Informationsveranstaltung Linguistische Informatik Korpus- und Computerlinguistik

Prof. Dr. Stephanie Evert
Lehrstuhl für Korpus- und Computerlinguistik
http://www.linguistik.uni-erlangen.de/







Sie dürfen zwei Transitionen erleben!

Stefan → Stephanie

Linguistische Informatik → Computerlinguistik



Korpuslinguistik ≠ Computerlinguistik?

bzw. Sprachtechnologie (Natural Language Processing, NLP)

Was ist "Linguistische Informatik" und wie lange gibt es sie noch?





Was genau ist eigentlich ein Korpus?

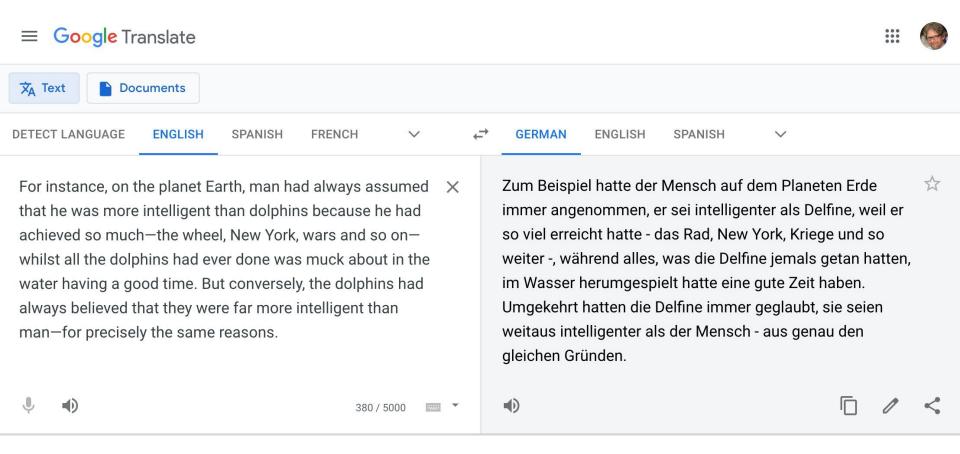
Das Korpus Die Korp**ora**

- Korpus¹ (im weiten Sinn)
 - = Sammlung von Sprachdaten / Texten in maschinenlesbarer Form
 - sehr große Korpora (≥ 100 M Wörter) sind besonders nützlich
 - Auswertung mit statistischen Methoden und maschinellen Lernverfahren
- Korpus² (im engen Sinn)
 - = Stichprobe authentischer Sprachdaten / Texte, die für eine bestimmte Sprache oder Sprachvarietät repräsentativ ist
 - z.B. literarische Korpora, Dialekte, gesprochene Sprache, IBK, ...
 - Basis für empirisch fundierte sprachwissenschaftliche Studien
 - zentral für Korpuslinguistik (im engen Sinn) und Digital Humanities
- Korpus- und Computerlinguistik
 - = Erstellung, maschinelle Verarbeitung und Auswertung von Korpora¹





Sprachtechnologie: Maschinelle Übersetzung





திருவனந்தபுரம்: கேரளாவின் பூரண மதுவிலக்கை நோக்கிய × பயணத்தின் ஒரு கட்டமாக மதுக்கடைகளுக்கு ஞாயிறு தோறும் விடுமுறை விடப் படுவது இன்று முதல் அமல் படுத்தப் பட்டுள்ளது. கேரளாவில் வரும் 10 ஆண்டுகளுக்குள் பூரண மதுவிலக்கை அமல் படுத்த அம்மாநில அரசு திட்டமிட்டுள்ளது. அதன் படி, ஆண்டு தோறும் பத்து பத்து சதவீதமாக மது விற்பனையைக் குறைக்க முடிவு செய்யப்பட்டுள்ளது. இதனால், கேரளாவில் புதிய மது பார்களுக்கு லைசென்சு வழங்கப்பட வில்லை. மேலும் 418 மது பார்களுக்கான லைசென்சு ரத்து செய்யப்பட்டு மது பார்களும் மூடப்பட்டன. காந்தி ஜெயந்தி தினமான கடந்த 2-ந் தேதி கேரளாவில் 10 சதவீத அரசு மதுக்கடைகள் மற்றும் மது பார்கள் முடப்பட்டதாக அம்மாநில அரசு தெரிவித்துள்ளது. இந்நிலையில், இத்திட்டத்தின் அடுத்த கட்டமாக ஒவ்வொரு ஞாயிற்றுக் கிழமையும் மது கடைகள், பார்களுக்கு விடுமுறை விட வேண்டும் என்றும் கேரள அரசு அறிவித்திருந்தது.

Thiruvananthapuram: Kerala ist ein Schritt auf dem Weg hin zu einer vollständigen matuvilakkai Bars über die Feiertage, als es in der Sonntags hat mit Wirkung ab heute eingeführt. Kerala, ist die Landesregierung plant, in den nächsten 10 Jahren zu implementieren, eine umfassende matuvilakkai. Dementsprechend wurde beschlossen, den Verkauf von Alkohol jährlich auf zehn Prozent zu reduzieren. So hat Kerala Lizenz für den neuen Wein-Bars gewährt. Siehe auch 418 Weinweinstuben geschlossen für die Lizenz wird abgebrochen. Gandhi Jayanti Tag, 10 Prozent der Bundesstaat Kerala auf den letzten 2 Bars und Weinstuben, hat die Landesregierung stillgelegt. In diesem Fall wird die nächste Phase des Projektes jeden Sonntag Spirituosen-Läden, Bars, und die Regierung hat erklärt, Kerala ein Feiertag sein sollte.









Künstliche Intelligenz: virtuelle Assistenten



IN 1939's CARTOON
"THE POINTER", THIS
GUY GOT A NEW,
MORE PEAR-SHAPED
BODY & PUPILS WERE
ADDED TO HIS EYES



Prof. Dr. Stefan Evert | Lehrstuhl Korpus- und Computerlinguis



Künstliche Intelligenz: virtuelle Assistenten

CONTRACT SAYS THAT WAGES WILL RISE OR FALL DEPENDING ON A STANDARD SUCH AS COST OF LIVING

THIS CLAUSE IN A UNION



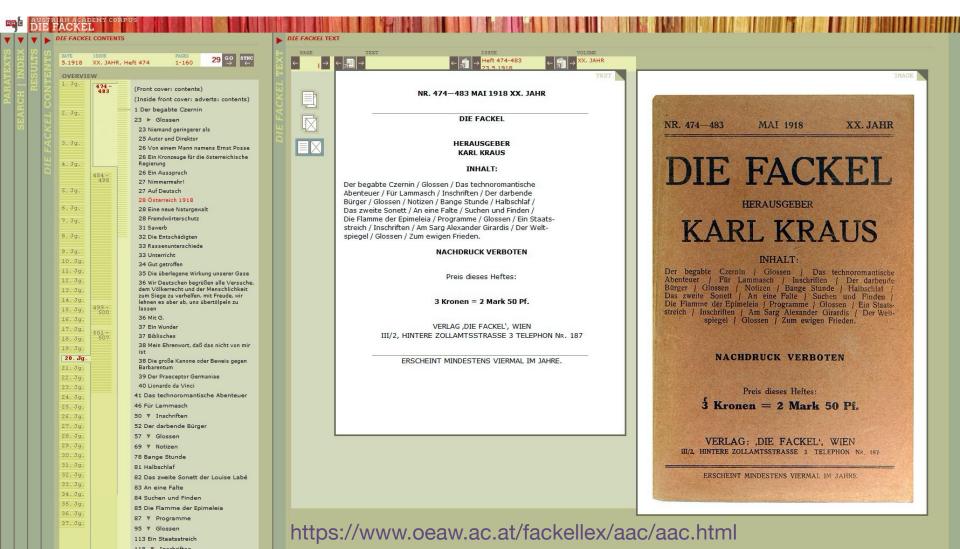


Prof. Dr. Stefan Evert | Lehrstuhl Korpus- und Computerlinguist



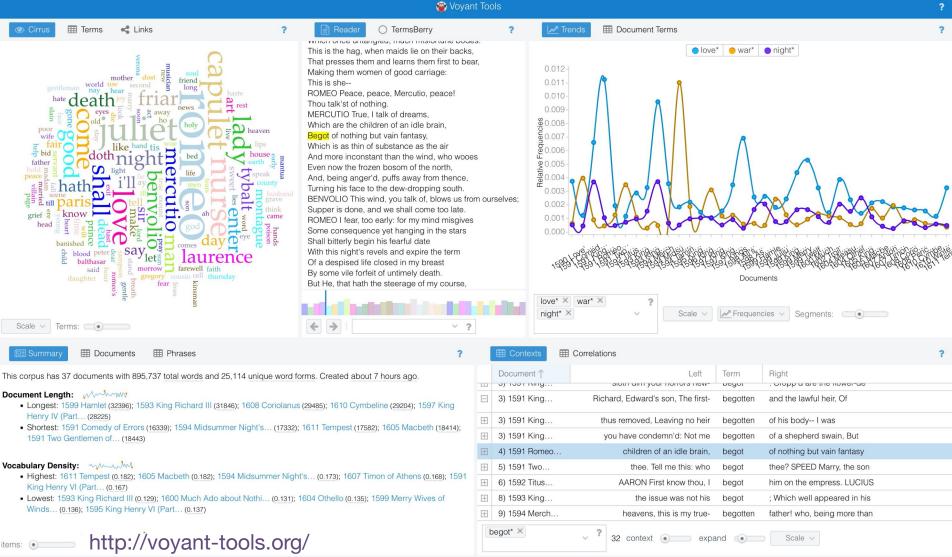


Digital Humanities: Digitale Editionen



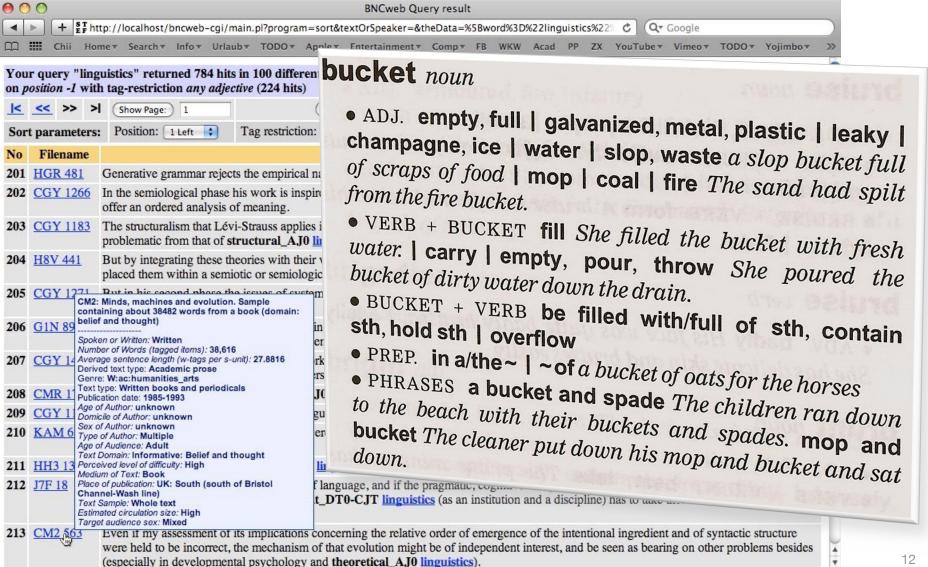


Digital Humanities: Analyse & Visualisierung



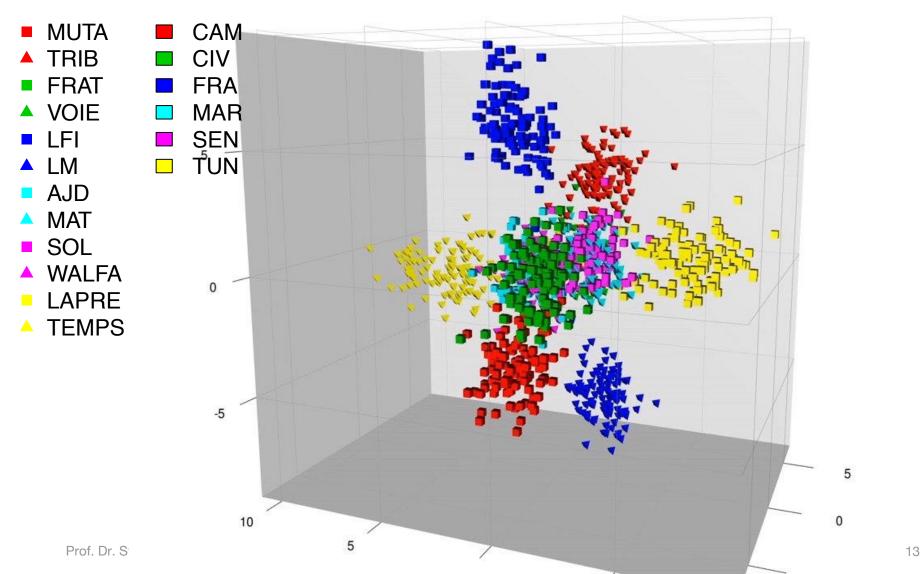


Sprachwissenschaft: Empirische Sprachbeschreibung



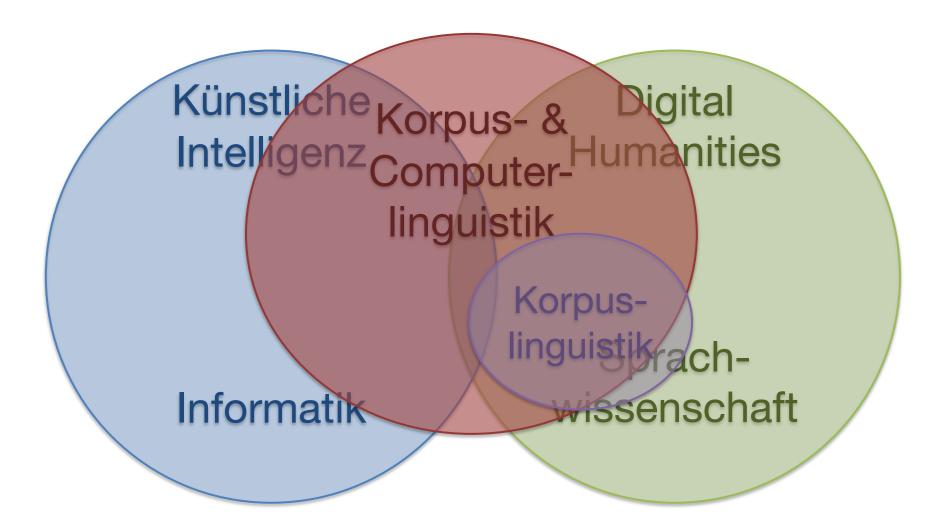


Sprachwissenschaft: Quantitative Linguistik





Wie war das jetzt mit Korpus- und Computerlinguistik?







Neuausrichtung des Studienfachs

- "Linguistische Informatik" → "Computerlinguistik"
- Fokus auf maschinelle Lernverfahren (insb. Deep Learning)
 - folgt dem Trend der modernen Computerlinguistik (und KI)
 - Schwerpunkt des Studiums: Operationalisierung von computerlinguistischen Aufgaben als Lernprobleme
 - "Baukasten" geeignet für zahlreiche Anwendungen in Sprachtechnologie, DH, Korpuslinguistik
- Programmierpraxis: Python
 - konsequenter Einsatz von Python und entsprechenden Frameworks im ganzen Studienverlauf
 - entsprechende Kenntnisse werden z.B. in praktischen Hauptseminaren vorausgesetzt
- Was wegfällt: Korpuslinguistik, Statistik
 - korpuslinguistische Theorie & Praxis sowie zugehörige Statistik
 → MA Linguistik, MA Digital Humanities
 - aber trotzdem intensive Arbeit mit Korpora als Trainingsdaten für maschinelle Lernverfahren, Basis für Text Mining, ...





Grundstudium

- Grundlagen der Computerlinguistik I III
 - GdCL I = klassische symbolische Ansätze, Mengenlehre, Logik
 - GdCL II = statistische Ansätze, maschinelle Lernverfahren, Wk-Theorie, lineare Algebra
 - GdCL III = Deep Learning = neuronale Netze, Tensoranalysis, praktische Implementierung
- Programmierung & Infrastrukturen I + II
 - Unix-Betriebssystem, Kommandozeile, Editor, reguläre Ausdrücke
 - Einführung in Python
 - Nutzung von Python-Bibliotheken und NLP-Standardwerkzeugen
- Einführungsmodul Linguistik
 - importiert aus Germanistik oder Anglistik
 - alternativ: DH-Modul 1 "Sprache & Text"
- Proseminar Computerlinguistik
 - Rezeption von Originalarbeiten, Präsentation, konstruktive Diskussion
 - Schreiben von Hausarbeiten, Literaturrecherche & Zitierung, LaTeX





Hauptstudium

- 3 Hauptseminare zu wechselnden Themen
 - 1x mit klassischer Hausarbeit
 - 1x mit mündlicher Prüfung
 - 1x mit praktischem Projekt + Projektbericht
- 1 Praxisseminar
 - angewandtes Gruppenprojekt (z.B. Teilnahme an Shared Task)
 - gemeinsamer Projektbericht in Form eines Konferenzbeitrags
 - Schwerpunkt auf Teamwork, Selbstorganisation, Arbeitsteilung, ...
- Angebot: mind. 2 Seminare / Semester
- Oberseminar (verpflichtend)
 - Besuch der Vorträge über 2 Semester
 - Essay über zwei ausgewählte Vorträge
- Praktikum (150h = 1 Monat Vollzeit)
 - bei externer Firma oder internes Forschungspraktikum





Erstfach ≠ Zweitfach

- Zweitfach: anwendungsorientierte Ergänzung zu geistes- oder sozialwissenschaftlichem Fach
 - komplett ohne Informatik-Importe (d.h. kein Gdl oder KonzMod)
 - Vermittlung der mathematischen und informatischen Grundlagen ganz auf CL zugeschnitten
- Erstfach: Interesse an Informatik und NLP-Methoden
 - + 20 ECTS statt Schlüsselqualifikationen
 - 7,5 ECTS Grundlagen der Informatik (verpflichtend)
 - 7,5 ECTS + 5,0 ECTS Wahlpflichtbereich
 - zahlreiche Module aus dem Bachelor Informatik stehen zur Wahl
 - Mathematik für Naturwissenschaftler (7,5) und Modellbildung & Statistik (5)

Studienplan **Linguistische Informatik**





PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

Semester 1 (WiSe)	Semester 2 (SoSe)	Semester 3 (WiSe)	Semester 4 (SoSe)	Semester 5 (WiSe)	Semester 6 (SoSe)
VL Grundlagen der CL 1 2 SWS; 2 ECTS	VL Grundlagen der CL 2 2 SWS; 2 ECTS	Proseminar Computerling. 2 SWS; 5 ECTS	HS Korpuslinguistik 2 SWS; 5 ECTS		•
Ü Grundlagen der CL 1 2 SWS; 3 ECTS	Ü Grundlagen der CL 2 2 SWS; 3 ECTS	Ü Computerling. Werkzeuge und Infrastrukturen 2 SWS; 5 ECTS	Ü Statistik 2 SWS; 5 ECTS		
Ü Arbeitstech- niken der CL 2 SWS; 2,5 ECTS	Grundseminar Programmierung 2 SWS; 5 ECTS	Aufbauseminar Programmierung 2 SWS; 5 ECTS	Hauptseminar theoretisch 2 SWS; 5 ECTS; Semester 4 oder 6	Hauptseminar praktisch 2 SWS; 5 ECTS	
Vorlesung Grundlagen der Informatik (GdI) 3 SWS; insg. 7,5 ECTS; Import		V Konzeptionelle Modellierung 2 SWS; 2,5 ECTS; Import; Sem. 2 – 4			
Tafelübung Gdl 2 SWS; s.o. ECTS; Import		Ü Konz. Mod. 2 SWS; 2,5 ECTS; Import; Sem. 2 – 4		Praktikur 1504-5-5	
Tutorensprechstd. Gdl 1 SWS; s.o. ECTS; Import				150h; 5 E	
12 SWS 15 ECTS	6 SWS 10 ECTS	10 SWS 20 ECTS	6 SWS 15 ECTS	2 SWS 5 ECTS	0 SWS 0 ECTS

Studienplan Computerlinguistik (Zweitfach)





Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Semester 6
(WiSe)	(SoSe)	(WiSe)	(SoSe)	(WiSe)	(SoSe)
VL Grundlagen	VL Grundlagen	VL Grundlagen	Hauptseminar I	Hauptseminar III	
der CL 1	der CL 2	der CL 3	2 SWS; 5 ECTS;	2 SWS; 5 ECTS;	
2 SWS; 2 ECTS	2 SWS; 2 ECTS	2 SWS; 3 ECTS	Hausarbeit	praktisches Projekt	
Ü Grundlagen	Ü Grundlagen	Ü Grundlagen	Hauptseminar II	Projektseminar	
der CL 1	der CL 2	der CL 3	2 SWS; 5 ECTS;	2 SWS; 5 ECTS;	
2 SWS; 3 ECTS	2 SWS; 3 ECTS	2 SWS; 7 ECTS	mündl. Prüfung	Teamprojekt	
S Grundkurs Programmierung 2 SWS; 5 ECTS	S Aufbaukurs Programmierung 2 SWS; 5 ECTS	Proseminar Computerling. 2 SWS; 5 ECTS	Oberseminar Com 2×1 SWS; 5 ECTS	puterlinguistik	
Einführung Linguistik 4 SWS; 5 ECTS; Import Angl./Germ.	DH-Modul 1: Sprache & Text 4 SWS; 5 ECTS; Import DGuS			Praktikum (extern 150 h = 1 Monat Vo unbenotet	•

10 SWS	6 SWS	6 SWS	5 SWS	5 SWS	0 SWS
15 ECTS	10 ECTS	15 ECTS	12,5 ECTS	12,5 ECTS	5 ECTS

Studienplan **Computerlinguistik (Erstfach)**





Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Semester 6
(WiSe)	(SoSe)	(WiSe)	(SoSe)	(WiSe)	(SoSe)
VL Grundlagen	VL Grundlagen	VL Grundlagen	Hauptseminar I	Hauptseminar III	
der CL 1	der CL 2	der CL 3	2 SWS; 5 ECTS;	2 SWS; 5 ECTS;	
2 SWS; 2 ECTS	2 SWS; 2 ECTS	2 SWS; 3 ECTS	Hausarbeit	praktisches Projekt	
Ü Grundlagen	Ü Grundlagen	Ü Grundlagen	Hauptseminar II	Projektseminar	
der CL 1	der CL 2	der CL 3	2 SWS; 5 ECTS;	2 SWS; 5 ECTS;	
2 SWS; 3 ECTS	2 SWS; 3 ECTS	2 SWS; 7 ECTS	mündl. Prüfung	Teamprojekt	
S Grundkurs Programmierung 2 SWS; 5 ECTS	S Aufbaukurs Programmierung 2 SWS; 5 ECTS	Proseminar Computerling. 2 SWS; 5 ECTS	Oberseminar Com 2×1 SWS; 5 ECTS	puterlinguistik	
Einführung Linguistik 4 SWS; 5 ECTS; Import Angl./Germ.	DH-Modul 1: Sprache & Text 4 SWS; 5 ECTS; Import DGuS		.1	Praktikum (extern 150 h = 1 Monat Vo unbenotet	

Grundlagen der Wahlpflichtbereich Informatik Informatik (GdI) ca. 6 SWS; 12,5 ECTS; Import Informatik (+ andere) 6 SWS; 7,5 ECTS; Import Informatik 10 SWS **12 SWS** 8 SWS 9 SWS 5 SWS 0 SWS 15 ECTS 17,5 ECTS 20 ECTS 20 ECTS 12,5 ECTS 5 ECTS



BA "Computerlinguistik" ab WS 2022/23

- Verfahren zur Änderung der Studien- und Prüfungsordnung läuft
- Studienanfänger*innen sollen in den BA Computerlinguistik wechseln
 - aber erstes Jahr nach PO des BA Linguistische Informatik
 - spezielle Angebote f
 ür Übergang (insb. Python-Programmierung)
 - GdI muss belegt werden (GOP!), ggf. Anrechnung als SQ
- Studierende im 3. Semester: Wechsel gut machbar
 - geeignete Anerkennungen von Modulen bis 4. Semester
 - v.a. zwei zusätzliche Hauptseminare + Oberseminar belegen
 - im Erstfach: Wahlbereich Informatik
- Studierende in höheren Semestern: eher nicht sinnvoll
 - Abschluss nach alter PO bleibt möglich (→ Äquivalenztabelle)





Übergangsregelungen

Computerlinguistik (neu)	Linguistische Informatik (alt)		
Bezeichnung	ECTS	Bezeichnung	ECTS
Grundlagen der Computerlinguistik I (traditionelle Verfahren)	5	Grundlagen der Computerlinguistik I	7,5
Programmierung und Infrastrukturen I	5	Programmierung I	5
Grundlagen der Computerlinguistik II (statistische Verfahren)	5	Grundlagen der Computerlinguistik II	5
Programmierung und Infrastrukturen II	5	Programmierung II	5
Grundlagen der Computerlinguistik III (Deep Learning)	10	Korpuslinguistik	10
Proseminar Computerlinguistik	5	Proseminar Computerlinguistik	5
Vertiefungsmodul Computerlinguistik I	5	Vertiefungsmodul Computerlinguistik theoretisch	5
Vertiefungsmodul Computerlinguistik II	5		-
Vertiefungsmodul Computerlinguistik III	5		-
Vertiefungsmodul Computerlinguistik	5	Vertiefungsmodul Computerlinguistik	5
Praktisch		praktisch	
Praktikum	5	Praktikum	5
Oberseminar Computerlinguistik	5		-
Grundlagen der Informatik (GdI) ¹	7,5	Grundlagen der Informatik (GdI)	7,5
Linguistische Grundkompetenzen	5	Werkzeuge und Infrastrukturen	5
Wahlpflichtmodul Informatik 1 ¹	7,5		-
Wahlpflichtmodul Informatik 2 ¹	5	Konzeptionelle Modellierung	5





Nach dem Studium

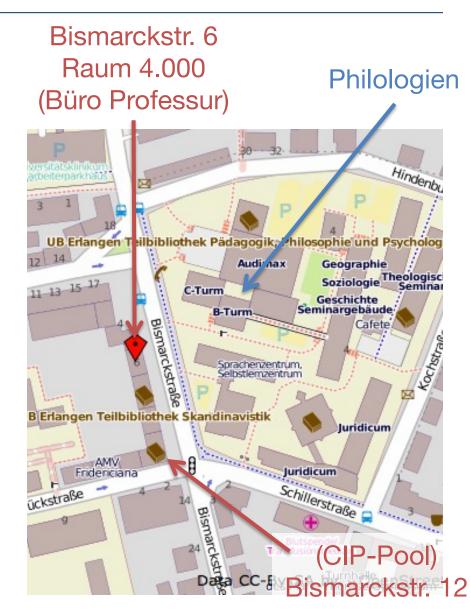
- Masterstudium / Promotion
 - Computerlinguistik
 - Sprachwissenschaft / Linguistik
 - Digital Humanities
 - Angewandte Informatik
- Tätigkeitsbereiche in der Wirtschaft (→ Sprachtechnologie)
 - Google, Microsoft, Facebook, Amazon, Twitter, Siemens, ...
 - Text Mining, Information Retrieval, Search Engines, ...
 - Lexikographie und Terminologie
 - Spracherkennung und Sprachsynthese, Dialogsysteme
 - Computergestützter Sprachunterricht (CALL)
 - Viele Start-Up-Unternehmen im IT-Bereich suchen Computerlinguisten!



Über uns

Unser Team

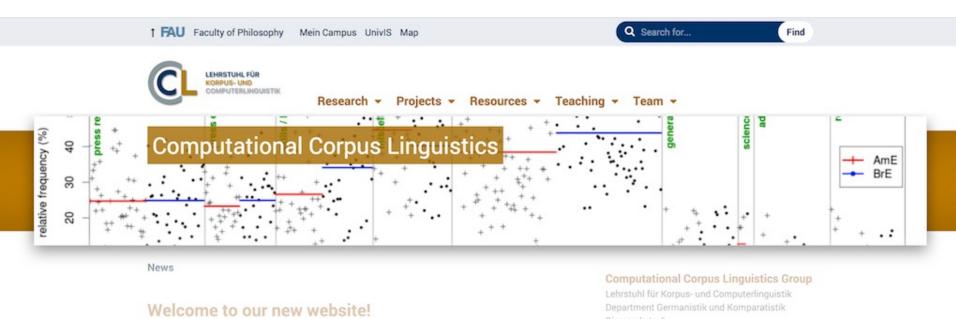
- Prof. Dr. Stephanie Evert stefan.evert@fau.de
- Dr. Besim Kabashi besim.kabashi@fau.de
- Dr. Thomas Proisl thomas.proisl@fau.de
- Philipp Heinrich, M.Sc. philipp.heinrich@fau.de
- Natalie Dykes, M.A. natalie.mary.dykes@fau.de
- Andreas Blombach, M.A. andreas.blombach@fau.de







Ganz wichtig: http://www.linguistik.fau.de/



- Sehr kleiner Studiengang (~ 50 Studierende) ... noch
- Aktive Fachschaftsinitiative:
 https://www.linguistik.phil.fau.de/teaching/fsi-computerlinguistik/
- FSI-Stammtisch
- Grundlagen- und Orientierungsstudium: https://www.ziwis.fau.de/lehreundangebote/grundlagen-und-orientierungsstudium/



Vorstellungsrunde









Stundenplan

siehe https://www.linguistik.fau.de/teaching/lehrveranstaltungen/

	Мо	Di	Mi	Do	Fr
08:00				08:15 - 11:45 Grundlagen der Informatik	
09:00				(Bauer) H7, H8, H10	
10:00		10:15 - 11:45 Konzeptionelle Modellierung		Arbeitstechniken der Computerlinguistik (Proisl) 0.320 Bismarckstr. 12	10:15 - 11:45 PS Aufbaukurs
11:00		(Lenz) H11		()	Programmierung Python (Proisl) 0.320 Bismarckstr. 12
12:00					
13:00					
14:00	14:15 - 15:45 Übung Grundlagen der	14:15 - 15:45 Computerlinguistische	14:15 - 15:45 Computerlexikographie		
15:00	Computerlinguistik 1 (Evert) 0.320 Bismarckstr. 12	Werkzeuge und Infrastrukturen (Kabashi) 0.320 Bismarckstr. 12	(Kabashi) 0.320 Bismarckstr. 12		
16:00	16:15 - 17:45 Übung Grundlagen der Computerlinguistik 1 (Evert) 0.320 Bismarckstr. 12	16:15 - 17:45 Proseminar Computerlinguistik (Evert) 0.320 Bismarckstr. 12	16:15 - 17:45 Oberseminar Computerlinguistik (Kabashi) 0.320 Bismarckstr. 12		
17:00	16:15 - 17:45 Wörter, Texte & Frequenzen: statistische Analyse von Sprachdaten (Blombach) 02.313				





Präsenzlehre (endlich!)

- Bis auf weiteres finden unsere LVen in Präsenz statt!
 - Wechsel zu virtueller LV kann einvernehmlich beschlossen werden.
- Hygienekonzept: 3G-Regelung
 - Nachweis: geimpft, genesen oder getestet
 - wird vor jeder LV kontrolliert (auch f
 ür Nutzung des CIP-Pools)
 - Maskenpflicht gilt auch am Sitzplatz
 - Kontaktdatenerfassung mit darfichrein.de (entfällt wohl ab 19.10.)
- Studierbarkeit soll gewährleistet bleiben
 - besondere Risikogruppe / keine Impfung möglich: Attest
 - wird als Sonderfall behandelt, z.B. Betreuung per E-Mail
- Aktuelle Info: https://www.fau.de/corona





Lehrangebot im WS 2021/22

- VL + Ü Grundlagen der Computerlinguistik 1
- Ü Arbeitstechniken der Computerlinguistik
- VL + TÜ + Tut Grundlagen der Informatik
- PS Computerlinguistik
- Ü Computerling. Werkzeuge und Infrastrukturen
- PS Aufbauseminar Python
- VL + Ü Konzeptionelle Modellierung
- Praktisches HS: Computerlexikographie
- Oberseminar Computerlinguistik
- S Wörter, Texte & Frequenzen
- VL + Ü Visualization
- VL + Ü Deep Learning
- S Forschungsdatenmanagement

• 1. Sem.

3. Sem.

5. Sem.

SQ



Pilotexperiment: Import-Hauptseminare

Alternative Hauptseminare (Anerkennung über Papierschein):

- Wörter, Texte & Frequenzen: statistische Analyse von Sprachdaten (Blombach, Heinrich | nur als praktisches HS)
- Chunks, constructions, and context how speakers express meaning (Herbst)
- "Künstliche Intelligenz" und juristische Argumentation Von Aristoteles zu Legal-Tech und Richterautomat?
 (Adrian | bereits ausgebucht, evtl. 1–2 Sonderplätze)
- Machine Learning: Introduction (Feigl, Löffler, Mutschler)
- Machine Learning: Advances (Feigl, Löffler, Mutschler)
- Automatic Analysis of Voice, Speech and Language Disorders in Speech Pathologies (Yang, Maier | Warteliste)





Informationen für Erstsemester Modul Grundlagen der Computerlinguistik I

- Vorlesung: Do 12:00 Mo 14:00 (Zeitfenster)
 - Bereitstellung als Screencasts von 45–60 Minuten Dauer
 - Dozent: Prof. Dr. Stefan Evert ⁶⁹
- Übung: Mo 14:15–15:45 / 16:15–17:45 (CIP-Pool)
 - Fragen und Diskussion zur Vorlesung
 - Veranschaulichung mit Beispielen und interaktiven Übungen
 - wöchentliche Übungsaufgaben zur schriftlichen Bearbeitung
 - Dozentin: Prof. Dr. Stephanie Evert
- Arbeitstechniken der CL: Do 10:15–11:45 (CIP-Pool)
 - Arbeit mit dem Betriebssystem Linux (inkl. Installation)
 - grundlegende Techniken und Werkzeuge für die computerling. Praxis
 - Dozent: Dr. Thomas Proisl
- **Tutorium**: n.V. (Tutor: Timm Weber)





Informationen für Erstsemester Modul Grundlagen der Informatik

- Vorlesung: asynchron
 - Bereitstellung als Videoaufzeichnungen
 - können "asynchron" angesehen werden
- Tafelübung: Do 10:15–11:45 (Livestream)
 - Fragen können vorab im Chat gestellt (und z.T. beantwortet) werden
 - Tafelübung als Livestream, der auch aufgezeichnet wird
 - kann vollständig asynchron belegt werden
- Tutorensprechstunden: viele Kleingruppen (Chat / virtuell)
 - Fragen zu Vorlesung, Tafelübung und Übungsaufgaben





Nächste Schritte

- LVen werden in der Regel durch StudOn-Kurse begleitet
 - Wichtig: StudOn-Kurs für Grundlagen der Computerlinguistik 1 https://www.studon.fau.de/crs4154004.html
 - dort aktuelle Informationen, Materialien, Zugang Videos + Zoom
- Gruppeneinteilung für die Übungen: jetzt!
- Lehrveranstaltungen in dieser Woche
 - Vorlesung: 1. Screencast bis n\u00e4chsten Montag schauen!
 - Donnerstag 21.10.: ATCL, Montag 25.10.: Übung Computerlinguistik
- Prüfungsanmeldung über meinCampus nicht vergessen!
 - Anmeldungszeitraum: 3 Wochen ca. Ende Nov. Anfang Dez.
- Accounts für CIP-Pool = IdM-Kennung
- Sprechstunde Prof. Dr. Stephanie Evert: hybrid (→ StudOn)
 - montags 10:00–11:00 (Voranmeldung) + 11:00–12:00 (offen)



Fragen?