesexamens am 16. Januar 1773 das Studium der Rechtswissenschaften auf der Karlsschule. Drei Jahre später wechselte er zur Medizin und wurde 1780<sup>[1]</sup> promoviert. Gleich mit seinem Theaterdebüt, dem 1782 uraufgeführten Schauspiel Die Räuber, gelang Schiller ein bedeutender Beitrag zum Drama des Sturm und Drang und der Weltliteratur.

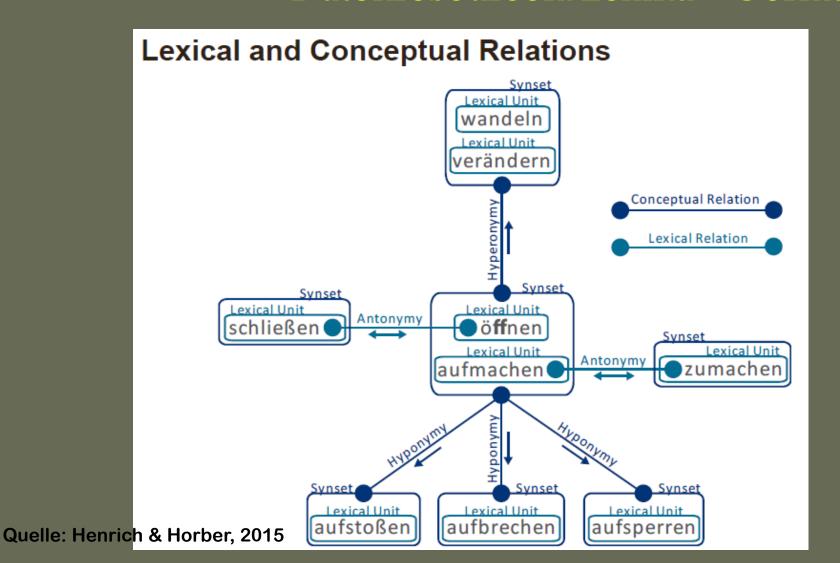


# Datenresourcen: Lexika

# GermaNet

- Breit abdeckendes deutsches WordNet-Äquivalent
- Semantisches Netz f
  ür deutsche Verben, Nomen,
   Adjektive, Adverbien (155k Einträge)
- Synsets und diverse semantische Relationen
- http://www.sfs.uni-tuebingen.de/GermaNet/

# Datenresourcen: Lexika – GermaNet



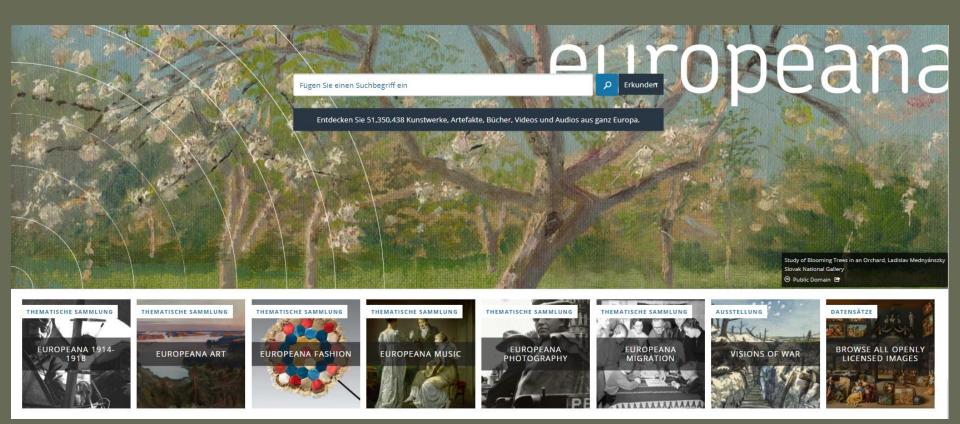
# **Portale**

# • Europeana

- Multimediale, multilinguale Digital Library zu europäischem Kulturerbe
- 27 Mio Bilder, 22 Mio Texte, 1,1 Mio Videos, 700k Audios,
- https://www.europeana.eu/portal/de

# Europeana

http://www.europeana.eu/portal/



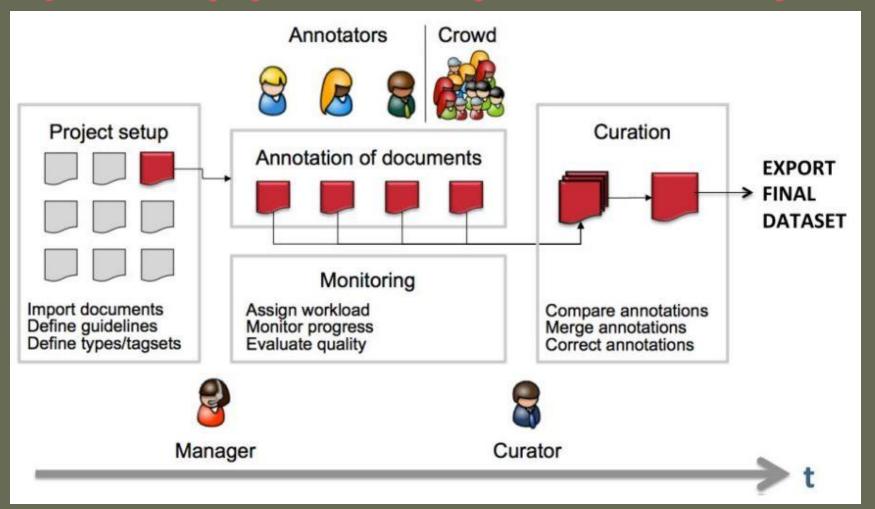
# Annotationssysteme

# WebAnno

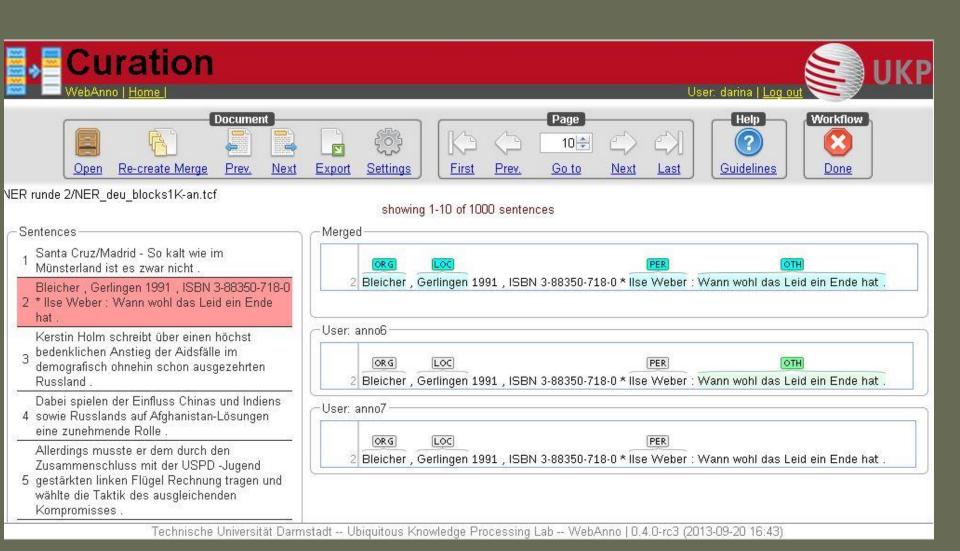
- Tool zur Unterstützung manueller Annotation
- https://webanno.github.io/webanno/

# Manuelle Annotation – Workflow für Metadatengewinnung

https://webanno.googlecode.com/svn/tags/latest-stable/docs/user-guide.html



# Annotationswerkzeug - WebAnno

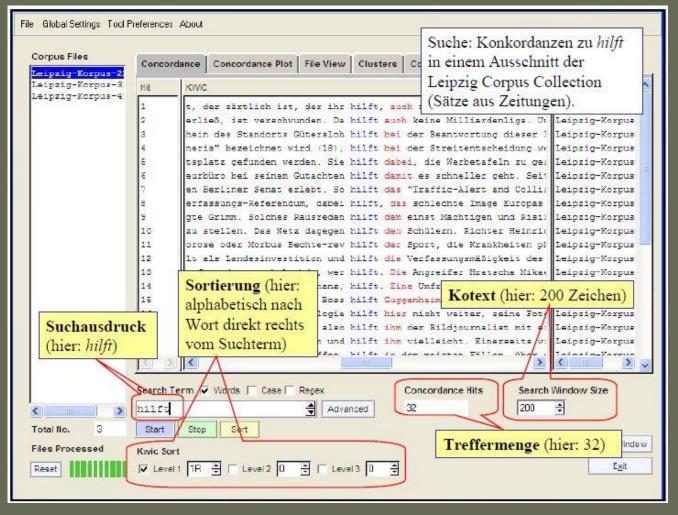


# Konfigurierte Systeme: Konkordanzer

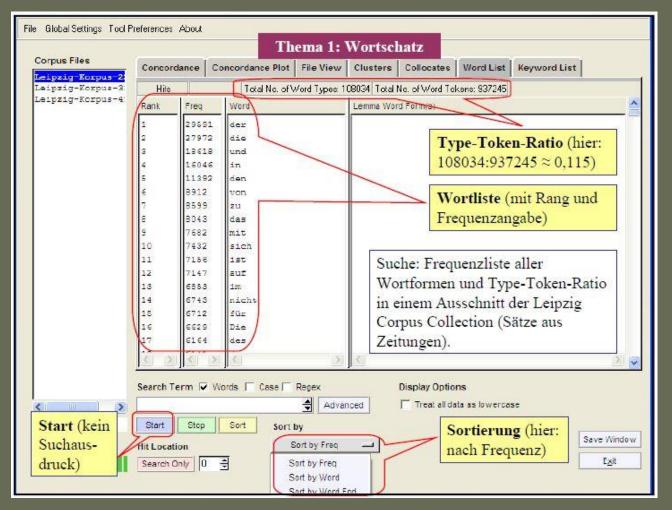
# • AntConc

- Automatische Konkordanzerstellung
- Einfache sprachstatistische Auswertungen
- http://www.laurenceanthony.net/software/antconc/

# Konfigurierte Systeme: Konkordanzer – AntConc



# Konfigurierte Systeme: Lexikalische Frequenzanalyse – AntConc



# Konfigurierbare Systeme: NLP-Pipelines

# • WebLICHT

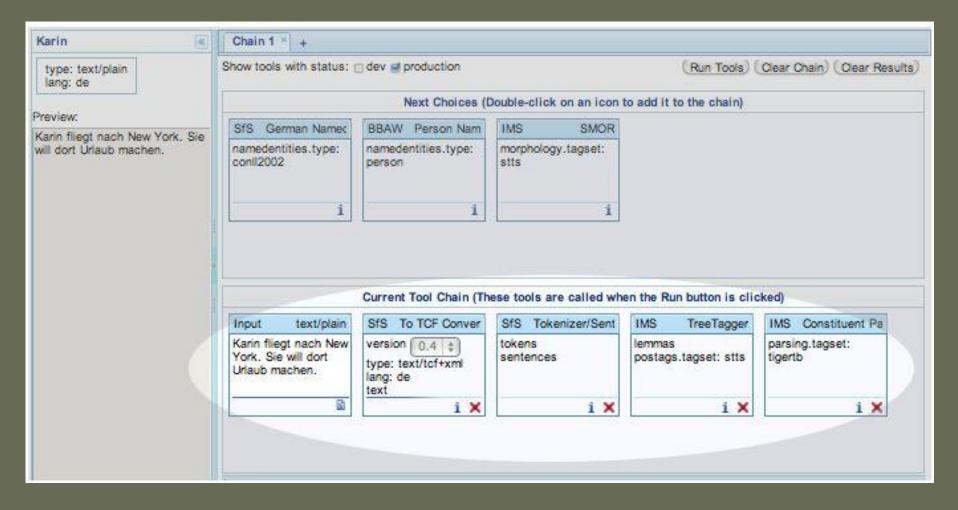
- Aufbau (komplexer) sprachanalytischer Systeme
- Konfiguration von CL-Komponenten
- http://weblicht.sfs.uni-tuebingen.de/ weblichtwiki/index.php/Main\_Page

# WebLicht: Komponenten für die automatische Annotation von Textkorpora (Computerlinguistik)

http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main Page

- Satzsegmentierer
  - Erkenner f
    ür einzelne S
    ätze
- Tokenizer
  - Erkenner für einzelne Textwörter
- POS-Tagger (part of speech)
  - Erkenner für die Wortart(en) einzelner Textwörter
- Morphologische Tagger und Lemmatisierer
  - Analyse morphologischer Merkmale, Erkenner für Lemmata
- Syntax-Parser und Chunker
  - Analyse der syntaktischen Struktur von Sätzen, Nominalgruppen-Erkennung
- WSD-Analyse (word sense disambiguation)
  - Disambiguierung einzelner Wortbedeutungen
- NE-Erkenner (named entities)
  - Erkenner für die semantischen Typen einzelner Textwörter
- Satz- und Wort-Alignierung
  - Erkenner für gleiche/ähnliche Sätze/Wörter in zwei verschiedenen Textressourcen

# Konfigurierbare Systeme: NLP-Pipelines – WebLicht



# Konfigurierbare Systeme: NLP-Pipelines – WebLicht



# Seminarleistungen

- Vortrag (mündlich)
  - 1-stündig
  - Elektronische Version (PDF, PPT) verfügbar machen
- Referat (schriftlich)
  - 15-20 Seiten Kerntext (mit Standardformaten)
  - Elektronische Version (PDF, DOC) verfügbar machen
  - Eidesstattliche Erklärung zur Eigenautorenschaft
    - Wir prüfen mit Plagiatserkennungs-Software
  - Abgabe: Ende Juli 2018

# Bemerkungen zu Referaten

#### • Aufbaumuster:

- Deck- bzw. Titelblatt mit vollständigen Angaben
- Inhaltsverzeichnis
- Einführung ins Thema, Motivation
- Themenabhandlung: grundlegende Formalisierungen, Verfahrensbeschreibungen (Algorithmen), Systemfunktionalitäten, Ressourcenmerkmale, Experimente/Evaluationen usw.
- Fazit mit kritischer Würdigung, offene Probleme ansprechen
- Bibliographie

#### • Zitationen:

- Alle verwendeten Quellen zitieren
  - Mit einem bibliographisch korrekten Zitat die jeweilige Quelle eindeutig beschreiben
  - Fachartikel <u>nicht</u> mit http://...foo.pdf-Link zitieren
  - Online-Quellen mit URLs und Datum des letztem Zugriffs
- Wikipedia ist keine zitierfähige wissenschaftliche Quelle!
- Eigenleistungen (Literatur, Beschäftigung mit konkreten Ressourcen/Systemen usw.) sind sehr erwünscht!!

# Wege zum Vortrag und Referat

- Email: Anmeldung von drei nach fallender
   Priorität geordneten Themenwünschen
  - First-come, first-served
- Email: Themenvergabe durch Dozenten
- Erste Literaturhinweise als "Saat" nach Bestätigung der Themenauswahl
- Themenbearbeitung durch Referenten
  - Mündlicher Vortrag zum vereinbarten Termin
  - Schriftliches Referat (unter Einhaltung der organisatorischen Verabredungen) zum vereinbarten Termin

# Ablaufplan

```
11.4. U. Hahn Einführung ins Thema
```

18.4. U. Hahn ----

25.5. U. Hahn ----

2.5. U. Hahn Themenvergabe

06.6. xxxxxxx yyy

13.6. xxxxxxx yyy

20.6. xxxxxxx yyy

27.6. xxxxxxxx yyy