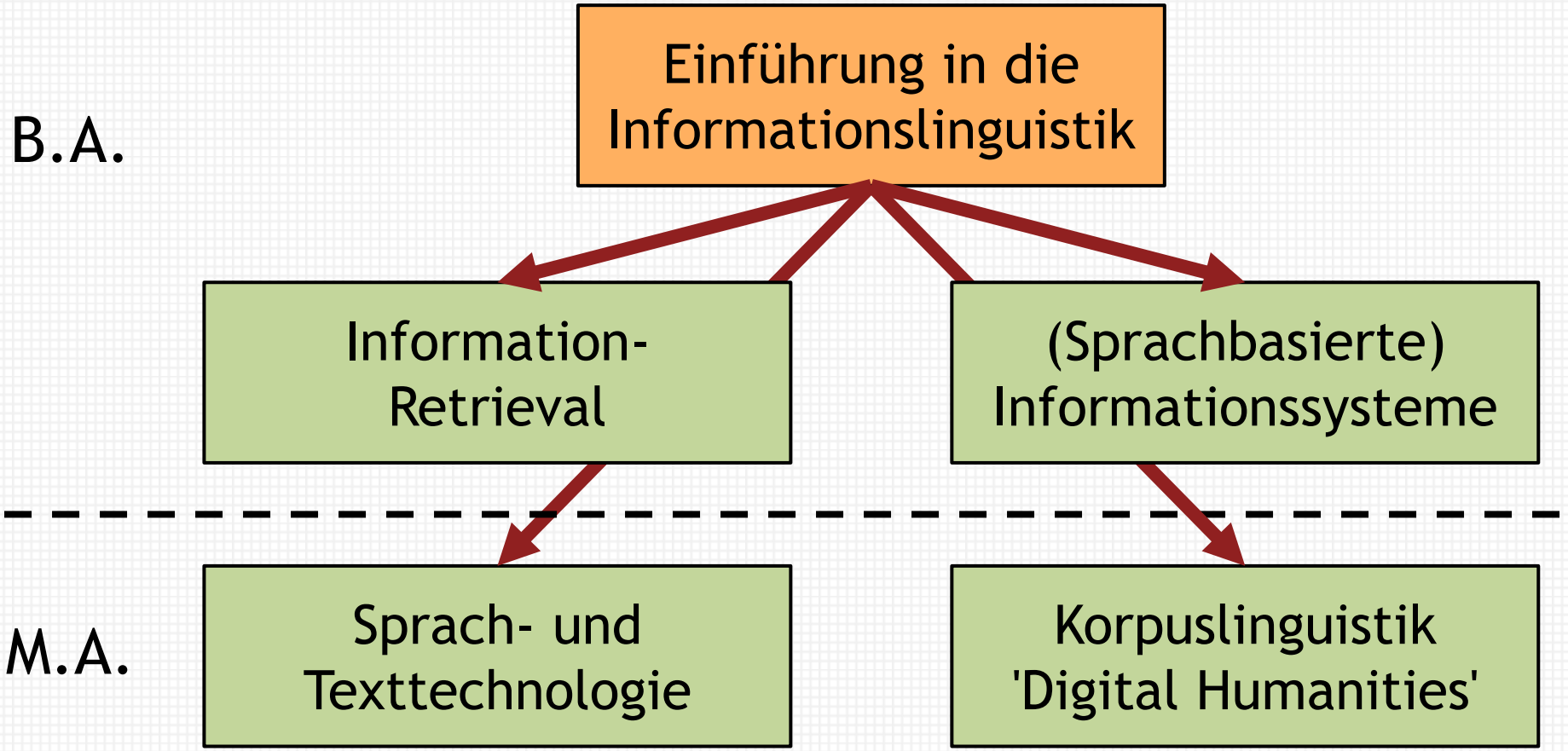
Einführung (Skript 2013)

Informationswissenschaft
Universität Regensburg
Jürgen Reischer

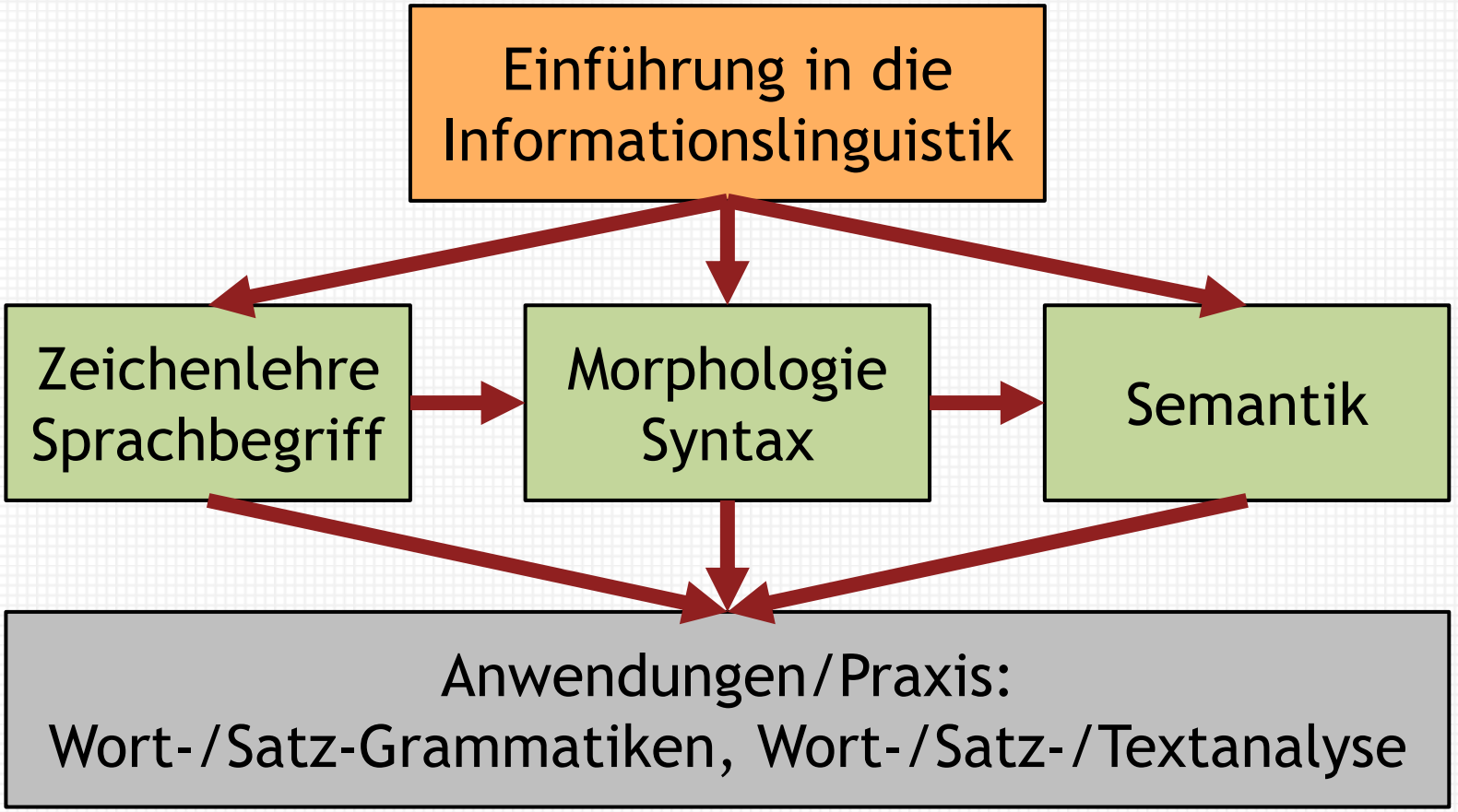
Informationslinguistik in Regensburg

- Übersicht aus informationswissenschaftlicher Sicht:



Informationslinguistik in Regensburg

- Übersicht aus sprachwissenschaftlicher Sicht:



Was ist Informationslinguistik?

- Informationslinguistik: *Sprachwissenschaft im Dienste der Informationswissenschaft* unter Nutzung von Verfahren und Erkenntnissen benachbarter wissenschaftlicher Disziplinen:
  - Computerlinguistik und Künstliche Intelligenz:
    - Computerlinguistik (Informatik und Sprachwissenschaft): Verfahren zur maschinellen Verarbeitung von Sprache;
    - Künstliche Intelligenz (Informatik und Kognitionswissenschaft): Einbettung der Sprache in die allgemeine Kognition (Denken/Problemlösen/Handeln) und Kommunikation (Interaktion mit Umwelt, Diskurs) des Menschen.

Was ist Informationslinguistik?

- Korpus- und Textlinguistik (spezielle Bereiche bzw. Herangehensweisen der Sprachwissenschaft):
  - *Korpuslinguistik*: sprachwissenschaftlicher Teilbereich, der sich mit der Sammlung und Analyse von Texten als repräsentativem Ausschnitt einer Sprache befasst;
  - *Textlinguistik*: sprachwissenschaftlicher Teilbereich, der sich mit der Analyse einzelner Texte und deren Eigenschaften befasst (formal-inhaltlicher Aufbau/Einordnung).
- Sprach- und Texttechnologie:
  - ziel- und praxisorientierte Realisierung sprachbasierter Anwendungen (z. B. Dialogsysteme);
  - ingenieursmäßige statt wissenschaftliche Herangehensweise (Funktionieren/Konstruieren statt Modellieren).

Was ist Informationslinguistik?

- Informationslinguistik als integraler Bestandteil und Teildisziplin der Informationswissenschaft:
  - (automatische) *Verarbeitung natürlicher Sprache* in Form von geschriebenen oder gesprochenen Texten in und für Informationssysteme (z. B. Information-Retrieval-Systeme wie Suchmaschinen oder OPACs);
  - informationslinguistische Verfahren und Modelle spielen eine entscheidende Rolle bei der Er- und Verarbeitung sprachlich verfasster Information (s. ausführlicher unten):
    - Textakquisition und Textproduktion;
    - Textverwaltung und Textnutzung.
  - ➡ *Konzentration auf Interaktionsmedium/-modalität der Sprache (im Gegensatz zur Medieninformatik).*

Was macht Informationslinguistik?

- *Textakquisition*:
  - *Spracherkennung*: automatische Überführung gesprochener in geschriebene Sprache;
  - *Texterkennung*: automatische Überführung gedruckter (analoger, nicht-digitaler) Texte/Werke in digitale Dokumente durch OCR ('Optical Character Recognition'):
    - z. B. Digitalisate in Bibliotheken (digitalisierte Bücher, Magazine etc.);
    - z. B. Google Books.

Was macht Informationslinguistik?

- *Textproduktion*:
  - *Fehlererkennung*: automatische Tipp- und Rechtschreib-fehlerkorrektur in Textverarbeitungen, Mailprogrammen usw.;
  - *Sprachverbesserung*: automatische lexikalische und grammatikalische Analyse zur Erkennung und Behebung verschiedener Arten sprachlicher Abweichungen:
    - *Lexik*: Terminologie-Prüfung zur Ersetzung ungebräuchlicher/ obsoleter oder umständlicher/komplizierter Ausdrücke durch Synonyme anhand eines Thesaurus;
    - *Grammatik*: stilistische und syntaktische Prüfung anormaler oder fehlerhafter Konstruktionen.

Was macht Informationslinguistik?

- *Textverwaltung*:
  - *Wissensrepräsentation*: Erschließung/Aufbereitung und Kodierung/Darstellung von Text(inhalt)en durch Metadaten (Daten, die Form und Inhalt des Textes selbst beschreiben):
    - *Indexierung*: automatische Ermittlung von Deskriptoren als inhaltsbeschreibende Ausdrücke (Terme);
    - *Zusammenfassung*: automatische Erzeugung von Abstracts/Extracts (Ermittlung interessierender und relevanter/wesentlicher Inhalte zur thematischen Textkomprimierung);
    - *Textklassifikation*: automatische Erkennung von Textsorte/-genre bzw. thematische Einordnung innerhalb einer Fachgebietsklassifikation.
  - *Ressourcengenerierung*: automatischer Aufbau von Thesauri/Wortnetzen oder Ontologien aus Textkorpora.

Was macht Informationslinguistik?

- *Textnutzung/Textrecherche*:
  - *Informationswiederfindung*: Auffinden von Texten mit interessierenden Inhalten über beschreibende Terme (Suchbegriffe) und Abfragesprachen (z. B. mittels 'und' bzw. 'oder' verknüpfte Suchterme):
    - Information-Retrieval-Systeme zur Suche von Dokumenten in einem Dokumentenbestand, Within-Document-Retrieval-Systeme zur Suche innerhalb eines Dokuments;
    - Frage-Antwort-Systeme für Daten-/Faktenbanken (z. B. Wolfram Alpha, s. unten).
  - *Sprachsteuerung*: Dialogsysteme zur sprachlichen Interaktion mit Informationssystemen (z. B. Navigations-, Auskunfts-, Empfehlungs-, Bestell- und Expertensysteme, Chatbots).

Phasen der Textverarbeitung

- Die zugrunde liegenden Texte müssen aus ihrer Rohform durch einen Annotationsprozess in eine für die Maschine verarbeitbare Form gebracht werden:
  - *Annotation* als Prozess und Ergebnis der *Hinzufügung linguistischer Information* (z. B. Wortart, Textsorte, Thema usw.): Erzeugung eines linguistischen 'Mehrwerts' für den Text;
  - *Repräsentation* (Kodierung/Darstellung) des Textes in einem strukturierten Format (z. B. XML/Unicode).

Phasen der Textverarbeitung

Phase	Aktion	Beschreibung
1	*Kodierung*	Vereinheitlichung verschiedener Zeichendarstellungsformen in das universelle Unicode-Format (z. B. Umkodierung ASCII-kodierter Texte)
2	*Tokenisierung*	Zerlegung eines Textes in einzelne sinntragende Einheiten (so genannte 'Tokens' im Sinne bedeutungstragender Zeichen)
3	*Normalisierung*	Deflexion von Wörtern (Reduktion auf Grund- und Stammformen), Vereinheitlichung von Zahlendarstellungen usw.
4	*Worterkennung*	Ermittlung zusammengehöriger Sinneinheiten nach Tokenisierung und Normalisierung (z. B. Telefonnummern), Erkennung und Zerlegung/Interpretation komplexer Wortbildungen (z. B. Komposita)

Phasen der Textverarbeitung

Phase	Aktion	Beschreibung
5	*Mehrwortterm-Erkennung*	Ermittlung phrasaler lexikalischer Einheiten (feste Wendungen im Sinne von Idiomen/Phrasemen, vor allem Nominalgruppen)
6	*Eigennamen-Erkennung*	Ermittlung benannter Entitäten ('Named Entity Recognition'), evtl. verzahnt mit Mehrwortterm-Erkennung
7	*Satzerkennung*	Ermittlung einzelner Sätze innerhalb eines Absatzes (Zerlegung eines Absatzes in syntaktische Sätze)
8	*Wortarten-Erkennung*	Ermittlung von Wortkategorien (Part-of-Speech-Tagging: 'Markierung der Wortart'): Trennung von Funktionswörtern ('synsemantische' Wörter ohne eigene Bedeutung) und Inhaltswörtern ('autosemantische' Wörter mit eigener begrifflicher Bedeutung)

Phasen der Textverarbeitung

Phase	Aktion	Beschreibung
9	*Parsing*	Ermittlung der syntaktischen Struktur von Sätzen und morphologischen Struktur von Wörtern (aufwändig!); 'flaches' Parsing zur Entdeckung von Nominalgruppen (thematisch relevante Einheiten)
10	*Pronomen-Auflösung*	Ersetzung von Pronomen durch ihre sprachlichen Referenzausdrücke (aufwändig!)
11	*Disambi-guierung*	Ermittlung der korrekten Bedeutung (Lesart) eines Wortes im Satz bei mehrdeutigen (ambigen) Aus-drücken ('Word Sense Disambiguation')
12	*Inhalts-erkennung*	thematische Segmentierung von Texten über text-linguistische Kohärenzmuster (z. B. via WordNet)
13	*Übersetzung*	Überführung eines Textes in eine andere Sprache

Einsatz- und Aufgabenbereiche

- Anwendungsbereiche im Information-Retrieval:
  - Indexierung (engl. 'indexing'):
    - Automatische Ermittlung beschreibender, inhaltstragender Terme aus Texten (Deskriptoren), die für die thematische Suche indiziert werden müssen (ähnlich dem Index eines Bibliothekskatalogs oder Buches, evtl. ergänzt durch den automatischen Aufbau eines Lexikons/Thesaurus);
    - Indexe stellen ein 'Konzentrat' der Inhalte und Themen eines Dokuments oder einer Dokumentensammlung dar, um die dort enthaltene Information für Suchanfragen des Nutzers verfügbar zu machen.

Einsatz- und Aufgabenbereiche

- Zusammenfassung (engl. 'summarizing'):
  - Automatische Erzeugung eines Textkondensats aus dem Originaltext als möglichst kurze und prägnante Vorschau auf die Inhalte eines Textes (ähnlich einem thematischen 'Thumbnail' für ein Textdokument);
  - die Zusammenfassung dient dabei auch der Verbesserung der Suchleistung eines Information-Retrieval-Systems, da nur auf die wirklich wesentliche Information eines Textes zugegriffen werden muss.
  - Beispiel zur Veranschaulichung (s. nächste Seite): Abstracts in verschiedenen Sprachen, die den Textinhalt in eigenen Worten wiedergeben.

Einsatz- und Aufgabenbereiche

Bernd Ludwig und Jürgen Reischer, Regensburg

Informationslinguistik in Regensburg

In ihrem Beitrag stellen die Autoren die Informationslinguistik als Teildisziplin der Informationswissenschaft vor, grenzen sie gegen benachbarte Fächer Theoretische Linguistik, Computerlinguistik und Maschinelle Sprachverarbeitung ab, zeigen aber zugleich auch Gemeinsamkeiten und Überschneidungsbereiche auf. Anwendungsbereiche, Verfahren und Produkte der Informationslinguistik werden in einem kurzen Überblick eingeführt. Einige davon, die im Zentrum der Forschung an der Universität Regensburg stehen, werden unter Bezugnahme auf aktuelle Arbeiten und Forschungsprojekte näher erläutert.

Deskriptoren: Linguistik, Informationswissenschaft, Hochschulausbildung, Forschung, Informationslinguistik

Linguistic information science in Regensburg
In their contribution, the authors introduce linguistic information science as a sub discipline of information science. They distinguish it from the related subjects theo-

Mots-clés: Linguistique, sciences de l'information, enseignement supérieur, recherche

1 Was ist Informationslinguistik?

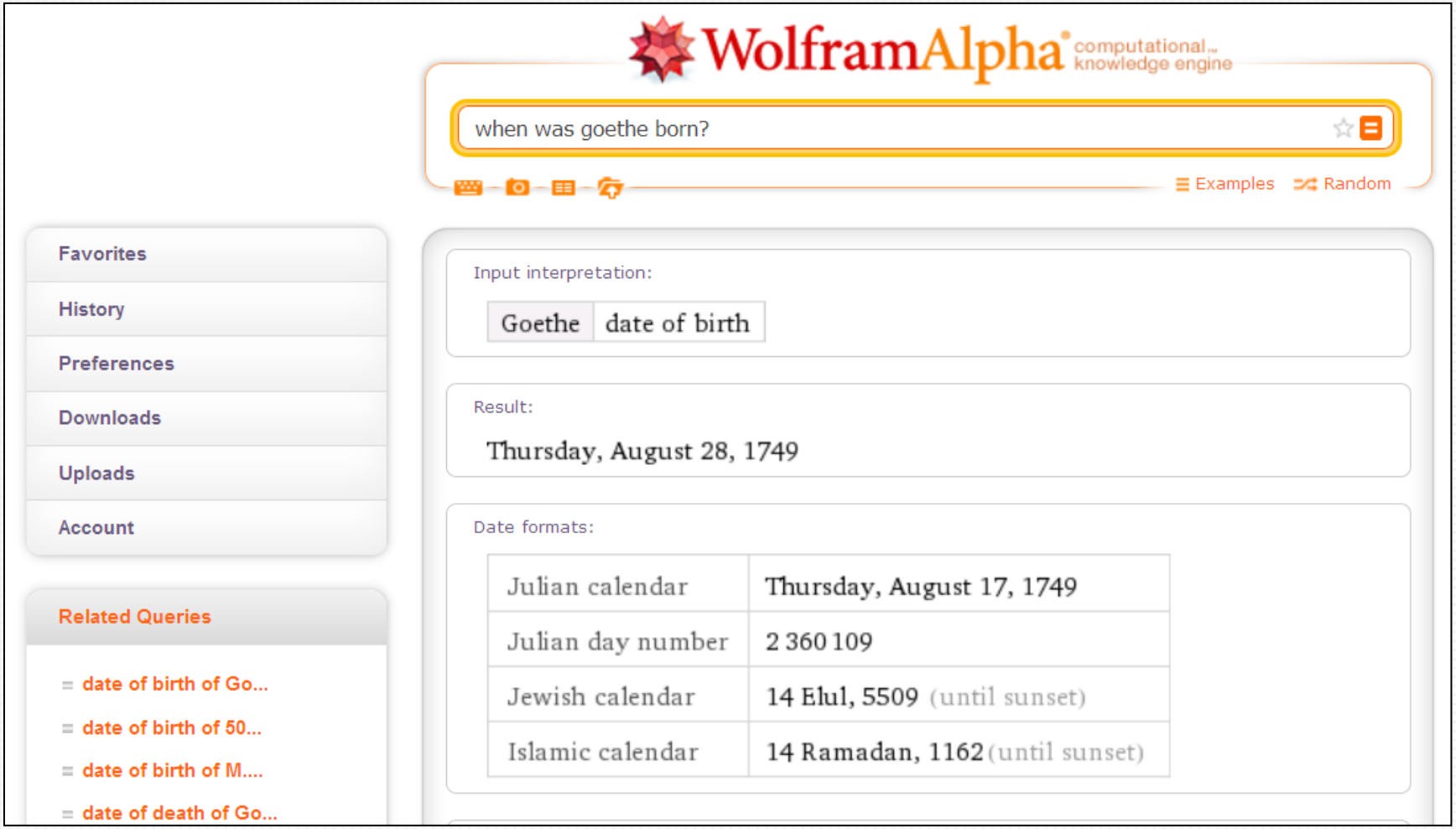
Informationslinguistik kann als diejenige Teildisziplin der Informationswissenschaft betrachtet werden, die sich mit der Verarbeitung natürlicher Sprache (Texte) in und für Informationssysteme befasst (vgl. [1], 2ff. [2], 219ff. [3], 1). In diesem Sinne spielt die Informationslinguistik eine entscheidende Rolle bei der Er- und Bearbeitung sprachlich verfasster Information (Texte) im Transformationsprozess von Daten oder Wissen zu Information. Transformationsprozesse finden sich in der Informationsverarbeitungskette bei der Erschließung (Analyse, Aufbereitung), Repräsentation (Aus-/Wiedergabe, Verwaltung) und Wiedergewinnung (Suche/‚Matching') von in

Ausschnitt aus Ludwig & Reischer (2012)

Einsatz- und Aufgabenbereiche

- Beauskunftung:
  - Automatisierung der Beantwortung allgemeiner oder häufig gestellter Faktenfragen (FAQ: 'Frequently Asked Questions') über eine natürlichsprachliche Schnittstelle (Auskunftssysteme, Expertensysteme, Frage-Antwort-Systeme [engl. QA- bzw. 'Question-Answering'-Systeme]);
  - dadurch können formalisierte (z. B. logische) oder rein stichwort-basierte Anfragen vermieden werden, d. h. die Schnittstelle zum Nutzer wird natürlicher und einfacher.
  - Beispiel zur Veranschaulichung: Faktenfrage "when was goethe born?" an das Frage-Antwort-System 'Wolfram Alpha'.

Einsatz- und Aufgabenbereiche

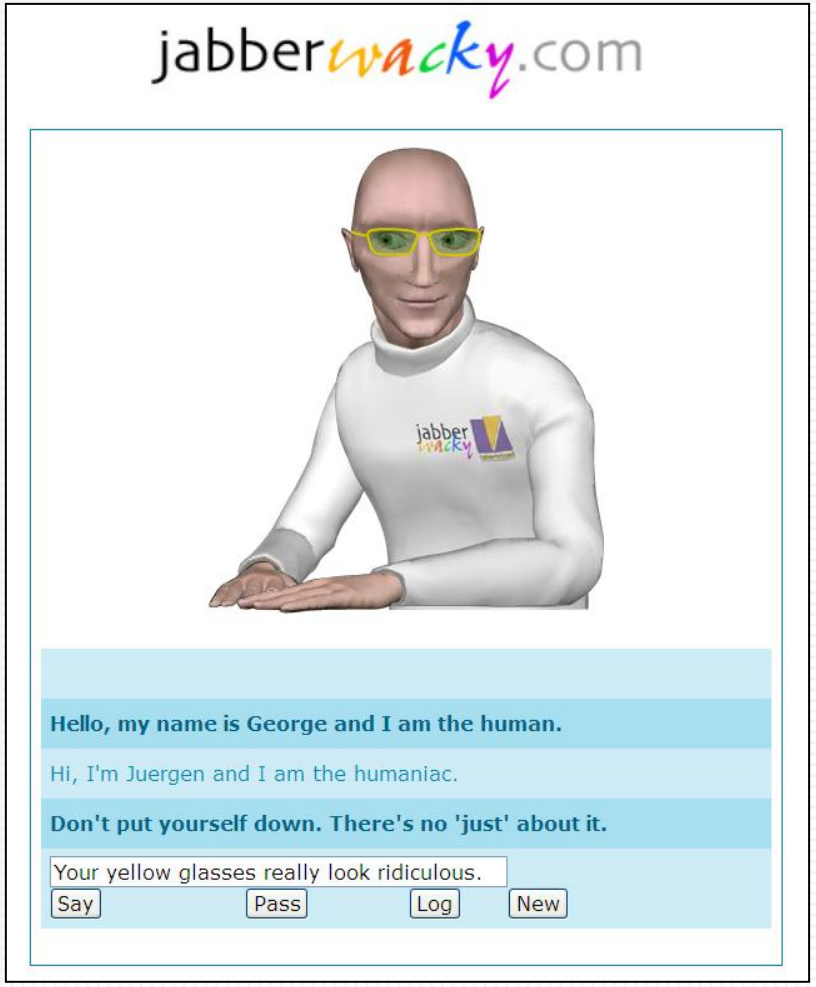


Screenshot www.wolframalpha.com (13.2.2013)

Einsatz- und Aufgabenbereiche

- Sprachinteraktion:
  - Automatisierung der Kommunikation bzw. sprachlichen Interaktion mittels Dialogsystemen oder Chatbots:
    - ernsthafte Anwendungen: gesprochensprachliche Navigation, Bestellungen/Reservierungen/Buchungen/Banking über Telefon;
    - spielerische Anwendungen: geschriebensprachlicher Smalltalk ('Chat') mit virtuellen Personen/Spielfiguren/Avataren (Simulation sprachfähiger Personen).
  - Generell Interaktion mit Informationssystemen, deren Ziel im Gegensatz zur Beauskunftung nicht primär in der Beantwortung von Fragen/Suchanfragen liegt.
  - Beispiel zur Veranschaulichung: Chatbot 'Stella' (ernsthaft) der UB Hamburg und Chatbot 'George' (spielerisch).

Einsatz- und Aufgabenbereiche



Screenshot www.jabberwacky.com (13.2.2013)

Screenshot www.sub.uni-hamburg.de/home.html (13.2.2013)

Überblick

- Themen des Kurses: Grundlagen und Anwendungen zur Verarbeitung und Analyse sprachlich kodierter Information:
  - Zeichenlehre (Semiotik):
    - Zeichenmodelle und Zeichentypen;
    - Zusammenhang Zeichen und Kommunikation;
    - Grundlegung der linguistischen Teilbereiche.
  - Lexik und Grammatik (lexikalische Einheiten und grammatikalische Regeln):
    - Allgemeiner Zusammenhang der linguistischen Teilbereiche untereinander und mit anderen Komponenten der menschlichen Kognition (Stellung der Sprache beim Denken und Handeln).

Überblick

- Linguistische Grundlagen:
  - Morphologie ('Formenlehre') und Syntax: Lehre von den elementaren Bausteinen der Sprache und den Regeln ihrer Zusammensetzung zu komplexeren Einheiten (Wort- und Satzbildung);
  - Semantik ('Bedeutungslehre'): Lehre von den Bedeutungen oder Funktionen sprachlicher Einheiten.
- Grundlegende Beschreibungsinstrumentarien/-modelle für Wörter und Sätze:
  - Reguläre Ausdrücke/Grammatiken: einfache Grammatiken zur Beschreibung von Wörtern und Wortbildungen bzw. Wortformenbildungen (Flexion);
  - Phrasenstrukturgrammatiken: einfache Grammatiken zur Beschreibung von Wortbildungen und Sätzen.

Überblick

- Begleitende Übungen zu allen Themen des Kurses:
  - Übungsaufgaben zur Vertiefung/Anwendung der theoretischen Grundlagen (z. B. Anwendungsfälle wie automatische Indexierung/ Zusammenfassung, Wortnetze);
  - Praxisteil zur grammatikalischen Beschreibung von Wörtern und Sätzen (unter Verwendung geeigneter Programme).
- Kursziel: Erwerb der Kompetenz zur Beschreibung und Analyse linguistischer Phänomene im Hinblick auf sprach- und informationswissenschaftliche Fragestellungen.
- Klausur als Abschlusstext mit Fragen im Stile der Übungs-aufgaben (idealerweise genügen die Übungsaufgaben als Klausurvorbereitung).

Ressourcen

- Internationale und deutsche Linguistik:
  - http://linguistlist.org;
  - http://www.linse.uni-due.de/linse;
  - http://www.dgfs.de (Deutsche Gesellschaft für Sprachwissenschaft).
- Informations- & Computerlinguistik bzw. Sprach- und Texttechnologie (Deutsch):
  - http://www.gscl.org (Gesellschaft für Sprachtechnologie und Computerlinguistik);
  - http://www.coli.uni-saarland.de/projects/stud-bib/ (Studienbibliographie Computerlinguistik).

Ressourcen

- Kursunterlagen:
  - Foliensätze: http://www.juergen-reischer.de (Sektion 'Manuskripte' in der unteren Hälfte);
  - Zusatzmaterialien: bei Bedarf auf Kurslaufwerk K: unter K:\PT\Infwiss\Kurse-JR\InfoLinguistik.
- Literatur:
  - Ludwig, B. & Reischer, J. (2012): Informations-linguistik in Regensburg. Information, Wissenschaft & Praxis 63(5), S. 292–296.