TECHNISCHE UNIVERSITÄT DARMSTADT

Sprachressourcen in der Lehre:
Erfahrungen, Einsatzszenarien, Nutzerwünsche

Prototypen und Processing Chains: Werkzeuge und Kompetenzen für die linguistische Sprachverarbeitung

Sabine Bartsch

Technische Universität Darmstadt

Institut für Sprach- und Literaturwissenschaft

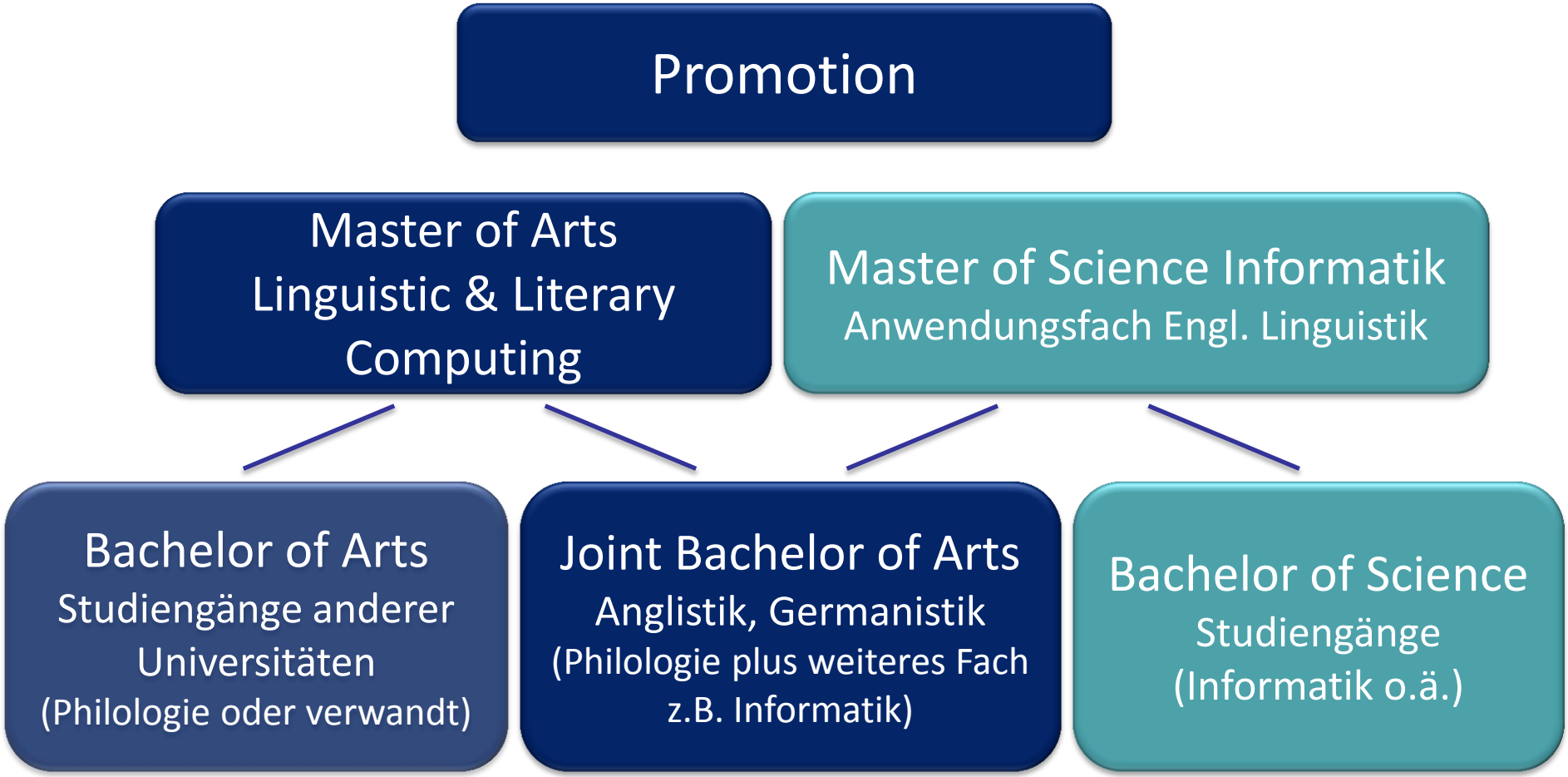
URI: http://www.linglit.tu-darmstadt.de

E-Mail: {lastname}@linglit.tu-darmstadt.de

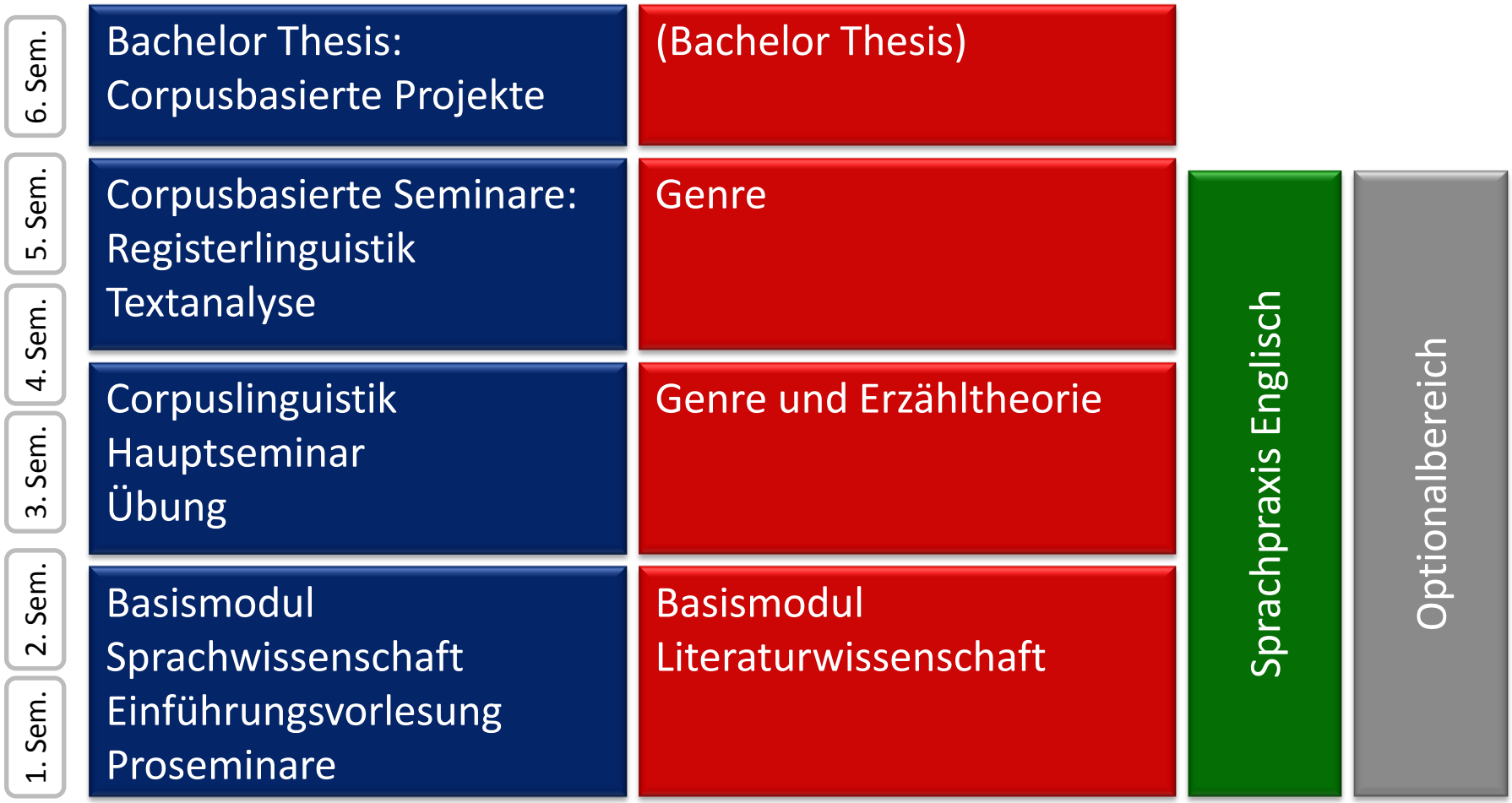
Rahmenbedingungen

- Studiengänge
- Zielkompetenzen und Erwartungshorizont
- Verwendete Sprachressourcen
- Lehre und F & L Transfer
- Erfahrungen und Lehren

Studiengänge & Zielgruppen



Bachelor of Arts Anglistik
Studienprogramm 180 CP



Zielkompetenzen: Bachelor of Arts

Corpusbasierte Seminare	• Corpusbasierte Seminare (Kookkurrenzphänomene, Registerlinguistik, Diskursphänomene) • Anwendung corpuslinguistischer Fertigkeiten • Corpusstatistik (Frequenzen, statistische Verfahren)
Corpuslinguistik: Lexis, Grammatik, Diskurs	• Corpusabfrage • Frequenzanalyse (Häufigkeitsverteilung, Kookkurrenz) • Automatische und manuelle Annotation • Tokenisierung, lx MWA, Satzerkennung • Corpora und Corpuskodierung
Basismodul Sprachwissenschaft	• Empirische Methoden • „Händische“ Analysen • Sprachsystem und Sprachtheorie • Grundlagen der Sprachwissenschaft

Verteilung auf Lernergruppen

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

Studentische Projekte

- Kollokationen in literarischen Texten am Bsp. 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 „flacher“ vs. „runder“ Charaktere

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
  - Problematik: Umgang mit anderen Alphabetsystemen und Kodierungen

Studentische Projekte

- Multimodale Analyse von Werbeartefakten aus der Kosmetikbranche (Printwerbung und YouTube)
  - Corpusaufbau: Printwerbung und YouTube Channels der Firmen
  - Transkription und Annotation mit automatischen Werkzeugen und Exmaralda
  - Auswertung von Unterschieden zwischen Produktlinien und Firmen (Interpersonale Relationen, Lexiko-Grammatik, Pseudowissenschaftlicher Jargon „Nanosomen-Komplex“)

Herausforderungen

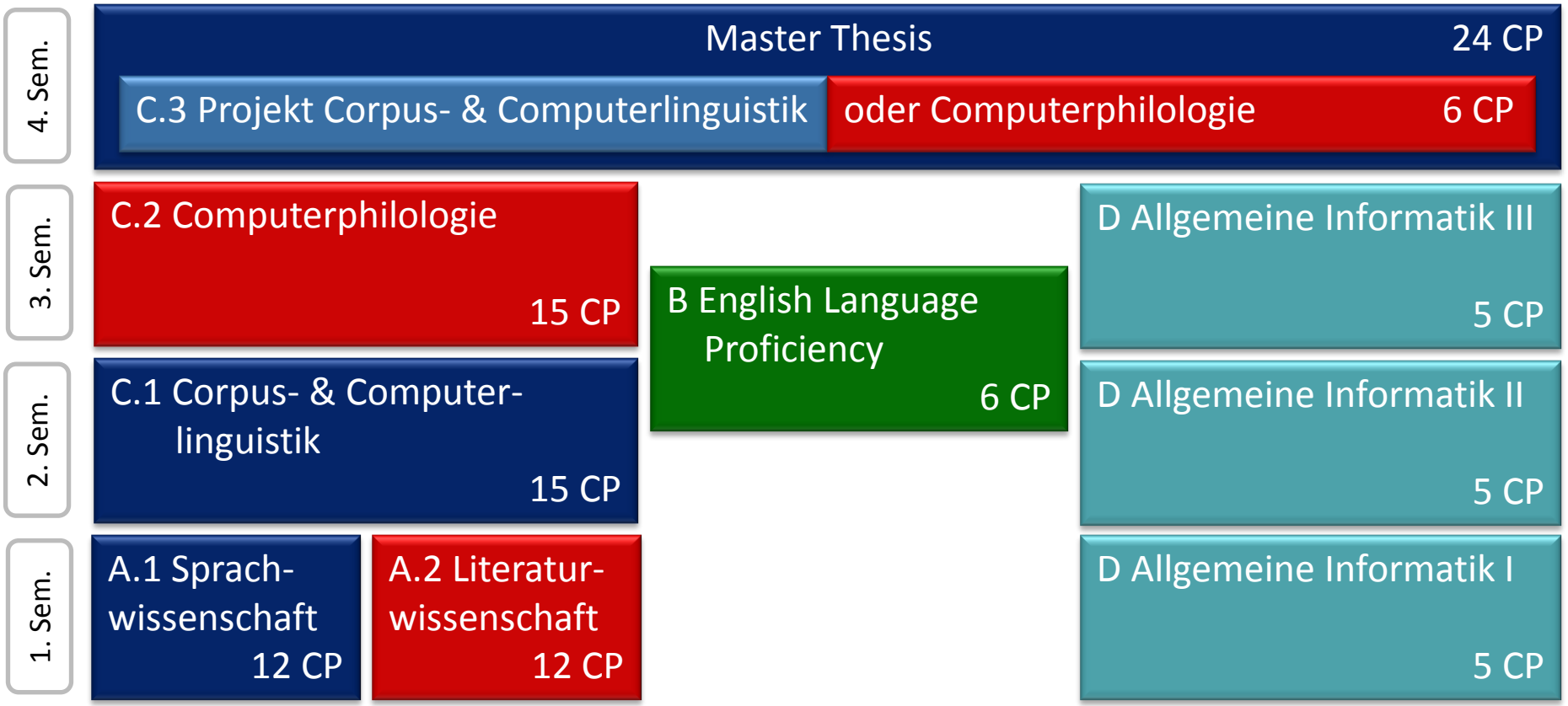
- Linguistische Fragestellungen
- Operationalisierung
- Auswahl, Aufbau und Benutzung geeigneter Ressourcen (Corpora, Annotation, Query)

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

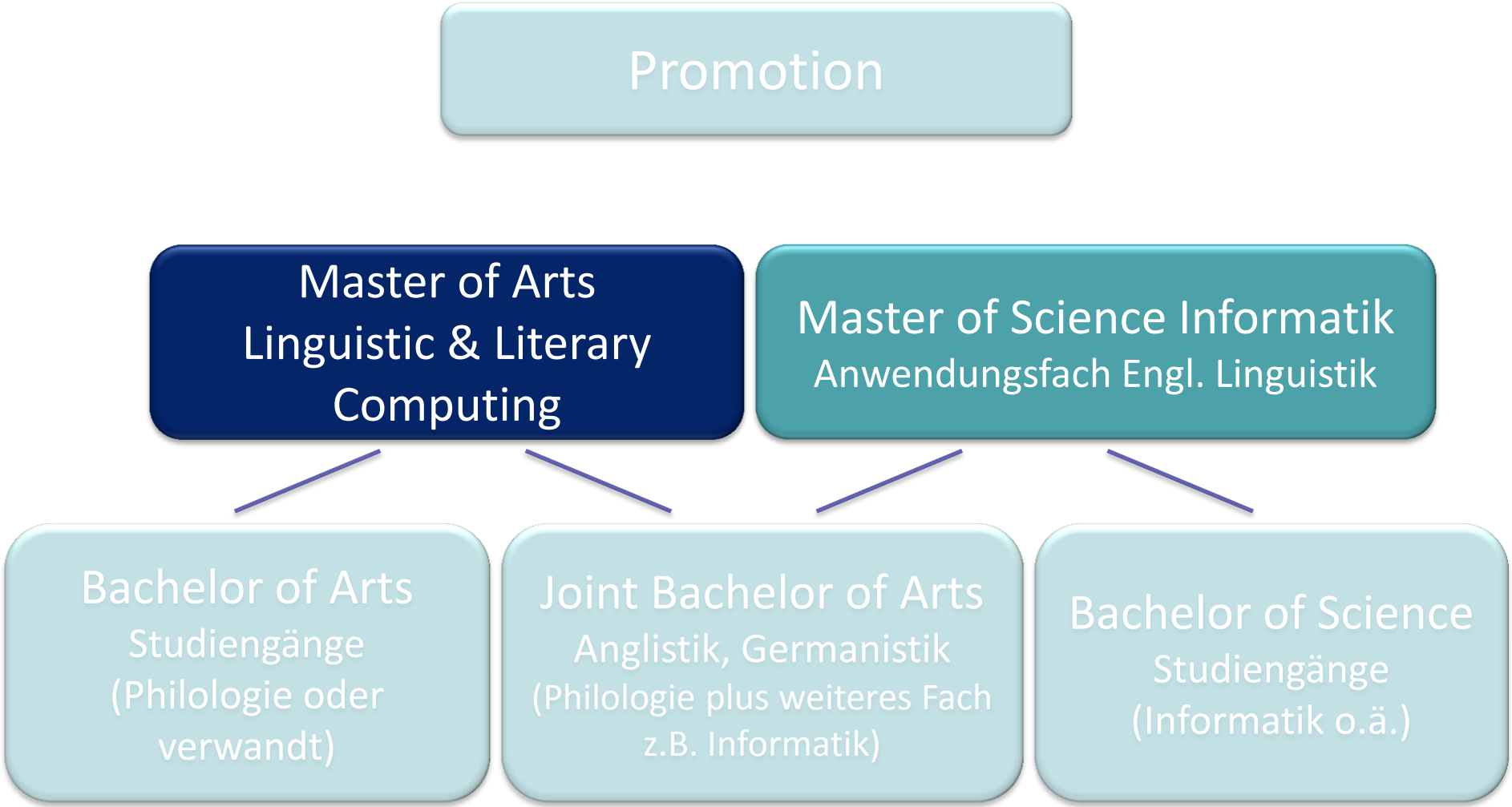
Ziele im philologischen Bachelor

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

Master of Arts
Linguistic & Literary Computing 120 CP



Studiengänge & Zielgruppen



Zielkompetenzen: Master of Arts Linguistic & Literary Computing

Informatik	• Einführung in die Allgemeine Informatik • Java Programmierung • NLTK (NLP mit Python) (in der Erprobung)
Corpus- und Computer-linguistische / Computerphilologische Seminare	• Registerlinguistik / Diskurslingustik • Corpora und probabilistische Verfahren • Computerphilologie (Edition, Lexikographie) • XML-Familie (XML, XSLT; TEI)
Computeranwendungen in der Linguistik	• Anwendungen (Annotation, MT, IR/IE, Diskursorganis.) • Ressourcenaufbau • Techniken und Werkzeuge • Fortgeschrittene Annotationsaufgaben
Sprachwissenschaft	• 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
  - Auswertung und Visualisierung der Merkmale (mit xslt)

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.lingilt.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:

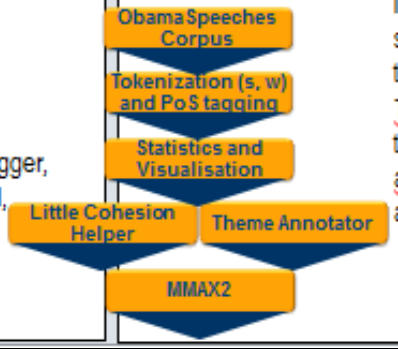
- Stand alone tools (Decision Tree Tagger, Theme Annotator, UAM Corpus Tool, MMAX2, etc.)
- GATE
- Natural Language Toolkit (NLTK)

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.

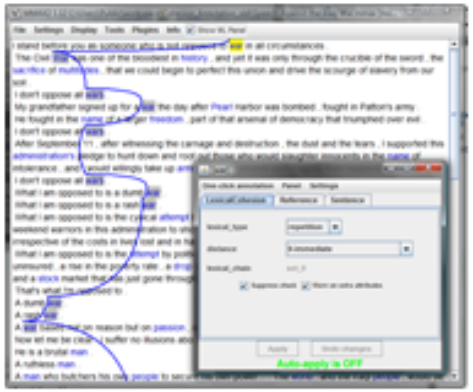


Fig. 2: MMAX project by LCH

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/

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

Zielkompetenzen im Master LLC

- Entwicklung eigener Workflows
- Entwicklung von Spezifikationen und Prototypen (LLC-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)

Eingesetzte Sprachressourcen

- Textcorpora
- Abfragewerkzeuge
- Annotationswerkzeuge
- Processing Pipelines
- Anforderungen

Verwendete 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

Verwendete 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)

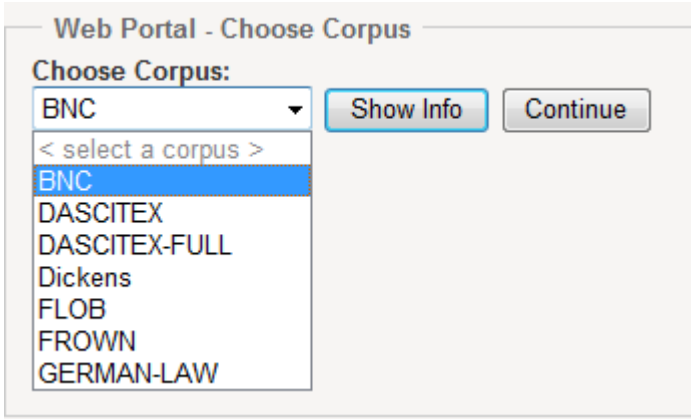
- Plain text Abfragen
- Abfragen über annotierte Corpora
- Baumbankabfragen (Tregex, TigerSearch)
- Abfragen über multilayer Annotationen

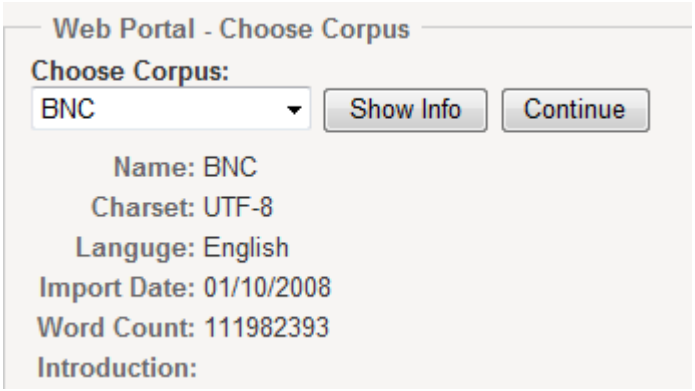
WEB CWB PORTAL

Eigenes IMS Open CWB Interface

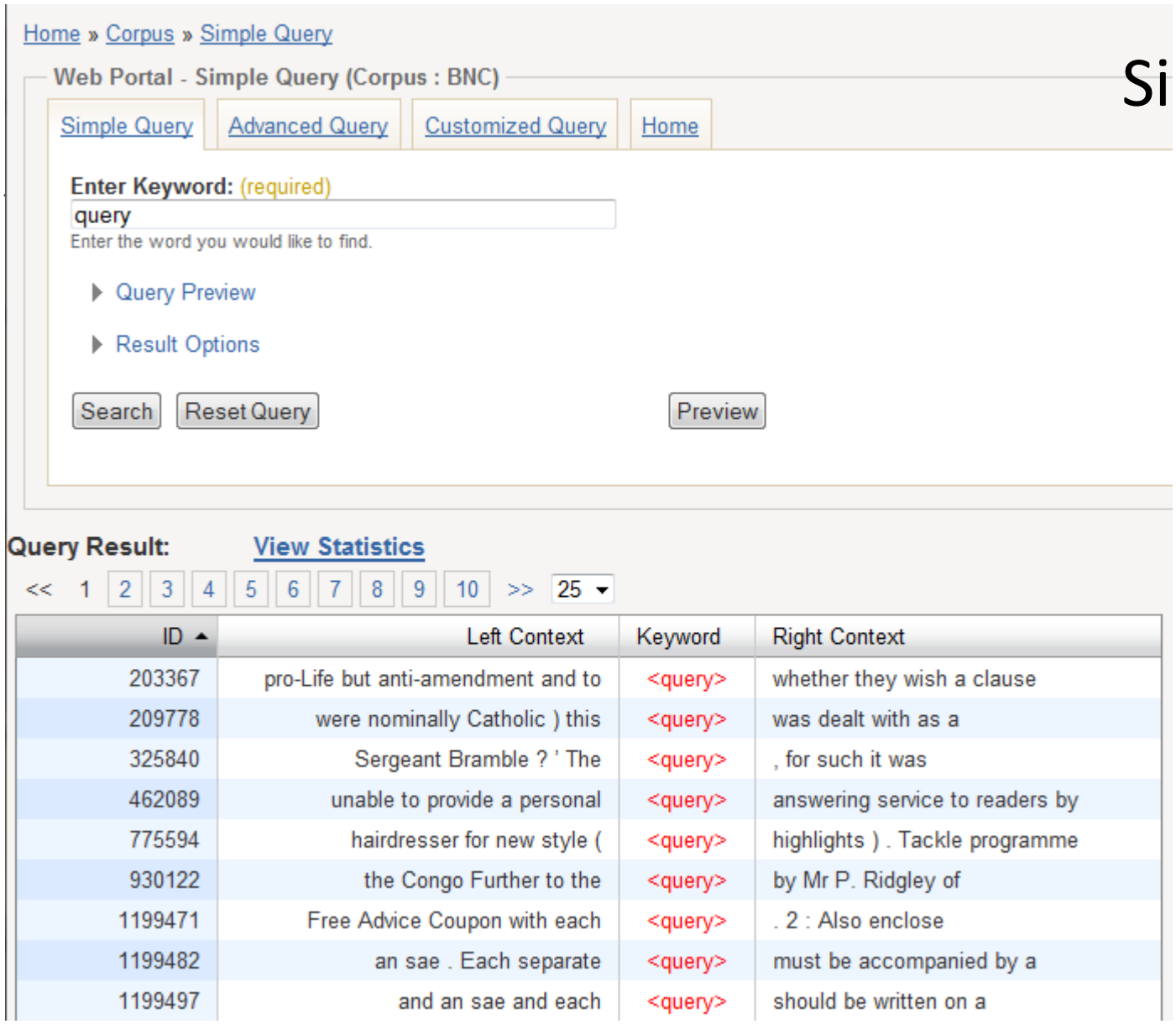
- Bereitstellung Copyright-geschützter Corpora
- Didaktische Unterstützung beim Erlernen einer Abfragesprache

Corpusauswahl



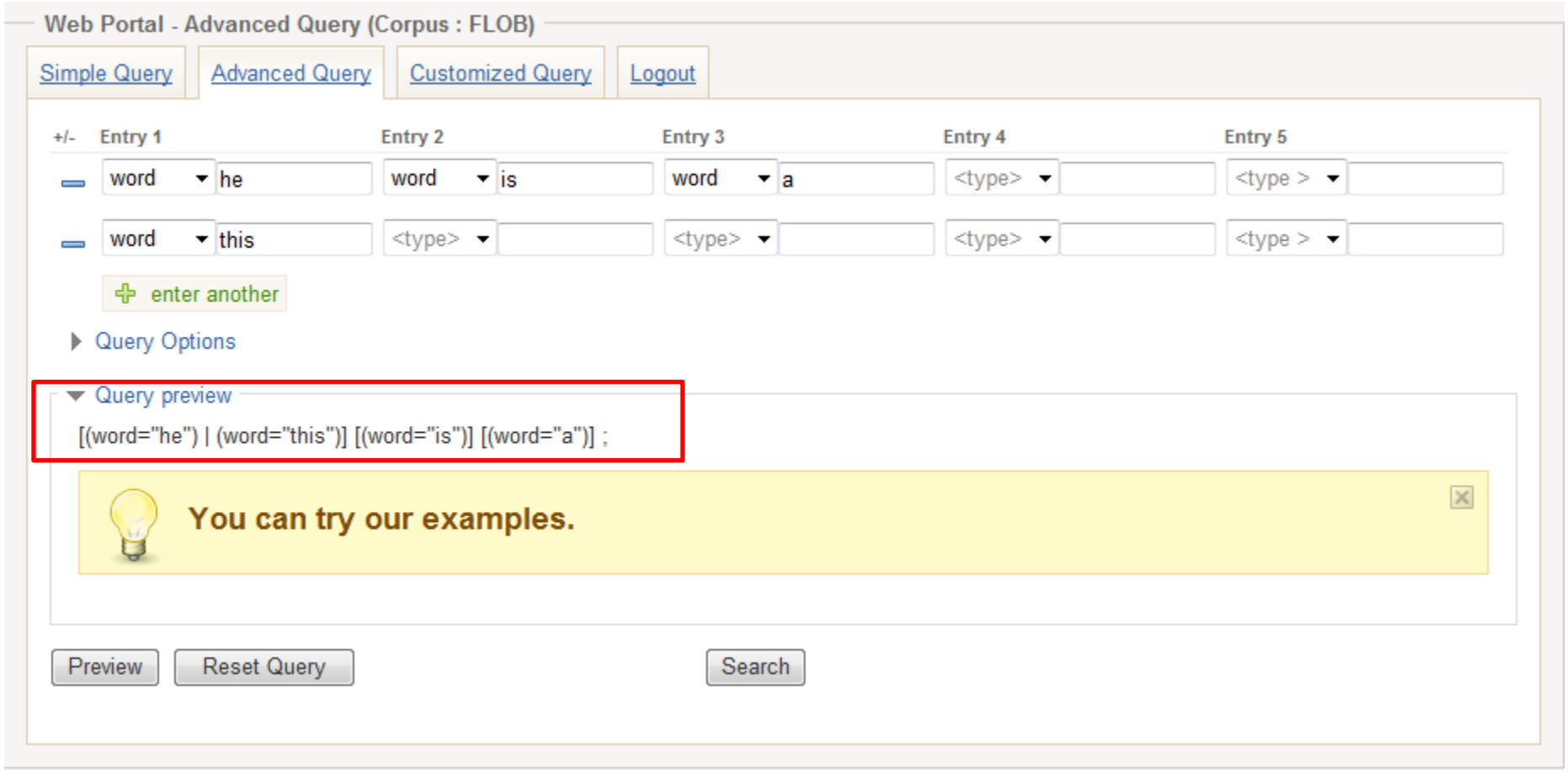


Simple query



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



Advanced query

Home » Corpus » Advanced Query

Web Portal - Advanced Query (Corpus : BNC)

Simple Query | Advanced Query | Customized Query | Home

+/-	Entry 1	Entry 2	Entry 3	Entry 4	Entry 5
	lemma have	lemma be	c5 VVG	<type>	<type >

enter another

Query Options

Query preview

```
[(lemma="have")] [(lemma="be")] [(c5="VVG")] ;
```

Result Options

Result Context:

5 Words Change

Sort Result By...:

First word on the left Sort

Save Result As XML/HTML/TEXT/CSV

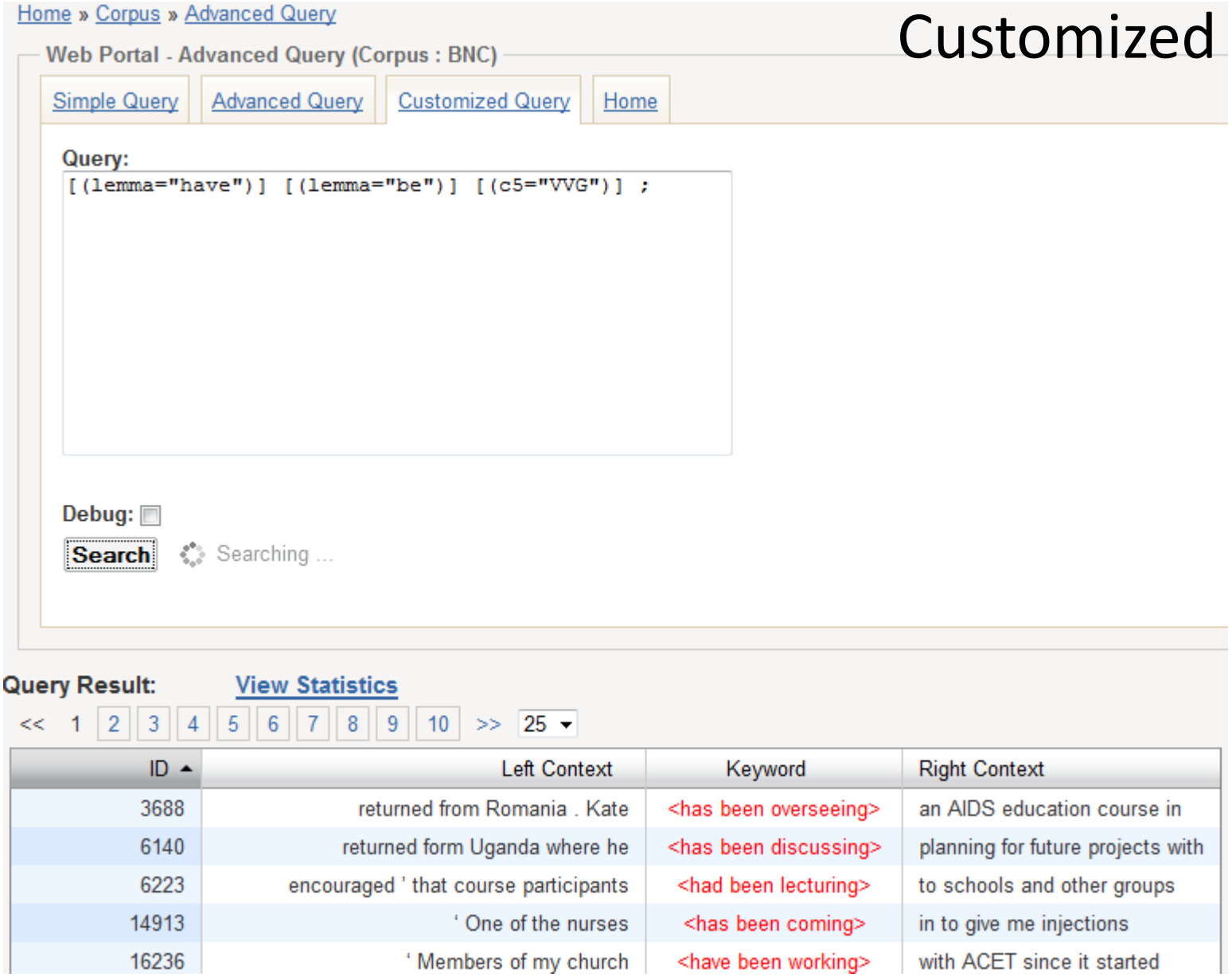
Search Reset Query Preview

Query Result: View Statistics

<< 1 2 3 4 5 6 7 8 9 10 >> 25

ID	Left Context	Keyword	Right Context
4056082	passing . ' ' You	<have been waiting>	! ' Erika stamped her
682481	dung steamed where the cattle	<had been standing>	, and Cameron , in
708990	But surely , if he	<had been spying>	, he would have had
850088	drinks up . ' I	<'ve been wondering>	, now it seems ,

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

Verwendete 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)

Verwendete Sprachressourcen

PROCESSINGPIPELINES

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

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

Verwendete Sprachressourcen

ANFORDERUNGEN

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

Verwendete 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
  - Lehrressourcen durch erhöhten Aufwand (andere Lehrformen, zeitlicher Aufwand)

Portal mit

- Materialien,
- Corpora,
- Query-Interface

über E-Learning Plattform und auf eigenen Servern

Sabine Bartsch (01-2011)

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)
- Volltextzugang zu allen Ressourcen
- Erhöhter Aufwand pro Lehrveranstaltung
- Gemischte Gruppen aus Linguisten / Philologen und Informatikern führen oft weiter

Erwartungen und Wünsche

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

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.

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/
- WordNet: http://wordnet.princeton.edu/