Linguisticsweb.org

TECHNISCHE UNIVERSITÄT DARMSTADT

Ressourcen für die Ausbildung in den Digital Humanities

Am Beispiel nachhaltiger Vermittlung von Kompetenzen für die linguistische und philologische Sprachverarbeitung

Sabine Bartsch

Technische Universität Darmstadt

Institut für Sprach- und Literaturwissenschaft

URI: http://www.linglit.tu-darmstadt.de

E-Mail: bartsch@linglit.tu-darmstadt.de

Student team members and contributors:
Franziska Horn
Michael Hanl

Former team members:
Ella Syndikus
Zhi Chen


LinguisticsWeb
tutorials, corpora, references
Search
English
Log In
What's New?
Recent Changes
The Corpus Web
LinguisticsWeb
Contributors
Search LinguisticsWeb
List of Topics
Contact Us
Webs
Computerphilologie
LinguisticsWeb
Linguistics Tutorials
Beginner's Tutorials
Advanced Tutorials
You are here: LinguisticsWeb > LinguisticsWeb Web > WebHome (03 Feb 2012, MichaelHanl)
Welcome to the LinguisticsWeb
Welcome to the linguisticsweb, a web of information for students and researchers in linguistics
computational corpus http information language linguistics linguisticsweb methods students tutorials web
Linguisticsweb.org is a project aimed at supporting and furthering study and research in corpus and computational linguistics by means of providing students and researchers with a variety of materials that enable them to familiarize themselves with methods and techniques in linguistics, natural language processing and other, language related research.This website provides students and researchers in linguistics with tutorials, how-tos, links, tools, corpus access and other types of information to help them learn corpus and computational linguistic methods, techniques and technologies for text analysis.
The website is created in collaboration with students and for students. It is focused on methods and technqiues as well as materials used in teaching linguistics, corpus linguistics and computational linguistics at TU Darmstadt. Its aim is to allow students to explore methods and tools as taught in courses, but also aims to encourage students to try out state-of-the-art tools and techniques on their own.

Webs

- Computerphilologie
- LinguisticsWeb

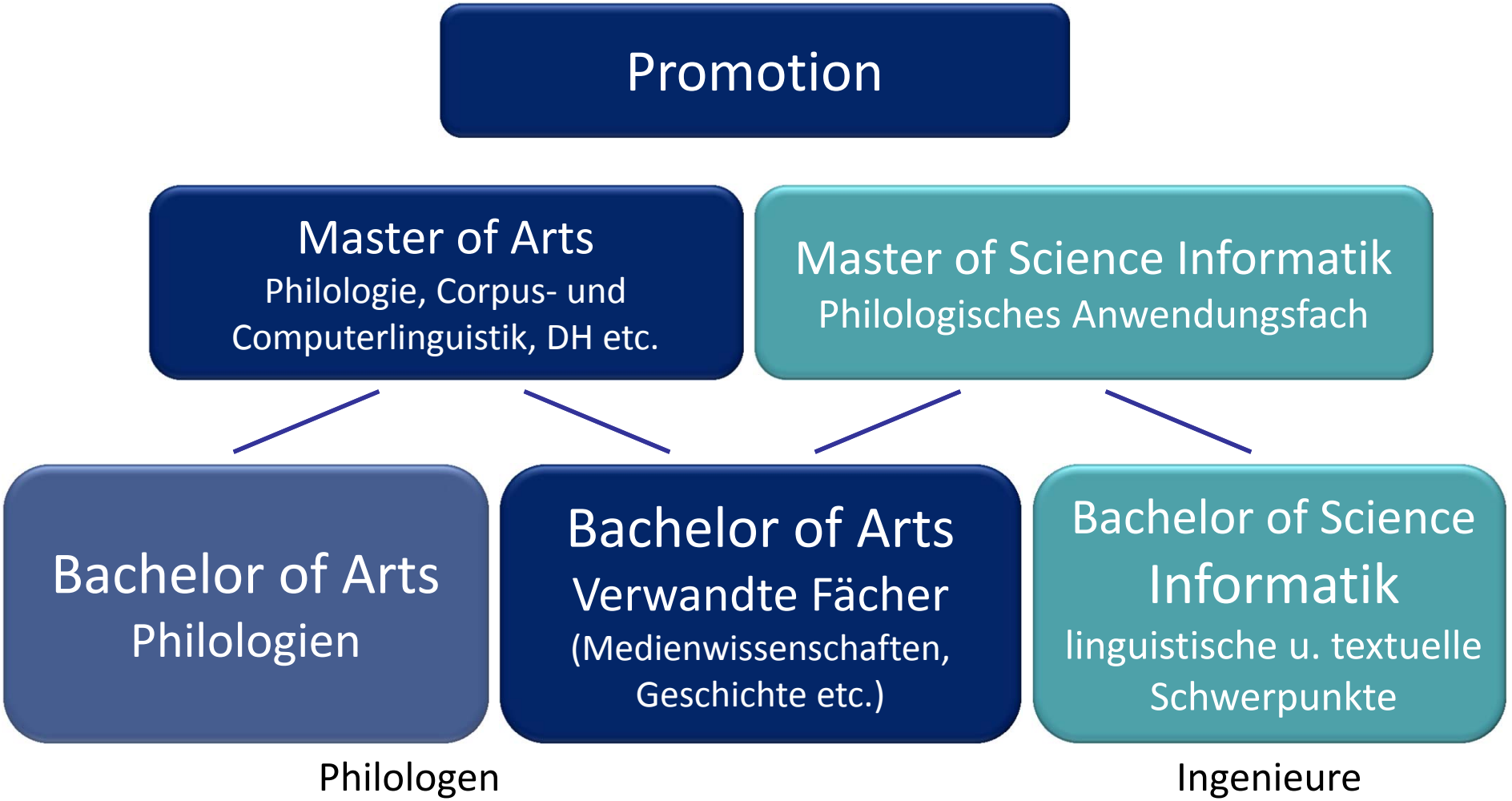
- Linguistic Tutorials
- Linguistic Glossary
- Linguistic Teaching
- External Resources
- Linguistic References

- Tutorials und How-tos
- Glossar linguistischer Termini
- Lehrmaterialen, Beispielinhalte
- Links zu externen Ressourcen
- Bibliographie

Rahmenbedingungen

- Zielsetzungen
  - Ziel-Studiengänge und Ziel-Gruppen
  - Zielkompetenzen und Erwartungshorizont
- Verwendete Sprachressourcen
- Ausbildung und F & L Transfer
- Erfahrungen und Lehren

Studiengänge & Zielgruppen



Zielkompetenzen auf BA Niveau: Beispiele

Stärker informatische Grundlagen	• Programmierung: Textmanipulation mit XSLT, Scriptsprachen • Grundlagen der Informatik: Datenstrukturen & Algorithmen
Fachnahe methodologische und technische Kompetenzen	• Textenkodierung und Verarbeitung (XML etc.) • Corpora, Corpusabfrage etc. • Empirische Methoden, „Händische“ Analysen • Methodologische Grundlagen
Philologische Grundlagen	• Corpuslinguistik und Computerphilologie • Sprachsystem und Sprachtheorie • Grundlagen der Sprach- und Literaturwissenschaft

Studentische Projekte

- Kollokationen in literarischen Texten zum Bsp. anhand der Werke von Charles Dickens
  - Basis: Bestehendes Corpus, Kollokationsstatistik
  - Statistische Kollokationsanalyse: Kollokationen von Begriffen aus dem Bereich der Körpermerkmale der zentralen und peripheren Protagonisten
  - Charakterisierung sog. „flacher“ vs. „runder“ Charaktere anhand von Kollokationen (z.B. *blue eyes, dark eyes sparkling eyes, brooding eyes, shiny eyes* etc.)

Studentische Projekte

- Vergleichende Analyse amerikanischer und russischer Präsidentenreden des frühen 21. Jh.
  - Vergleichscorpora: Nutzung des bestehenden Darmstädter ObamaSpeeches Corpus plus Aufbau und Annotation eines eigenen Vergleichscorpus russischer Präsidentenreden
  - Keywordanalyse, Themenanalyse etc.
  - Problematik: Umgang mit anderen Alphabetsystemen und Kodierungen; Werkzeuge für *languages other than English*, in diesem Fall Russisch

Herausforderungen

- Linguistische Grundbegriffe und Fragestellungen
- Operationalisierung
- Auswahl, Aufbau und Benutzung geeigneter Ressourcen (Corpora, Annotation, Query)

- Basale technische Fertigkeiten aufbauen
- Vorurteile über Technologie abbauen
- Betreuung und Beratung bei Installation
- Begleitung und Beratung der Projekte

Ziele im philologischen Bachelor

- Verständnis für linguistische Fragestellungen und deren Operationalisierung
- Auswahl geeigneter Daten und Werkzeuge
- Verständnis für den Aufbau der Daten
- Sicherer Umgang mit linguistischen Ressourcen
  - Notwendige Schritte zur Aufbereitung von Daten zur Annotation
  - Abfolge von Werkzeugen (Tokenisierung – Tagging usw.)
  - Geeignete Query-Szenarien und -Techniken einplanen

Ziele und Ressourcen

Bsp. Bachelor of Arts Anglistik

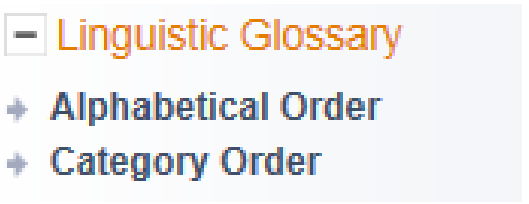
Corpora: Standardcorpora: British National Corpus, BROWN, LOB, FROWN, FLOB;
eigene Corpora, z.B. aus Texten des Oxford Text Archive

Annotation: Automatische Annotation: Tokenizer, POS tagger, Parser;
Manuelle Annotation → *stand-alone* Werkzeuge

Query: Frequenz, Konkordanz, lexikalische / grammatische Muster

- Philologische Grundlagen
- Ressourcen
- Literatur

Linguisticsweb.org



What's New?
- Recent Changes
- The Corpus Web

LinguisticsWeb
- Contributors
- Search LinguisticsWeb
- List of Topics
- Contact Us

Webs
- Computerphilologie
- LinguisticsWeb

+ Linguistics Tutorials

− Linguistics Glossary
- Alphabetical Order
- Category Order

+ Linguistics Teaching

+ External Resources

+ Linguistics References

Linguistic Glossary

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

This glossary is a ongoing compilation of important terms and concepts in linguistics. It is originally based upon the manuscript of an introductory linguistics books written at TU Darmstadt by Leslie Siegrist and Sabine Bartsch. Work on this glossary has been ongoing for some years and is going to continue to serve the information requirements of students of linguistics. The glossary aims to cover all terms introduced in introductory linguistics modules as well as more advanced terms. The alphabetical entry into the list of terms if probably the most straightfoward approach, however, all terms have been classified according to the field of linguistics they originate in (phonetics & phonology, morphology, syntax, semantics etc.) as well as their rooting in specific theories of linguistics, e.g. Systemic Functional Linguistics.

A

Affective Meaning (Semantics)

Affixation (Morphology)

Allomorph

Affixation Hide

Hyperonym to:
- Prefixation
- Suffixation
- Infixation

(Morphology)

Allophone (Phonology)

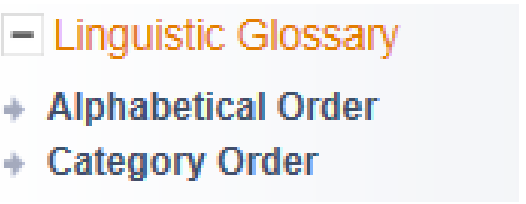
Allophonic Transcription (Phonology)

Alphabetisation (Morphology)

Alveolar Ridge (Phonology)

Annotation (Computer linguistics)

Linguisticsweb.org



What's New?
- Recent Changes
- The Corpus Web

LinguisticsWeb
- Contributors
- Search LinguisticsWeb
- List of Topics
- Contact Us

Webs
- Computerphilologie
- LinguisticsWeb

+ Linguistics Tutorials

− Linguistics Glossary
- Alphabetical Order
- Category Order

Glossary Search by Category

In order to display glossary entries per category, please select a category and press "Search":

please select a category | Search

Phonetics

Traditional phonetics - the study of the physical / material (acoustic, auditory, articulatory) aspects of speech sounds; the scope of modern phonetics does not allow such a rigid restriction to the physical aspects of language but includes functional aspects in the scope of phonetics as well

Articulatory Phonetics

Auditory Phonetics

Articulatory Phonetics Hide

Branch of phonetics concerned with the organs involved in the production of speech sounds; describes and classifies speech sounds based on the organs of speech involved in their production and the manner of their production; note: the human organs of speech have other primary functions in breathing and consumption of food, language production is only a secondary function.

-- SabineBartsch - 22 Jul 2011

Collocation analysis

Contents:

Collocations are a pervasive feature of native-like language use in any language. Any given stretch of language contains and is comprised of a large number c less fixed and recurrent word combinations that allow little or no variability in terms of the lexical items involved. The following set of examples illustrates the phenomenon:

- *confirmed bachelor*
- *handy tool*
- *give a talk*
- *take a bath*
- *attend a meeting*
- *commit a crime*
- *commit to memory*

Definition of collocation

A number of attempts at defining collocation as a linguistic phenomenon are evident from the literature. The following information is a condensed version of the theoretical background of the concept of collocation as discussed in Bartsch, 2004. *Structural and Functional Properties of Collocations in English.* Tübingen:

The oldest dates back to a Harris in 1750 who uses the term in the following context that is to be found in the earliest attested use of the term on the Oxford E Dictionary (OED, Second Edition):

1750 Harris Hermes ii. iv. Wks. (1841) 197 The accusative .. in modern languages .. being subsequent to its verb, in the collocation of the words.

This definition is closer to what we might call colligation today, i.e. the preference of certain lexical items or classes of lexical items for particular grammatical

Is it *fully*, *completely* or *totally* aware?

Validating collocation candidates using statistical measures

Analysis by Franziska Horn

Contents:

1 Collocations in the context of foreign language learning

Is it *fully*, *completely* or *totally* aware? Which of these phrases is the more common one or are all phrases widely used and accepted within the English langu

This question refers to the challenge that learners of a language are confronted with. Each of these phrases contains the adjective *aware* and an adverb whic indicates a state of high awareness. *Fully*, *completely* and *totally* share a similar meaning which illustrates a common problem for non-native speakers: Often no apparent reason for learners why one expression is more adequate than another but collocation-related mistakes are obvious for native speakers.

The website Englishclub.com presents fully aware as an example for a collocation. To prove this assumption we analyze our three collocation candidates usi

External Resources

There are several resources available for students and learners of linguistics alike, which can be used for teaching, as well as bases for linguistic analysis. The following page summarizes some of the most prevalent and important resources the LinguisticsWeb and external sources have to offer. To get an overview over available corpora, see Corpus Overview.

database der dictionary en frame german http langua lemid linguisticsweb online py res semantic trier uni universit wbgui www wörterbuch

Contents

Dictionaries

Resource	Organisation	Language	Description
Canoo.net	Canoo AG in collaboration with the University of Basel, the Vrije Universiteit Amsterdam and the IDSIA Lugano	DE	Online language service which offers information about word mean German Spelling, inflection, word formation and morphology, datab includes approximately 250 000 entries.
Deutsches Rechtswörterbuch (DRW)	Heidelberg Academy of Sciences and Humanities	DE	Dictionary of legal terms with approximately 90 000 entries. The DR aims to cover a time period starting from the beginning of written cu until 1800.
Deutsches Wörterbuch by Jacob and Wilhelm Grimm	Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften at the University of Trier, in collaboration with the Berlin-Brandenburg Academy of Sciences and Humanities	DE	Online edition of the Deutsches Wörterbuch (DWB) by Jacob and Wilhelm Grimm, lists approximately 300 000 entries.
DWDS	Berlin-Brandenburg Academy of Sciences and Humanities	DE	Lexical database combining several dictionaries: DWDS (Digitales Wörterbuch der deutschen Sprache) which is based on the Wörterl der deutschen Gegenwartssprache (WDG), Deutsches Wörterbuch

Linguistics References

Search over the entire References by using the **Reference-Search-Topic**

Introduction to Systemic Functional Linguistics

The presented references consist of basic readings for students. It is primary aimed for students during their first year in linguistics to get an insight into the conecpt and theories of Systemic Functional Linguistics and terms.

[Eggins04]

Suzanne Eggins. *An Introduction to Systemic Functional Linguistics*. Continuum International Publishing Group, London, 2004.

[Halliday04]

Michael Halliday. *Introduction to Systemic Functional Grammar*. Oxford University Press, 2004.

Corpus Linguistics

Understanding the construction, usage and analysis of corpora is essential to new areas and fields of linguistics. References concerning usage, analysis of data and adjacent fields of research are stated within this collection.

[Aijmer09]

Karin Aijmer. Parallel and comparable corpora. In Anke Lüdeling and Merja Kytö, editor, *Corpus Linguistics. An International Handbook*. de Gruyter, Berlin, 2009.

[McEneryXiao07]

A. M. McEnery and R. Z. Xiao. Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: Translation and the Linguist (Translating Europe)*. Multilingual Matters Ltd, Clevedon, UK, 2007. [.pdf]

Tutorials

The collection of all references used throughout the tutorials of the LinguisticsWeb.

LinguisticsWeb
tutorials, corpora, references

Linguistic Tutorials
- Beginner's Tutorials
- Advanced Tutorials

Basics

Generic Tools

Text editors, programming editors and a variety of other generic tools are useful to support linguistic analysis tasks. A set of useful tools is introduced here. More...

Regular Expressions

Regular expressions (regex) are patterns that can be used for querying strings of text. A basic understanding of them enables you to formulate complex queries to analyze corpus data. More...

System Requirements

Many software tools in corpus and computational linguistics have some basic requirements that have to be fulfilled as a prerequisite such as a Java Develpment Kit or other programming language environments. More...

Windows Shell Tutorial

The Windows Shell, also known as the Command Line, offers access to commands and tools that are not accessible via the graphical user interface, but have to be called by means of command line parameters. More...

LinguisticsWeb
tutorials, corpora, references

Linguistic Tutorials
- Beginner's Tutorials
- Advanced Tutorials

Taggers

QTag

QTag is a language independent, pure probabilistic part-of-speech tagger written by Oliver Mason (2008). More...

StanfordPoSTagger

The Stanford PoS Tagger is a probabilistic part of speech tagger developed by the Stanford Natural Language (NLP) group. More...

TreeTagger

The TreeTagger is a statistical part of speech tagger; it was develped by Helmut Schmid at the Institut für Maschinelle Sprachverarbeitung at the University of Stuttgart. More...

Manual annotation tools

EXMARaLDA

EXMARaLDA can be described as a system for computer assisted transcription and annotation of spoken language. It can be used for construction and analysis of spoken language corpora. More...

MMAX2

MMAX2 is a tool for manual annotation of relations between textual units. More...

UAMCorpusTool

The UAM Corpus Tool comprises a set of tools for linguistic annotation of texts which can be done manually and semi-automatically. More...

Query

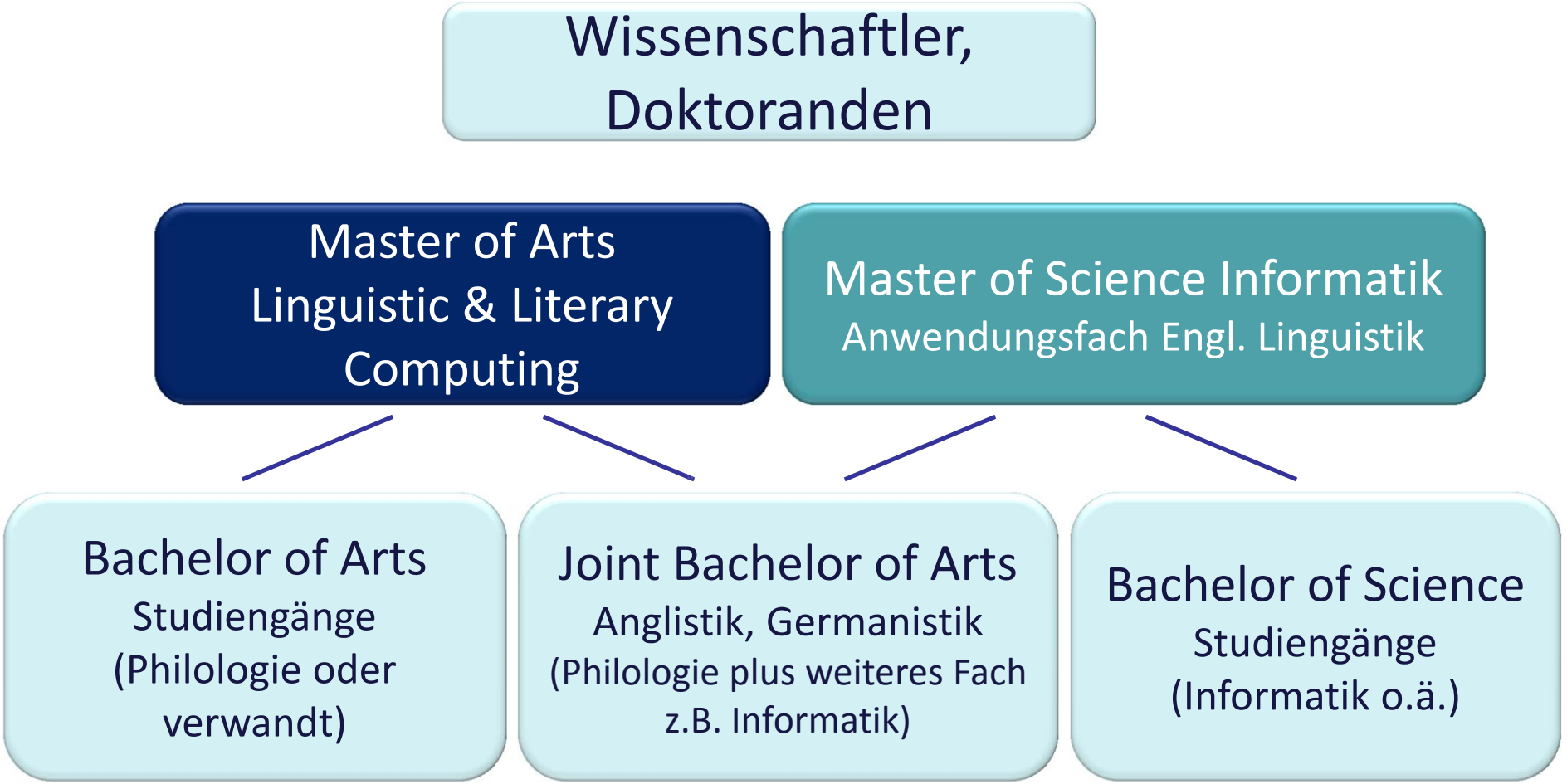
Concordancers

Concordancing is one of the most basic tasks in corpus inspection and analysis as it allows a selective and structured view on corpus data. More...

IMSOpen Corpus Workbench

The IMS Open Corpus Workbench provides corpus storage and query functionality based on a very efficient corpus index and a powerful regular expression language. More...

Studiengänge & Zielgruppen



Zielkompetenzen: Bsp. Master of Arts Linguistic & Literary Computing

Informatik	• Einführung in die Allgemeine Informatik • Java Programmierung • NLTK (NLP mit Python)
Corpus- und Computer-linguistische / Computerphilologische Seminare	• Registerlinguistik / Diskurslingustik • Corpora und probabilistische Verfahren • Computerphilologie (Edition, Lexikographie) • XML-Familie (XML, XSLT; TEI)
Corpus- und Computer-linguistische / Computer-philologische Grundlagen	• Anwendungen (Annotation, MT, IR/IE, Diskursorganis.) • Ressourcenaufbau • Techniken und Werkzeuge • Fortgeschrittene Annotationsaufgaben
Philologien	• Sprachsystem und Sprachtheorie • Empirische Methoden

Verteilung auf Lernergruppen

	MA Linguistic & Literary Computing	MSc Informatik, Anwendungsfach Engl. Linguistik
Corpora:	Standardcorpora plus Aufbau eigener Corpora, z.B. Obama Speeches Corpus, Literarische Corpora	
Annotation:	**abstraktere Phänomene** Diskursphänomene, z.B. (semi-)automatische Annotation von Kohäsion oder Thema-Rhema	
Komplexere Annotation:	Processing chains, Pipelines; „roll-your-own“; Multilayer Annotation (Exmaralda, MMAX2, TEI)	
Query:	Erweiterte Kenntnisse, multilayer Query (Exmaralda, MMAX2) auch Programmierung mit Python & NLTK, XSLT	

Studentische Projekte

- Automatically detecting gender allocation in A.L. Kennedy‘s „Failing to fall“
  - Formulierung von operationalisierbaren Kriterien für die Genderzuordnung der Protagonisten
  - Aufbau des Corpus und Annotation geeigneter Merkmale
  - Task: Mehrebenenannotation → diverse linguistische Merkmale → diverse Werkzeuge
  - Auswertung und Visualisierung der Merkmale (mit xslt)

Studentische Projekte

- ObamaSpeechesCorpus – Aufbereitung:
  - Html
  - Plain text
  - XML
  - GATE DataStore
- POS, Parsing, RST, Kohäsion
- Software: Little Cohesion Helper
  - Werkzeug, das auf Basis von NLTK, WordNet und MMAX2 lexikalische Kohäsion automatisch annotiert und manuell nachbearbeitbar macht

ObamaSpeeches.com: Building and Processing a Corpus of Political Speeches A student project

Sabine Bartsch, Stefania Degaetano, Tomasz Grubba, Nina Petrychka, David Sullivan, Christoph Tragl, Claudio Weck

Institut für Sprach- und Literaturwissenschaft, Hochschulstrasse 1, 64289 Darmstadt, URL: http://www.linglit.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Introduction

This poster presents a student project aiming at integrating annotation tools for a discourse analysis of a corpus of speeches by US President Barack Obama. The project entails corpus collection, encoding, annotation and query.

The linguistic aim of the project is to learn more about the characteristics of a set of political speeches in terms of established register features (Biber 1988, 1995) as well as their discourse structure in terms of topic development within speeches, use of cohesive devices (Halliday & Hasan 1976), rhetorical structure modeled on the basis of Rhetorical Structure Theory (RST) (Mann & Thompson 1987) and thematic development (Matthiessen 1995, Halliday 2004).

Issues are the interplay between different tools in light of heterogeneous data formats, and the integration of automatic annotation procedures as pre-processing steps for manual annotation tasks.

Aim: development of a processing chain that allows the linguist to explore the relevant properties of the corpus at different levels of linguistic organization.

Approach: integration of automatic and manual annotation tasks by means of NLTK.

2. The ObamaSpeeches Corpus (OSC)

120 speeches by US-President Barack Obama

Time span: 2002-2009.

Source: www.ObamaSpeeches.com

Source format: html

Derived formats for linguistic processing:

- plaintext
- html
- XML (TEI P5)
- GATE data store

3. Methods: Multi-level corpus annotation

Annotation requirements:

- Corpus metadata
- Tokenization
- Part of speech tagging
- Cohesive chains
- Rhetorical structure
- Thematic structure

Data format: multi-layer standoff

Tools explored:

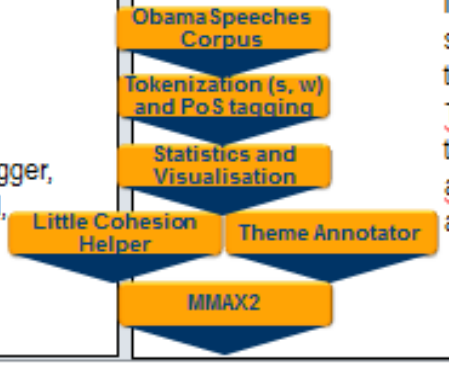
- Stand alone tools (Decision Tree Tagger, Theme Annotator, UAM Corpus Tool, MMAX2, etc.)
- GATE
- Natural Language Toolkit (NLTK)

3. Multi-level corpus annotation (ctd.)

Tool	Feature evaluation
Stand alone tools	+ powerful, reliable - tool integration - heterogeneous data
GATE	+ well-integrated - usability (new tools) - stability + homogeneous data
NLTK	+ powerful, flexible + tool integration + usability + homogeneity possible

4. Adopted approach

The Natural Language Toolkit (NLTK) is used as a basis for an implementation of automatic annotation steps whose output is prepared for further manual processing with MMAX2.



NLTK enables the integration of many standard annotation tools (e.g. the Punct-Tokenizer, Unigram tagger) as well as an api to resources such as WordNet.

4. Automatic support for manual annotation: The Little Cohesion Helper

As an example module developed with NLTK, the Little Cohesion Helper is presented here. Based on the NLTK / Python interface to WordNet, the Little Cohesion Helper (LCH) (Weck, Tragl 2009), this tool was developed to automatically identify and annotate cohesive ties in free text and prepare the output for further manual processing.

MMAX2 is the tools of choice for the annotation of cohesion, a task that has previously be shown to be amenable to automatic support on the basis of resources such as WordNet (Teich, Fankhauser 2006).

LCH integrates all pre-processing steps such as tokenization, pos-tagging with cohesion annotation . LCH produces as its output an MMAX2 project that allows further manual processing (see Figure x.x). It produces statistics on different types of cohesive relations, distance of relations and chain length (see Fig. x.x).

4. The Little Cohesion Helper (ctd.)

Fig. 1: LCH GUI and code

LCH can be used on the command line or Python's IDLE or through a GUI.

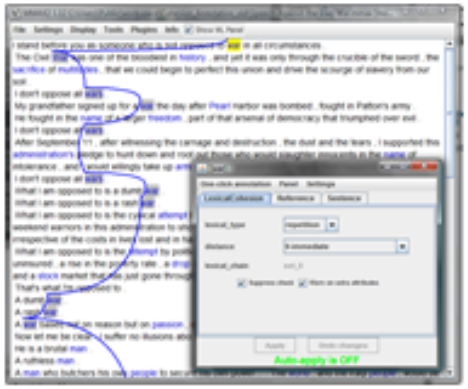


Fig. 2: MMAX project by LCH

The user can select all types of cohesive ties described in Halliday & Hasan (1976) for identification.

5. Additional features and future work

NLTK is also used for basic text statistics and visualizations thereof e.g. as a wordcloud.

Thematic structure annotated automatically by means of the Theme Annotator (Schwarz et al. 2008) can also be integrated into MMAX2 projects.

Query of the data currently proceeds by the MMAX2 query & statistics facilities. In the future, ANNIS2 will be employed to hold the data and allow for more advanced query.

References

AnnoLab,: http://www.annolab.org

ObamaSpeeches.com URL: http://www.obamaspeeches.com

MMAX2 URL: http://mmax2.sourceforge.net/

NLTK. URL: http://www.nltk.org

Steven Bird, Ewan Klein, and Edward Loper. 2009. **Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit.** O'Reilly.

Halliday, MAK, Ruqaiya Hasan. 1976. **Cohesion in English.** Harlow: Longmen.

PAULA Interchange Format for Linguistic Annotations, URL: http://www.sfb632.uni-potsdam.de/~d1/paula/doc/

4 Annotation with MMAX2

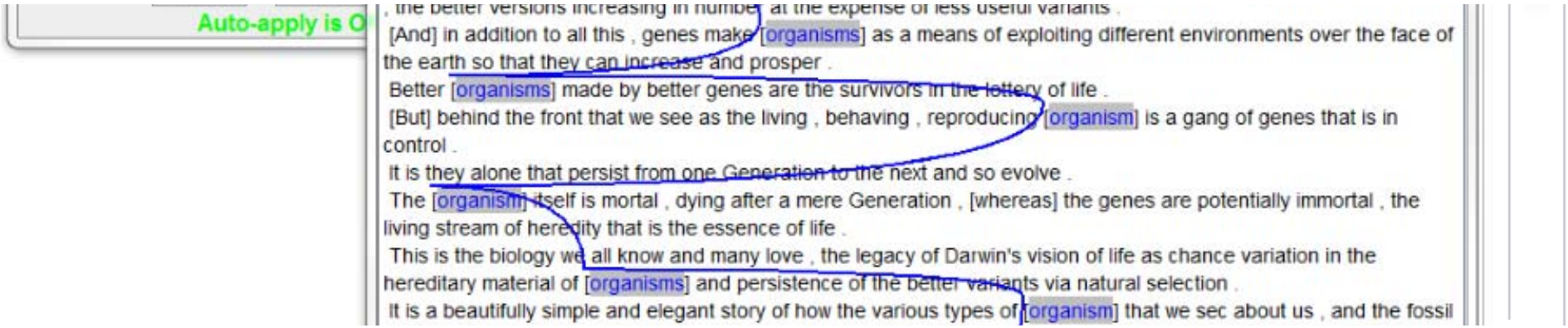
5 Summary

MMAX2 is especially suited for discourse annotation tasks. Its main advantages are its flexibility and configurability as well as the visualization of the discourse relations. Also, its configuration and annotation files are stored in xml, so are amenable to further processing outside of MMAX2 and are human readable. Having said that, its configurability is also its greatest obstacle that makes it difficult for inexperienced users to use the tool out of the box. It is therefore a good idea to start students and other first time users off with an example of a finalized project to ge a feel for what it is capable of.

References

[MullerStrube06]

Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197-214. Peter Lang, Frankfurt a.M., Germany, 2006. [DOI]



Beginner's Guide to WordFreak & the OpenNLP Tools

How-to by Michael Hanl

add annotation annotations batch bibliome corresponding create detection endnote english file files folder http iks jar linguistic load main menu models named need open opennlp pos project select sentence set step text tool tools use want window wordfreak

Contents

▼ About WordFreak and the OpenNLP-tools | ▶ Installation | ▶ Working with WordFreak

1 What are WordFreak and the OpenNLP Tools?

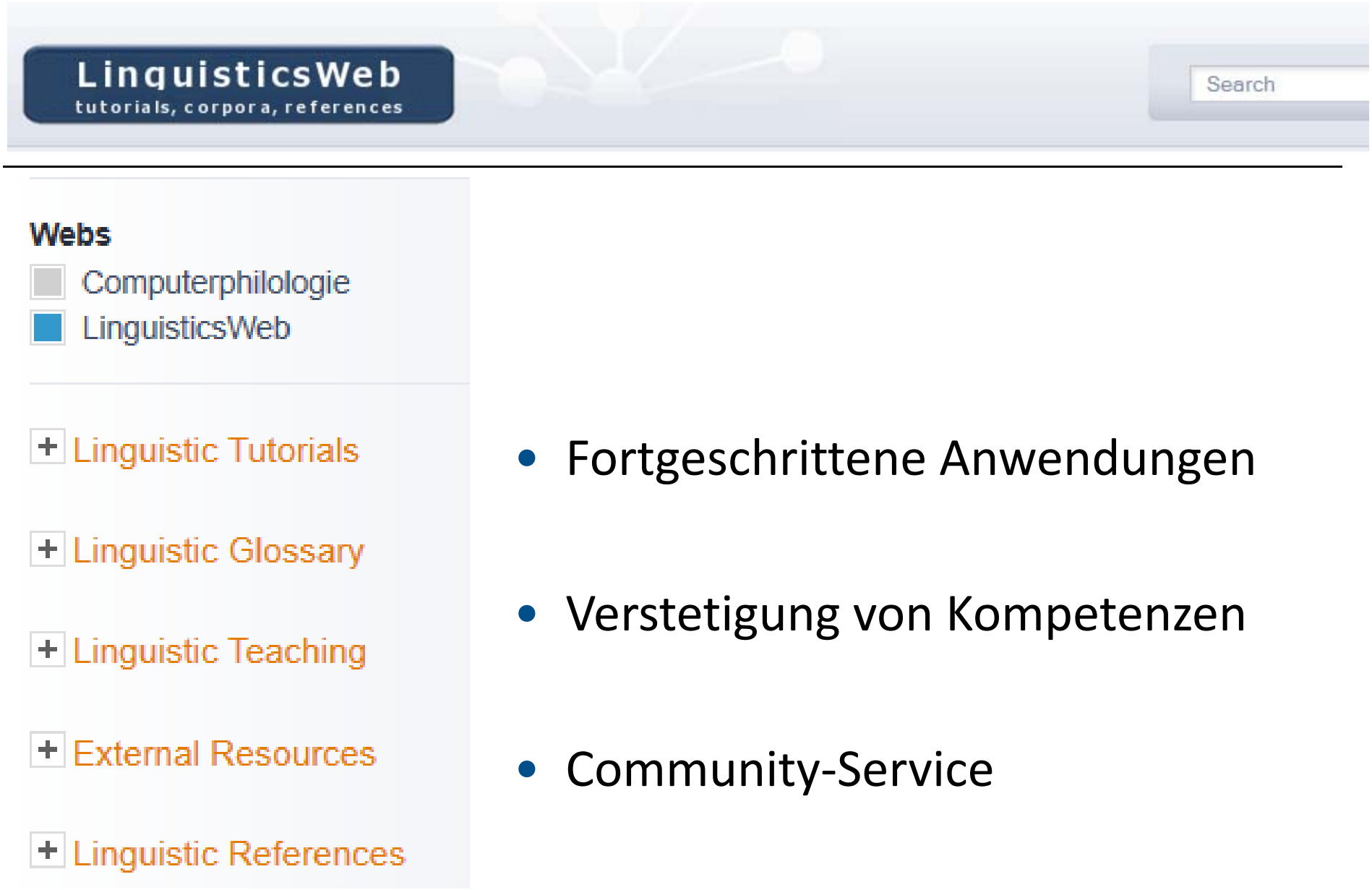
*"*WordFreak *is a java-based linguistic annotation tool designed to support human, and automatic annotation of linguistic data as well as employ active-learning for human correction of automatically annotated data. It provides basic functionality to check automatic annotation manually or create annotations into the interchangeable format XML. But if you want work with a more advanced set of tools than basic POS-Tags, you need to extent* WordFreak. *One set of*

Zielkompetenzen im Master

- Entwicklung eigener Workflows
- Entwicklung von Spezifikationen und Prototypen (DH-Studenten)
- Fähigkeit zum Aufbau eigener Sprachressourcen
- Durchführung gemeinsamer Seminarprojekte und ggf. Publikation
  Bsp.: Theme-Annotator (Schwarz et al. 2008),
  LittleCohesionHelper (Tragl & Weck 2009) plus
  Obama Speeches Corpus (Bartsch et al. 2009)
- Gemischte Gruppen: LLC und Informatiker

Vorteile der gemischten Gruppe

- Zusammenführung unterschiedlicher Ausgangskompetenzen
- Einüben von gegenseitigem Verständnis
- Interdisziplinäre Kommunikationsfähigkeit
- Simulation der Teamstruktur in Forschungs-projekten (Linguisten / Philologen plus Informatiker)



- Fortgeschrittene Anwendungen
- Verstetigung von Kompetenzen
- Community-Service

Basics

Generic Tools
Regular Expressions
System Requirements
Windows Shell Tutorial

Taggers

QTag
StanfordPoSTagger
TreeTagger

Parsers

Stanford NER
Stanford Parser
Stanford Tregex

Frameworks

GATE
WordFreak-OpenNLP

Manual annotation tools

EXMARaLDA
MMAX2
UAMCorpusTool

Taggers

TreeTagger-Adv

Parsers

Frameworks

LinguisticsWeb
tutorials, corpora, references

Linguistic Tutorials
- Beginner's Tutorials
- Advanced Tutorials

Parsers

Stanford NER

The Stanford Named Entity Recognizer (NER) developed by Stanford University (2002-2006) and licensed under the GNU GPL is an application which locates and classifies named entities in a text into predefined categories. More...

Stanford Parser

The Stanford Parser is a statistical natural language parser from the Stanford Natural Language Processing Group. Used to parse input data written in several languages such as English, German, Arabic and Chinese it has been developed and maintained since 2002, mainly by Dan Klein and Christopher Manning. More...

Stanford Tregex

TregEx is a platform independent tool from Stanford University for querying syntactic pattern in tree data structures parsed by Stanford Parser. More...

Frameworks

GATE

The basic idea behind Gate is to facilitate the processing of data by using different processing tools like tokenizers, taggers and parsers as a framework to join different processing types and level out divergences in output and input formats. More...

WordFreak-OpenNLP

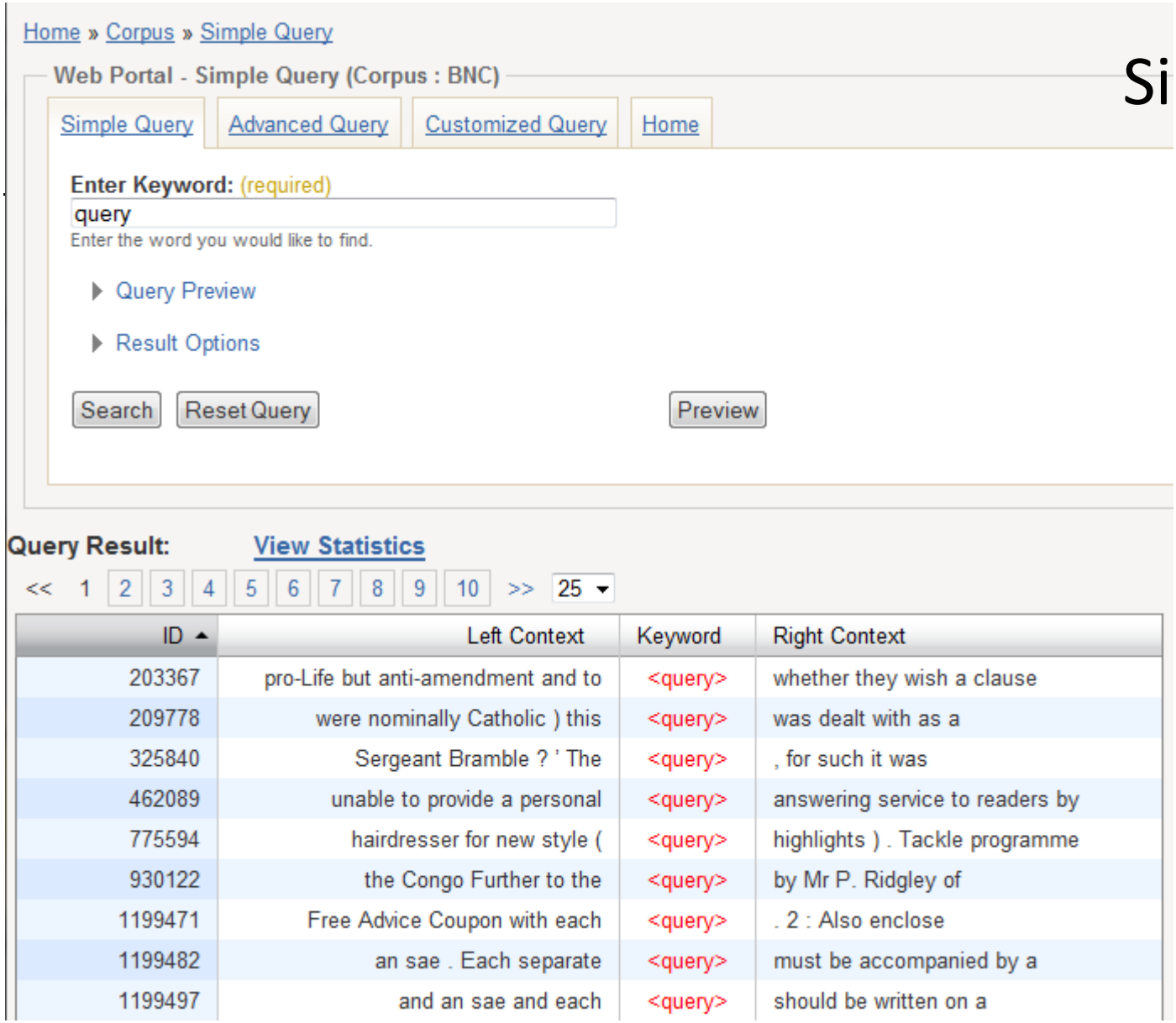
WordFreak is a java-based linguistic annotation tool designed to support human and automatic annotation of linguistic data as well as employ active-learning for human correction of automatically annotated data. More...

WEB CWB PORTAL

IMS Open CWB Interface für die Lehre:

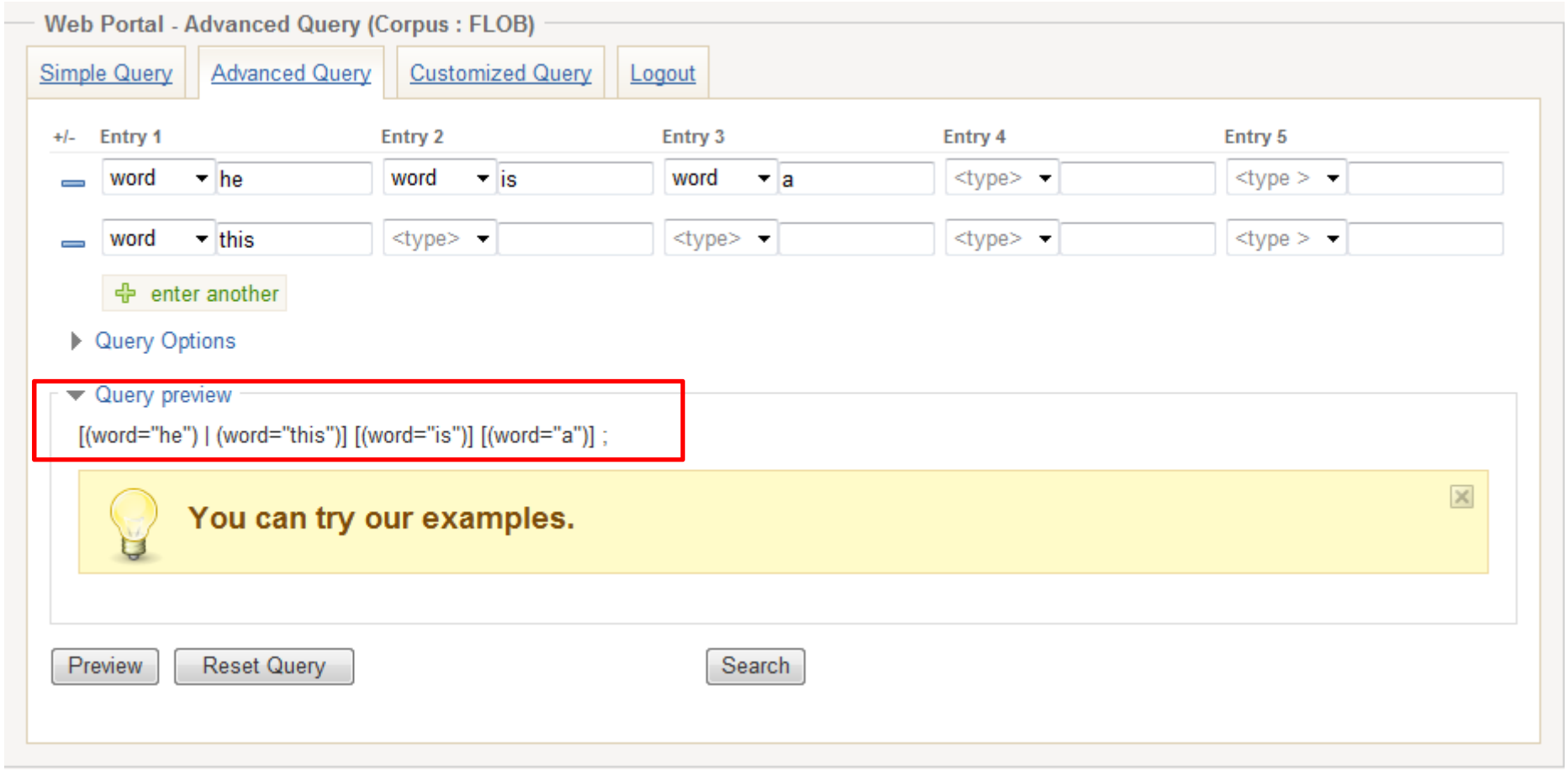
- Bereitstellung Copyright-geschützter Corpora für die Lehre
- Didaktische Unterstützung beim Erlernen einer Abfragesprache wie CQP

Simple query

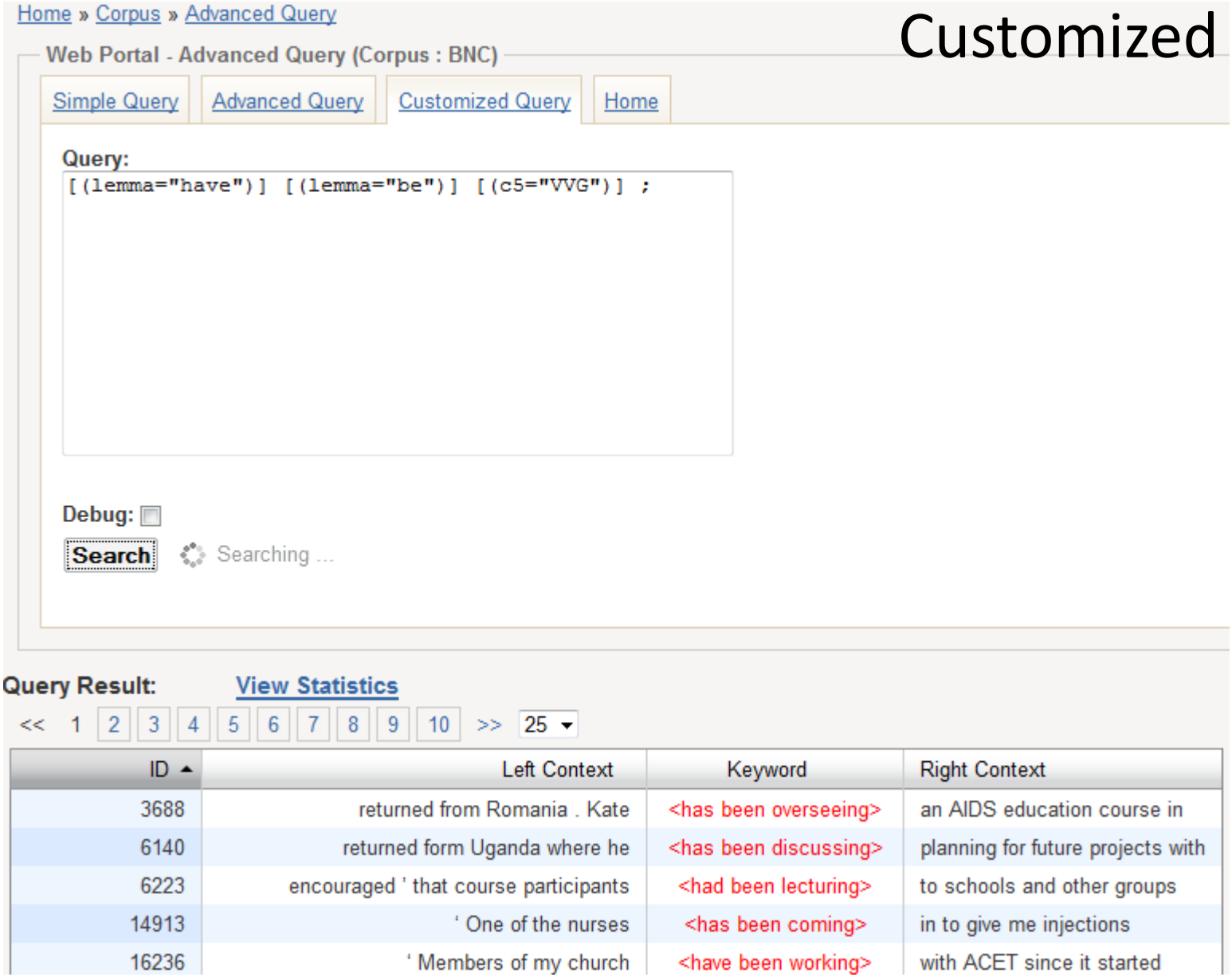


ID	Left Context	Keyword	Right Context
203367	pro-Life but anti-amendment and to	<query>	whether they wish a clause
209778	were nominally Catholic) this	<query>	was dealt with as a
325840	Sergeant Bramble ? ' The	<query>	, for such it was
462089	unable to provide a personal	<query>	answering service to readers by
775594	hairdresser for new style (	<query>	highlights) . Tackle programme
930122	the Congo Further to the	<query>	by Mr P. Ridgley of
1199471	Free Advice Coupon with each	<query>	. 2 : Also enclose
1199482	an sae . Each separate	<query>	must be accompanied by a
1199497	and an sae and each	<query>	should be written on a

Advanced query



Customized query



ID	Left Context	Keyword	Right Context
3688	returned from Romania . Kate	<has been overseeing>	an AIDS education course in
6140	returned form Uganda where he	<has been discussing>	planning for future projects with
6223	encouraged ' that course participants	<had been lecturing>	to schools and other groups
14913	' One of the nurses	<has been coming>	in to give me injections
16236	' Members of my church	<have been working>	with ACET since it started

Erfahrungen und Lehren

- Philologische Fragestellungen vor Werkzeugen
- Frühes Kennenlernen empirischer Methoden an kleinen, manuell annotierten Corpora
- Freie Zugänglichkeit der Werkzeuge und Daten (im CIP-Pool **und** von ausserhalb der Universität; Werkzeuge auch auf persönlichen Rechnern)
- Volltextzugang zu allen Ressourcen
- Synergien durch gemischte Gruppen aus Linguisten / Philologen und Informatikern
- Erhöhter Aufwand pro Lehrveranstaltung

Erwartungen und Wünsche

- Handling (Lehrende / Studierende)
- Wartbarkeit: Installation und Service
- Nachhaltigkeit: Wiederverwendbarkeit , gesicherte Verfügbarkeit von Software, Daten, Dokumentation
- Zugänglichkeit: Lösungen für Copyright / Lizenzgebühren / Plattformunabhängigkeit
- Verstetigung von Kompetenzen (auch bei Lehrenden)
- Ausbau der Lehrressourcen in der Methoden-ausbildung (institutionell, technisch und personell) in den *digital humanities*

Computerphilologie Web

In der Beschreibung der Computerphilologie wird unter dem Begriff Digitial Humanities die Gesamheit derzeitiger Bestrebungen der Datenvernetzung bzw. Aufbereitung und Digitalisierung von Texten verstanden. Der Gegenstand dieser Bestrebungen befasst sich im Genaueren mit infrastrukturellen Fragen der Bereitstellung, dem Auffinden und dem Abrufen von Daten für die philologische Gemeinschaft, als auch die systematische und strukturierte Aufbereitung von Daten hinsichtlich ihrer speziellen Erfordernisse. Diese Projekte, die teilweise mit DFG-Mitteln gefördert werden, sind zum Beispiel:

- TextGrid
- Dariah (Digital Research Infrastructure for the Arts and Humanities)
- Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen
- Wörterbuchnetz.de: digitale Nachschlagewerke, Verzeichnisse (Enzyklopädien, Wörterbücher, etc.)

Die Computerphilologie ist Teil des noch jungen Forschungsbereichs Digital Humanities, worunter diverse Forschungsaktivitäten subsumiert werden, die mit der Verarbeitung digitaler Daten in verschiedenen geisteswissenschaftlichen Disziplinen assoziiert werden. Dazu gehören neben der Digitalisierung und digitalen Erfassung von Daten und Artefakten, die Gegenstand geisteswissenschaftlicher Forschung sind, auch die Reflexion der Veränderung wissenschaftlicher Praktiken in der Analyse, Auswertung und Repräsentation wissenschaftlicher Forschungsdaten und Ergebnisse. Wesentliches Merkmal der Entwicklungen innerhalb der Digital Humanities ist jedoch vor allem auch die Ermöglichung neuer Formen des Erenntnisgewinns, der ohne die digitalen Daten und Praktiken nicht möglich wären.

Referenzen

- Bartsch et al. 2009. “ObamaSpeeches.com: Building and Processing a Corpus of Political Speeches. A student project.“ Poster im Rahmen eines Workshops zum Thema: *Processing Pipelines* im Rahmen der Jahrestagung der GSCL (Gesellschaft für Sprachtechnologie und Computerlinguistik). Studentisches Projekt von Sabine Bartsch, Christoph Tragl, Claudio Weck, Stefania Degaetano, Tomasz Grubba, Nina Petrychka, David Sullivan. Universität Potsdam, 29. Sept. – 2. Okt. 2009.
- Schwarz et al. 2008. "Theme Annotator: A rule-based approach to automatic Theme-Rheme identification", mit Lara Schwarz, Richard Eckart, Elke Teich. *Proceedings of the 9th Conference on Natural Language Processing (KONVENS 2008)*. Berlin, New York: Mouton de Gruyter.

Zusätzliche Information zu den eingesetzten Ressourcen

- Textcorpora
- Abfragewerkzeuge
- Annotationswerkzeuge
- Processing Pipelines
- Anforderungen
- Links zu den eingesetzten Werkzeugen

Eingesetzte Sprachressourcen

TEXTCORPORA

- Monolinguale Corpora (BNC, ICAME corpora, ICE International Corpora of English etc.)
- Multilinguale Corpora
- Textarchive (OTA, Project Gutenberg)
- Elektronische Editionen und Wörterbücher

- Corpuscompilation
- Corpuskodierung (Unicode etc.)
- XML-Familie und Standards (XML, XSLT; TEI) für strukturierte Textressourcen und Metadaten

Eingesetzte Sprachressourcen

ABFRAGEWERKZEUGE

- Webinterfaces (Mark Davies‘s Corpus interfaces)
- Stand-alone Konkordanzprogramme
- IMS Open Corpus Workbench mit CorpusWeb (eigenes Interface für Corpusabfragen mit CQP)
- ANNIS2 (Corpusimport nicht trivial)

- Plain text Abfragen
- Abfragen über annotierte Corpora
- Baumbankabfragen (Tregex, TigerSearch)
- Abfragen über multilayer Annotationen

Eingesetzte Sprachressourcen

ANNOTATIONSWERKZEUGE

- Automatische Annotationswerkzeuge
  - POS Tagger (TreeTagger, Stanford POS)
  - Syntaktische Parser (Stanford Parser)
  - Diskursannotation (OpenNLP Tools, eigenes automatisches Kohäsionsannotationswerkzeug)

- Manuelle Annotationswerkzeuge
  - Systemic Coder / RST Tool
  - Multilayer Annotation (Exmaralda, UAM Corpus Tool, MMAX2)
  - TextGrid-Werkzeuge (Edition, Text-Bild)

Eingesetzte Sprachressourcen

PROCESSINGPIPELINES

- Integrierte Toolchains / kompatibele Werkzeuge
  - Stanford NLP Tools
  - OpenNLP Tools, LingPipe
  - GATE (ANNIE) / UIMA und Eclipse
  - VREs: TextGrid

- Kompatibele, kombinierbare Werkzeugsets
- Einheitliche Programmierung / Annotation / Ein- und Ausgabeformate

Eingesetzte Sprachressourcen

ANFORDERUNGEN

- Fachwissenschaftlich
  - Linguistische Fragestellungen
  - Linguistische Theorien
- Methoden
  - Corpuslinguistik
  - Daten in der Linguistik
  - Corpora, Werkzeuge, Herangehensweisen

Eingesetzte Sprachressourcen

ANFORDERUNGEN

- Ressourcen (Werkzeuge / Daten)
  - Plattformunabhängig
  - Frei verfügbar
  - Lokal installierbar, extern zugänglich
  - Kompatibele Formate
  - Werkzeuge zur Formattransformation
- Ressourcen (Institutionell)
  - Technische Ressourcen an den Universitäten und universitätsübergreifend
  - Schaffung von Lehrressourcen zur Verstetigung von Wissen und zum Wissenstransfer

Tools

- Stanford NLP: http://nlp.stanford.edu/
- OpenNLP Tools: http://incubator.apache.org/opennlp/
- LingPipe: http://alias-i.com/lingpipe/
- GATE: http://gate.ac.uk/
- Apache UIMA: http://uima.apache.org/
- TextGrid: http://www.textgrid.de/
- NLTK: http://www.nltk.org/
- TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Manual annotation tools

- Exmaralda: http://www.exmaralda.org/
- MMAX2: http://mmax2.sourceforge.net/
- RST Tool: http://www.wagsoft.com/RSTTool/
- UAM Corpus Tool: http://www.wagsoft.com/CorpusTool/

Query

- ANNIS 2: http://www.sfb632.uni-potsdam.de/d1/annis/
- Concordancer for Windows: http://www.linglit.tu-darmstadt.de/index.php?id=linguistics
- IMS Open Corpus Workbench mit CorpusWeb: http://cwb.sourceforge.net/
- WordSmith Tools: http://www.lexically.net/wordsmith/

Corpora and other resources

- British National Corpus: http://www.natcorp.ox.ac.uk/
- Brown corpus: http://icame.uib.no/brown/bcm.html
- LOB corpus: http://khnt.hit.uib.no/icame/manuals/lob/index.htm
- Mark Davies‘ Concordance View: http://corpus.byu.edu/
- NLTK (Natural Language Toolkit) corpora: http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml
- WordNet: http://wordnet.princeton.edu/