Einführung in die Computerlinguistik und Sprachtechnologie

Vorlesung im WiSe 2018/19
(B-GSW-12)

Prof. Dr. Udo Hahn

Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

http://www.julielab.de

Allgemeine Hinweise

- Vorlesung: Mi, 10-12h (Humboldt 8, SR 1)
- Übung zV: Fr, 8-10h (Fürstengrab. 1, SR 275)
  - beginnt am 19.10.
- Vorlesungsmaterialien im Netz
  - http://www.julielab.de/ ⇨ „Students“
- B-GSW-12 besteht aus VL+ÜB und Seminar!
- Sprechstunde: Mi, 12-13h (nA) (FG 30, R 004)
- Email: udo.hahn@uni-jena.de
- URL: http://www.julielab.de
- Fachliteratur ist überwiegend in Englisch

Bitte …

- … Handys/Smartphones ausschalten
- … 90 Minuten ohne Mail- und Tweet-Check sind möglich

  „Digital detox“
- … kein Picknick

Institut für Germanistische Sprachwissenschaft der FSU Jena

- Lehrstuhl für Theoretische Linguistik – Grammatiktheorie
  - Prof. Dr. Peter Gallmann bzw. n.n.
- Lehrstuhl für Angewandte Linguistik – Computerlinguistik
  - Prof. Dr. Udo Hahn
- Professur für Pragmatik
  - Prof. Dr. Pia Bergmann
- Professur für Phonetik & Sprechwissenschaft
  - Prof. Dr. Adrian Simpson
- Professur für Geschichte der deutschen Sprache
  - Prof. Dr. Eckhard Meineke

Computerlinguistik in Jena (1/2)

- Institutionell: Teil der Germanistischen Sprachwissenschaft
  - aber einzelsprachübergreifende Methodik
  - besondere Anwendungsdomänen:
    - Naturwissenschaften: Biologie + Medizin
    - Sozial- und Wirtschaftswissenschaft
    - Digital Humanities
- Integration in die Informatik:

  Neben- bzw. Anwendungsfach für
    - B.Sc.: Informatik, Angewandte Informatik
    - M.Sc.: Informatik, Computational Science

Computerlinguistik in Jena (2/2)

- Aktive Forschergruppe
  - Lehrstuhl für Computerlinguistik = Jena University Language & Information Engineering (JULIE) Lab
    - Hohe internationale Visibilität (Publikationsdichte)
  - Deutsche Forschungsgemeinschaft (DFG)
    - Aktuell: (1/5) SFB 1076 AquaDiva – Biodiversität in der Critical Zone
    - Aktuell: Graduiertenkolleg Modell ‚Romantik' [Digital Humanities]
  - Bundesministerium für Bildung & Forschung (BMBF)
    - Aktuell: Nationale Förderinitiative „Systemmedizin" (J – L – AC)
    - Frühere Projekte: Forschungs-Cluster JenAge – Nationaler Forschungskern, StemNet
  - Förderinitiativen der Europäischen Union
    - Frühere Projekte: MANTRA (SA), CALBC (SA), BOOTStrep (STREP), ..
- Ausgründung von Start-up-Firmen
  - *Averbis, TexKnowlogy*
- Jobs, Jobs, Jobs … etwa als studentische Hilfskraft
- Themen, Themen, Themen … BA- oder MA-Arbeit, Dissertation

Weitere Veranstaltungen

- Seminar zu B-GSW-12
  - SoSe 2019
- Vorlesung/Übung ASQ-DH
  - Einführung in Digital Humanities: Grundlagen der Informatisierung der Geisteswissenschaften
  - Di, 17-19, Humboldt 8, SR 3

Computer (und Menschen!) tun sich schwer mit Sprache(n) …

Die *pykka* Sprache

- Güney pykka-i tassas pel Criftek ut pykka-e coggy pons Criftek
  - coggy (1)
  - Criftek (2)
  - Günny (1)
  - pel (1)
  - pons (1)
  - pykka-i (1)
  - pykka-e (1)
  - tassas (1)
  - ut (1)

Lexikografische Ordnung

Häufigkeits-zählung

Die *pykka* Sprache

- Günny pykka-i tassas pel Criftek ut pykka-e coggy pons Criftek
  - Perspektive des Computers/Menschen auf diese Äußerung:
    - uninterpretierbare Buchstaben-/Lautsequenz
  - Fehlt: Spezifikation von Wortbedeutung (Lexikon)
  - Fehlt: Regeln für Wortverknüpfung (Syntax)
  - Fehlt: Regeln für die Verbindung Syntax/Semantik
- Günny pykka-i tassas aus Criftek und pykka-e coggy nach Criftek
  - Pel → aus, ut → und, pons → nach
  - Lediglich ein Syntaxskelett

Die *pykka* Sprache

- Günny pykka-i tassas aus Criftek und pykka-e coggy nach Criftek
- Deutsche Wortäquivalente:
  - { Deutschland, Costa-Rica }
  - { exportieren, importieren }
  - { Optoelektronik, Banane }
- Deutschland importiert Bananen aus Costa-Rica und exportiert Optoelektronik nach Costa-Rica

Von *pykka* ins Deutsche I

- Günny pykka-i tassas aus Criftek und pykka-e coggy nach Criftek
- Deutsche Wortäquivalente:
  - [Deutschland = Günny, Costa-Rica = Criftek]
  - [importieren = pykka-i, exportieren = pykka-e]
  - [Banane(n) = tassa(s), Optoelektronik = coggy]
- Standard-Interpretation:

  Deutschland importiert Bananen aus Costa-Rica und exportiert Optoelektronik nach Costa-Rica

Von *pykka* ins Deutsche II

- Günny pykka-i tassas aus Criftek und pykka-e coggy nach Criftek
- Deutsche Wortäquivalente:
  - [Costa-Rica = Günny, Deutschland = Criftek]
  - [importieren = pykka-i, exportieren = pykka-e]
  - [Banane = tassas, Optoelektronik = coggy]
- Non-Standard-Interpretation:

  Costa-Rica importiert Bananen aus Deutschland und exportiert Optoelektronik nach Deutschland

Konstituenten der Analyse/ Produktion natürlicher Sprache

- Inventar von Wörtern (Lexikon) und ihrer Bedeutungen (lexikalische Semantik)
- Verknüpfungsregeln für Wörter (Syntax)
- Kompositionelle Ableitung der Bedeutung eines Satzes (Satz-Semantik) aus den lexikalischen Bedeutungen der Wörter und der Syntaxstruktur (semantische Interpretation)
- Evaluation der semantischen Interpretation auf der Basis von Hintergrundwissen (Enzyklopädie, Alltagswissen usw.)

Computerlinguistik I

- Linguistik: Gegenstandsbereich sind (überwiegend) natürliche Sprachen
  - Deutsch, Englisch, Französisch, ...
- Beispiele für formale Sprachen
  - $L = \{a^n b^n, n \in \mathbb{N}\}$

    $= \{ab, aabb, aaabbb, aaaabbbb, \ldots\}$
  - jede Programmiersprache, Auszeichnungssprache
    - JAVA, C++, ..., XML, HTML, ...
  - jede Logik
    - Aussagenlogik, Prädikatenlogik, Typenlogik, ...
  - Differentialgleichungen, Integrale, Vektoren, ...

Computerlinguistik II

- Beschreibungen und Formalisierungen entsprechen den Anforderungen, die sich aus der Verarbeitung durch Computer ergeben
  - keine natürlichsprachige Beschreibung (à la Duden oder Grammatik für Fremdsprachenerwerb), sondern formalisiert und damit explizit
  - explizite Spezifikation von Verfahrensbeschreibungen (Algorithmen), die von einer (abstrakten) Maschine ausgeführt werden können
  - Beachtung formaler (komplexitätstheoretischer) Eigenschaften der Beschreibung: Berechenbarkeit, Entscheidbarkeit, „Rechen-Kosten“ (Zeit, Speicher)

Computerlinguistik III

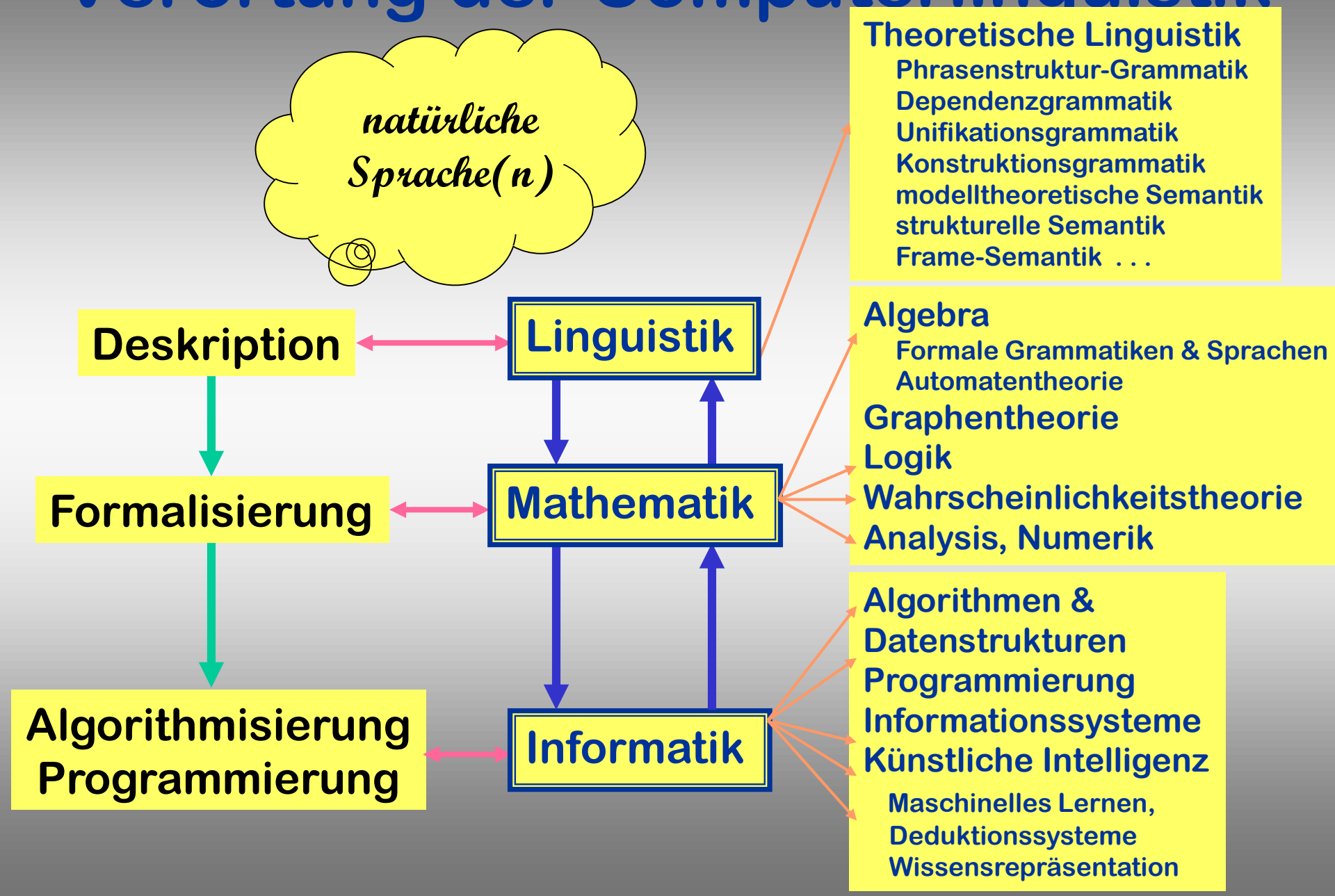
- Fundierung computerlinguistischer Beschreibungen durch Bezug auf theoretische und methodische Prinzipien der Linguistik und Informatik
  - Linguistische Grammatikmodelle vs. formale Grammatikmodelle der Informatik
  - Automatenmodelle der Informatik als Grundlage des Parsings natürlicher Sprache
  - Lexikonmodelle und Suchverfahren in Lexika
  - Semantische Repräsentationsformalismen vs. Wissensrepräsentationssprachen (Beschreibungslogik)

- Notabene: die Relevanz der Informatik nimmt aktuell zu, die der Linguistik ab !

Computerlinguistik IV

- Realisierung dieser Beschreibungen durch ihre Implementation in einem natürlichsprachlichen (Teil-)System entsprechend informatischer Standards
  - Computerlinguistik ist keine naiv „programmierte“ Linguistik
    - Programmiertechnologien (z.B. objekt-orientiert)
    - Daten(bank)technik (Speicher- und Zugriffsmethoden)
  - Software Engineering
    - Portierbarkeit (Domänenwechsel)
    - Wiederverwendbarkeit (Middleware: UIMA usw.)
    - Robustheit (NL ist ein sehr komplexes, nach wie vor nur partiell beschriebenes System)

Verortung der Computerlinguistik



Computerlinguistik-Standorte

www.ims.uni-stuttgart.de/info/SitesEurope.html#Germany

Computerlinguistik-Standorte

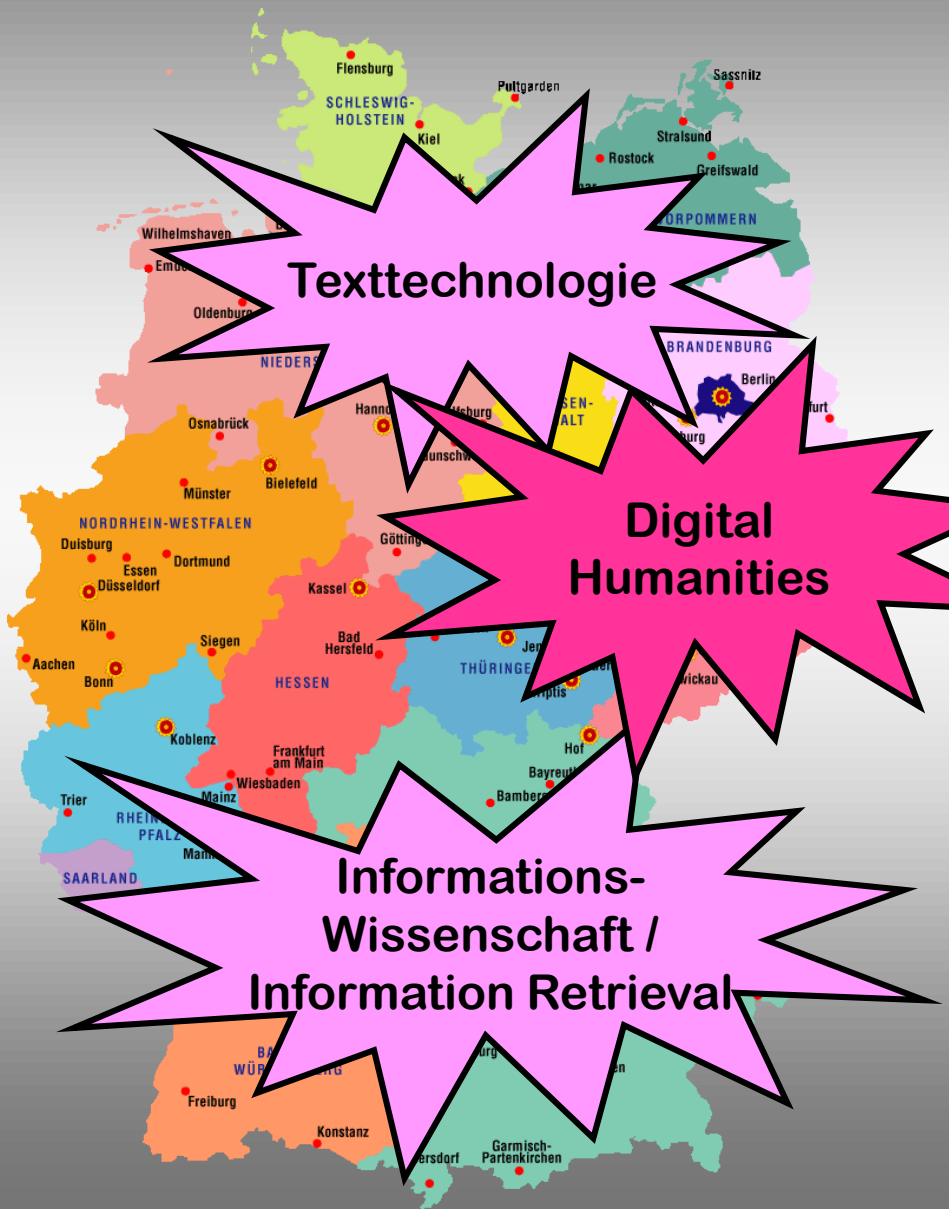
www.ims.uni-stuttgart.de/info/SitesEurope.html#Germany

24 [48]

- U Saarbrücken (6)
- U Stuttgart (3)
- U Heidelberg (5)
- RWTH Aachen
- U München (2)
- TU Darmstadt (4)
- U Jena
- U Tübingen (3)
- U Bielefeld (4)
- U Potsdam (2)
- U Bremen
- U Bochum (2)
- U Erlangen-Nbg.
- U Osnabrück (2)
- U Hamburg (3)
- KIT Karlsruhe
- U Duisburg-Essen
- U Leipzig
- U Magdeburg
- U Düsseldorf
- U Gießen
- U Hildesheim
- U Koblenz

Computerlinguistik-Standorte

www.ims.uni-stuttgart.de/info/SitesEurope.html#Germany



U Saarbrücken (6)
U Stuttgart (3)
U Heidelberg (5)
RWTH Aachen
U München (2)
TU Darmstadt (4)
U Jena
U Tübingen (3)
U Bielefeld (4)
U Potsdam (2)
U Bremen
U Bochum (2)
U Erlangen-Nbg.
U Osnabrück (2)
U Hamburg (3)
KIT Karlsruhe
U Duisburg-Essen
U Leipzig
U Magdeburg
U Düsseldorf
U Gießen
U Hildesheim
U Koblenz

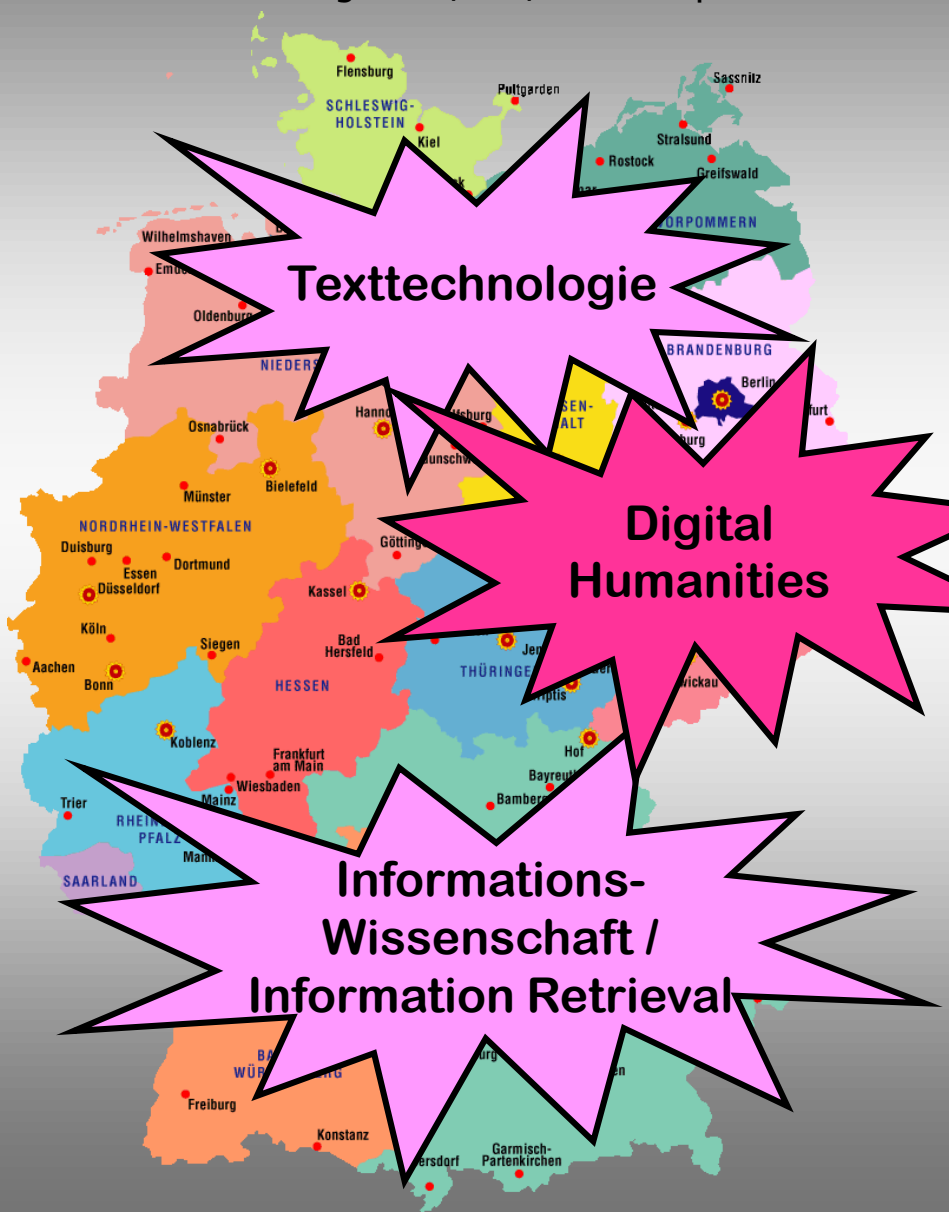
TU Darmstadt (2)
U Frankfurt/M. (2)
U Leipzig
U Bamberg
U Köln
U Passau
U Jena
HU Berlin
U Stuttgart
U Konstanz
U Dortmund
U Kassel
U Würzburg
U Göttingen
U Münster

U Regensburg
U Hildesheim
U Düsseldorf
U Dortmund
BU Weimar
U Bamberg
U Kaiserslautern
TU Dresden

http://www.dig-hum.de/

Computerlinguistik-Standorte

www.ims.uni-stuttgart.de/info/SitesEurope.html#Germany



Texttechnologie

Digital Humanities

Informations-Wissenschaft / Information Retrieval

U Saarbrücken (6)
U Stuttgart (3)
U Heidelberg (5)
RWTH Aachen
U München (2)
TU Darmstadt (4)
U Jena
U Tübingen (3)
U Bielefeld (4)
U Potsdam (2)
U Bremen
U Bochum (2)
U Erlangen-Nbg.
U Osnabrück (2)
U Hamburg (3)
KIT Karlsruhe
U Duisburg-Essen
U Leipzig
U Magdeburg
U Düsseldorf
U Gießen
U Hildesheim
U Koblenz

+ 23 [25]

TU Darmstadt (2)
U Frankfurt/M. (2)
U Leipzig
U Bamberg
U Köln
U Passau
U Jena
HU Berlin
U Stuttgart
U Konstanz
U Dortmund
U Kassel
U Würzburg
U Göttingen
U Münster

U Regensburg
U Hildesheim
U Düsseldorf
U Dortmund
BU Weimar
U Bamberg
U Kaiserslautern
TU Dresden

http://www.dig-hum.de/

Warum Compute ?

(auch mit einer Antwort