

DOKTORATSPROGRAMM

18./19. November 2021, Institut für Mehrsprachigkeit, Fribourg

Mehrsprachigkeit: Erwerb, Bildung und Gesellschaft

Universität Freiburg (CH) | Institut für Mehrsprachigkeit

Workshop: Digitale Korpuslinguistik der Schweizer Landessprachen

Prof. Dr. em. Jürgen Rolshoven, Universität zu Köln

Moderation: Prof. Dr. Matthias Grünert, Universität Freiburg (CH)

Zielsetzungen des Workshops

Der Workshop behandelt exemplarisch sprachliche Daten der Landessprachen Deutsch, Rätoromanisch, Französisch und Italienisch einschliesslich der jeweiligen Dialekte bzw. regionalen Varietäten. Er wendet sich an Forschende und Studierende aus den Sprach- und Literaturwissenschaften und aus der Sprachdidaktik unter besonderer Berücksichtigung der Mehrsprachigkeit. Der Workshop verbindet in seinen Schwerpunkten korpuslinguistische Grundlagen und Prinzipien mit aktiver Umsetzung und Anwendung; er nimmt Fragestellungen der Teilnehmenden auf und trägt zur Erweiterung von Problemlösekompetenzen bei.

Thematische Schwerpunkte

1. Digitale Textarchive und Korpora der Landessprachen

Die grundlegende Datenbasis für korpuslinguistische Fragestellungen sind digitale Texte. Sie sind in unterschiedlicher Weise ausgezeichnet. Systematisch nach etablierten und konventionalisierten Regeln ausgezeichnete digitale Texte stellen Korpora im engeren Sinne dar. Gegenstand des ersten Schwerpunkts ist die Erstellung digitaler Texte durch Verfahren maschinellen Textlesens (OCR Optical Character Recognition), die Präsentation vorhandener digitaler Texte der Landessprachen und die Prozessierung dieser Texte für Konkordanzen, Mustersuchen und Musterbildung.

2. Auszeichnungen und Auszeichnungssysteme

Dem folgt die Aufbereitung von Texten durch Auszeichnung, beispielsweise für morphologische oder syntaktische Fragestellungen, für die die ausgezeichneten Texte weiter prozessiert werden. Als Auszeichnungsformat hat sich ein Formalismus der Text Encoding Initiative (TEI) etabliert. Dabei werden sowohl bereits vorhandene TEI-ausgezeichnete Korpora als auch die Erzeugung von ausgezeichneten Texten zum Gegenstand.

Doktoratsprogramm Mehrsprachigkeit

Der Workshop ist Teil des Doktoratsprogramms Mehrsprachigkeit: Erwerb, Bildung und Gesellschaft (Modul B, angew. Linguistik). Es wird vom Institut für Mehrsprachigkeit und dem Studienbereich Mehrsprachigkeitsforschung und Fremdsprachendidaktik der Universität Freiburg in enger Zusammenarbeit mit dem Center for the Study of Language and Society der Universität Bern sowie Professor*innen der Universitäten Genf und Lausanne durchgeführt.

Teilnahme und Anmeldung

Teilnahmeberechtigt sind in erster Linie Doktorierende der Universität Freiburg und der Partneruniversitäten Bern und Lausanne. Das Angebot kann auf Doktorierende anderer Universitäten ausgeweitet werden.

Anmeldungen sind bis zum 24. Oktober 2021 an folgende Adresse zu richten:
philippe.humbert@unifr.ch

Ein Abstract zum eigenen Doktoratsprojekt von ½ – max. 1 Seite ist der Anmeldung beizulegen.

Ort

Institut für Mehrsprachigkeit, Murtengasse 24 / Rue de Morat 24, 1700 Fribourg, Raum K1.03

→ www.institut-mehrsprachigkeit.ch

3. Prozessierungen und Auswertungen

Die ausgezeichneten Texte bilden die Grundlage für weitere, spezialisierte Verarbeitung und Auswertung. Dabei führen quantitative Verfahren zur Beantwortung qualitativer Fragestellungen unter möglichem Einschluss einer größeren Zahl von Faktoren und deren Zusammenhängen, die sich sehr viel leichter maschinell erschließen lassen.

4. Semantik und Sprachsysteme

Der Schwerpunkt Semantik und Sprachsysteme beleuchtet Formen der Musterbildung und des maschinellen Lernens in natürlichen Sprachen auf Grund digitaler Korpora vor dem Hintergrund erweiterter linguistischer Notationen und Theoriebildungen; er verbindet eine forschungsorientierte Semiotik mit außerordentlich erfolgreichen Anwendungen IT-industrieller Natur. Anwendungen dieser Art sind z.B. die Suchmaschinen von Google oder Facebook, maschinelle Übersetzungen von Google Translate oder DeepL, die unter dem Schlagwort Künstliche Intelligenz vermarktet werden. Die Verbindung zu einer semiotisch basierten Linguistik fördert nicht alleine linguistische Erkenntnis, sondern bietet auch die Chance, implizites, in neuronalen Netzen verborgenes Wissen mit linguistischen Mitteln symbolisch und somit explizit verifizierbar zu machen oder – im Gegenteil – verborgene Fehler durch explizite Falsifikation aufzudecken.

5. Potentiale für Forschungsfragen und -themen der Teilnehmenden

Die Interessen und Fragestellungen der Teilnehmenden, die im Laufe des Workshops präsentiert werden, werden abschliessend hinsichtlich ihrer Lösungspotentiale diskutiert.

Über Prof. Dr. em. Jürgen Rolshoven

Studium der Romanistik, Allgemeinen Sprachwissenschaft und Geschichte an der Universität zu Köln, 1977 Promotion in Computerlinguistik, 1984 Habilitation über Selbstlernende Systeme. Von 1990–2018 war Jürgen Rolshoven Professor für Sprachliche Informationsverarbeitung und Leiter der mit seiner Berufung gegründeten Abteilung für Sprachliche Informationsverarbeitung der Universität zu Köln. Zu seinen Schwerpunkten gehören maschinelle Sprachverarbeitung, maschinelles Lernen, Selbstorganisation, e-learning und Digitalisierung. In diesen Bereichen hat er verschiedene Projekte realisiert, unter anderem die Annotierte Rätoromanische Chrestomathie und Pledari Grond (Interaktives Wörterbuch Deutsch-Rumantsch Grischun). Nach seiner Emeritierung ist er weiterhin mit verschiedenen Projekten tätig, unter anderem im Zusammenhang mit dem Rätoromanischen (Wörterbücher der Idiome und des Rumantsch Grischun, Digitalisierung).

Programm

Donnerstag, 18. November

9.00 – 09.15	Begrüssung und Einführung
9.15 – 10.00	Bestehende Sprachkorpora
10.00 – 10.30	Präsentation von Doktoratsprojekten
10.30 – 11.00	Kaffeepause
11.00 – 12.30	Korpuserstellung Tools und Anwendungen
12.30 – 13.45	Mittagspause
13.45 – 14.45	Auszeichnungen und Auszeichnungssysteme Part of Speech Tagging
14.45 – 15.15	Präsentation und Diskussion von Doktoratsprojekten
15.30 – 16.30	Syntax und Strukturen
16.30 – 17.00	Kaffeepause
17.15 – 18.45	Öffentlicher Vortrag von Prof. Rolshoven

Freitag, 19. November

9.00 – 10.00	Prozessierungen Konkordanzen und reguläre Ausdrücke
10.00 – 10.30	Diskussion von Doktoratsprojekten
10.30 – 11.00	Kaffeepause
11.00 – 12.30	Vektoren (Word2Vec)
12.30 – 14.00	Mittagspause
14.00 – 15.30	Programmieren im Grossen: Pipes (batches)
15.30 – 16.00	Kaffeepause
16.00 – 17.00	Lösungspotentiale für die Doktoratsprojekte

Covid-19

Aufgrund der weiterhin bestehenden Unsicherheit der epidemischen Lage sind kurzfristige Anpassungen des Programms oder Schutzmassnahmen möglich. Bitte erkundigen Sie sich auf unserer Website

► www.institut-mehrsprachigkeit.ch
-> Agenda