XX. LIPP Symposium 2014

Linguistik 2.0 – Die Herausforderung der „Digital Humanities“

Programm

Mi, 5.2.2014

18 Uhr	Felix Wichmann (Universität Tübingen) *Machine learning methods for system identification in sensory psychology*
19 Uhr	Umtrunk

Donnerstag, 6.2.2014

9.00 - 9.45	Hans-Jörg Schmid (LMU) & Annette Mantlik (Universität Heidelberg) *Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles*
9.45 – 10.30	Klaus Schulz (CIS, LMU) *Zur Nachkorrektur historischer OCR-erfasster Texte*
11.00 – 11.45	Susanne Oberholzer (Universität Zürich) *PRAAT, F5, soscisurvey.de, sql – Das Zusammenspiel von Sprachgebrauchs- und Spracheinstellungsforschung mit neuen Technologien*
11.45 – 12.30	Christian Riepl (ITG, LMU) *Das Projekt „Biblia Hebraica transcripta“: Eine kollaborative Forschungsdatenbank*
	Mittagspause
14.00 – 14.45	Katharina Franko (LMU) *":D Ciao ciao Italia hallo Deutschland! <3"* *Eine Analyse deutsch-italienischer Facebook-Statusmitteilungen*
14.45 – 15.30	Daphné Kerremans (LMU) *The NeoCrawler: Identifying English Neologisms on the Internet and Investigating their Diffusion in the Online Speech Community*
	Kaffeepause

16.00 – 16.45	Thomas Krefeld (LMU) & Desislava Zhekova (LMU) *Quantitative Typology and Dialect Dynamism: Analysis of Data from Contemporary Media*
16.45	Ausblick

Linguistik 2.0

Die Schlagworte, mit denen die großen wissenschaftsgeschichtlichen Strömungen belegt werden, sind in der Regel ambivalent, denn sie bezahlen ihre schnelle und flächendeckende Verbreitung mit erheblichen inhaltlichen Unschärfen. Der Ausdruck 'digital humanities' ist in dieser Hinsicht paradigmatisch: Er bezeichnet einerseits alle Optionen, die sich für die Geistes- und Sozialwissenschaften aus den so genannten Neuen Medien ergeben (neue Formen der sprachlichen Kommunikation, neue Wege der Datenerhebung und -analyse, neue Möglichkeiten der Forschungskooperation) und andererseits die daraus entstehende Notwendigkeit neuer theoretischer Modellierungen. Im Übrigen ist der Ausdruck erst aufgekommen, nachdem sich ganze Forschungsrichtungen, wie zum Beispiel die Korpuslinguistik, längst konstituiert hatten; digitale Techniken wurden oft mit großer Selbstverständlichkeit adaptiert und integriert, ohne dass der Eindruck eines wirklichen Paradigmenwechsels entstanden wäre. Mit 'digital humanities' wird also keineswegs ein Beginn, sondern ein progressiver Übergang markiert, der sich vor allem mit der Durchsetzung der sozialen Medien (Web 2.0) radikal beschleunigt hat. Zweck des Symposiums ist es, aktuelle sprachwissenschaftliche Projekte zu präsentieren, die überhaupt nur unter den medialen Bedingungen der letzten Jahre möglich wurden, und sie gleichzeitig in ihrer Eigenständigkeit zu reflektieren.

Abstracts

Felix Wichmann (Tübingen)

Machine learning methods for system identification in sensory psychology

Over the last decades research in sensory psychology has witnessed a transition from the phenomenological descriptions of perception to its quantitative analysis. Ultimately, we strive not only to re-describe perception in quantitative or statistical terms, but we aspire to quantitative process models of perception. This last step has thus far proven difficult, because as a prerequisite for psychophysical process models it is necessary to know to what extent decisions in behavioral tasks depend on specific stimulus features, the perceptual cues: Given the high-dimensional input---images, sequences of images or sound streams---which are the stimulus features the sensory systems base their computations on? In engineering terms, we require tools for system identification in sensory psychology. Over the last years I was involved in the development of inverse machine learning methods for (potentially nonlinear) system identification in sensory psychology, and we applied our methods to identify regions of visual saliency (Kienzle et al., 2009), to gender discrimination of human faces (Wichmann et al., 2005; Macke & Wichmann, 2010), and to the identification of auditory tones in noise (Schönfelder & Wichmann, 2012; 2013). In my

presentation I will concentrate on how stimulus-response data can be analyzed, and how to prevent both over-fitting to noisy data and how to enforce sparse solutions.

Hans Jörg Schmid (München) & Annette Mantlik (Universität Heidelberg)

Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles

The talk aims to demonstrate the potential of analyzing large datasets extracted from the Internet for investigating the linguistic knowledge of authors from earlier periods. It implements the traditional approach in historical corpus linguistics but goes beyond it in two fundamental ways. Firstly, it investigates usage patterns of individual authors in relation to general contemporary trends, taking into consideration typical formal, semantic and pragmatic characteristics of the given construction at the time. And secondly, it investigates cases where a given construction is **not** used by an author, even though it could or 'should' have been used if one considers the average frequency of a construction at a given time, given certain communicative intentions and text-types.

More than 800 attestations of the construction N+BE+*that* (e.g. *the problem was that …, the truth is that …*) dating from 1384 to 1871 were harvested from the following corpora available on or extracted from the Internet:

1. The quotation database of the online version of the *Oxford English Dictionary* (*OED3*)
2. Texts downloaded from *Project Gutenberg* amounting to approx. 19 million words
3. The *Parsed Corpus of Early English Correspondence* (PCEEC)
4. The *Old Bailey Corpus* (OBC)
5. The *Helsinki Corpus*
6. The *Paston Letters*
7. Letters by Jane Austen downloaded from *Oxford Text Archive*

Each of these attestations was coded with regard to the following variables:

1. CORPUS (source)
2. DATE
3. NOUN LEMMATIZED (normalized form of the noun)
4. NOUN TYPE (semantic; on two levels of granularity)
5. TEXT TYPE (on two levels of granularity)
6. TYPE OF DETERMINER (on two levels of granularity)
7. FUNCTION (reporting, narrative, argumentative, descriptive, explanatory)
8. INTERVENING MATERIAL BETWEEN N AND *THAT* (yes/no)
9. TENSE (present/past)
10. HAPAX (yes/no)
11. NUMBER OF TOKENS OF THE NOUN IN THE DATASET

With the help of a student project group from the Institute of Statistics, LMU, three types of statistical analyses were performed to analyze this dataset:

1. Following Gries and Hilpert (2008 and 2012), an agglomerative hierarchical cluster analysis inductively generating historical periods from the data (rather than superimposing conventional periods or 50-year brackets)
2. A logistic regression modelling variance and central tendencies during these periods.

3. An original method allowing a comparison of individual authors with general tendencies of periods.

In addition to the quantitative analysis, a detailed qualitative analysis of the data of selected authors is carried out. The combination of both types of analysis unveils interesting usage patterns which provide insights into the ways in which knowledge of the construction was entrenched in the minds of historical authors. Implications of the results for the study of language change are also discussed.

Klaus Schulz (CIS, LMU)

Zur Nachkorrektur historischer OCR-erfasster Texte

Bei der textuellen Erfassung historischer Dokumente mittels Verfahren der Optischen Charaktererkennung (OCR) treten oft viele Erkennungsfehler auf.

Diese Fehler behindern viele Formen der computergestützten Textanalyse und können dazu führen, dass die Texte im Rahmen geistes- oder sozialwissenschaftlicher Untersuchungen schlecht oder gar nicht nutzbar sind. Da das rein manuelle Erfassen der Texte (durch "double keying") zu teuer ist, bietet sich eine manuelle Nachkorrektur der OCR-erfassten Texte als Option an. Wir stellen ein am CIS entwickeltes Open Source Tool zur interaktiven Postkorrektur von OCR-erfassten historischen Dokumenten vor. Zwei Eigenschaften des Systems tragen in besonderer Weise zur Effizienz der Nachkorrektur bei. (1) Eine graphische Benutzerschnittstelle ermöglicht diverse parallele Sichten auf OCR-Textanteile und korrespondierende Teile der Seitenbilder. (2) Spezielle Sprachtechnologie, die im Hintergrund eingesetzt ist, berechnet ein statistisches Modell für (vermutete) häufige OCR-Fehler des Eingabedokuments und dort auftretende spezielle historische Schreibweisen. Basierend auf diesem Modell werden vermutete OCR-Fehler und Serien ähnlicher Fehler im Eingabedokument dem Benutzer angezeigt. Konkordanzsichten zur Überprüfung vermuteter Fehlerserien erlauben eine rasche Inspektion und eine Korrektur in einer Aktion. Praktische Benutzertests an drei europäischen Großbibliotheken haben ergeben, dass das Tool signifikant zu einer beschleunigten Korrektur historischer OCR-erfasster Texte beiträgt. Das Tool ist unter Github frei erhältlich.

Susanne Oberholzer (Universität Zürich)

PRAAT, F5, soscisurvey.de, sql – Das Zusammenspiel von Sprachgebrauchs- und Spracheinstellungsforschung mit neuen Technologien

Das Zusammenspiel von Sprachgebrauchs- und Spracheinstellungsforschung mit neuen Technologien soll am Beispiel eines aktuellen Dissertationsprojektes vorgestellt werden, das in dieser Form nur dank der neuen Technologien durchgeführt werden konnte. In diesem Projekt werden Sprachgebrauch und Spracheinstellungen in der Deutschschweiz untersucht. Ziel des Projektes ist es aufzuzeigen, wie die beiden Varietäten Dialekt und Standarddeutsch in einem speziellen Kontext, dem der Kirchen, von den Sprecherinnen und Sprechern eingesetzt werden und welche Rolle die Spracheinstellungen für die Wahl der jeweiligen Varietät spielen. Dafür wurden sowohl qualitative wie quantitative Daten erhoben.

Zur Datenerhebung und -auswertung kamen folgende Methoden und Tools zum Einsatz: Zur Untersuchung des Sprachgebrauchs wurden Tonaufnahmen von reformierten und katholischen Gottesdiensten (in verschiedenen Deutschschweizer Kantonen) im mp3-Format erstellt. Diese wurden in PRAAT transkribiert und annotiert. Die Sprachgebrauchsdaten außerhalb des konkreten Gottesdienstes sowie Spracheinstellungen der Pfarrerinnen und Pfarrer, die die aufgezeichneten Gottesdienste geleitet hatten, wurden anhand von leitfadengesteuerten Interviews erhoben. Diese Interviews wurden ebenfalls aufgezeichnet (mp3). Die Aufnahmen wurden anschließend in F5 transliteriert und getagged. Schließlich wurde zur Validierung dieser kleinen Stichprobe in fünf Kantonen der Deutschschweiz eine groß angelegte Fragebogenerhebung bei allen reformierten Pfarrerinnen und Pfarrern durchgeführt, die sowohl Fragen zum Sprachgebrauch als auch zu Spracheinstellungen enthielt. Die Daten wurden online über das Tool www.soscisurvey.de erhoben.

Die Daten aus PRAAT (Gottesdienst-Transkripte), aus F5 (Interview-Transkripte) sowie aus soscisurvey.de (Antworten aus Onlinebefragung) wurden allesamt in eine sql-Datenbank überführt. Mithilfe der Datenbank können die zur Analyse der Daten nötigen Abfragen getätigt werden, so sind beispielsweise Code-Switchings in den Gottesdiensten leicht zu eruieren und zu kategorisieren (Zitat vs. freie Rede, Sprecher, Gottesdienstteil), oder wo nötig (quantitative Frageblöcke der Onlinebefragung) zur weiteren Bearbeitung nach SPSS exportiert werden.

Es soll in diesem Beitrag also aufgezeigt werden, wie das Zusammenspiel von klassischen Erhebungsmethoden (Tonaufnahmen, Interviews, Fragebogen) mit modernen Technologien eine in gewissen Teilen automatisierte und schnelle Auswertung der Daten ermöglicht.

Christian Riepl (ITG, LMU)

Das Projekt „Biblia Hebraica *transcripta*“: Eine kollaborative Forschungsdatenbank

Der Vortrag stellt das Mitte der 1980er Jahre von Wolfgang Richter initiierte Projekt „Biblia Hebraica *transcripta*“ vor, führt in den aktuellen Stand der Forschungsdatenbank BHtDB 3.0 ein und zeigt deren Potenzial im Kontext der „Digital Humanities“.

In einem reflektierenden Rückblick auf die ersten Projektphasen, insbesondere unter dem Aspekt der Kooperation einer geisteswissenschaftlichen mit informatischen Disziplinen, werden zunächst Transkription, Zeichenkodierung, Datenstrukturen, Referenzsystem, Segmentierung und Tokenisierung behandelt. Sodann werden die auf den gesamten Textkorpus angewendeten automatischen Analyseverfahren zur Morphologie und Morphosyntax sowie die rechnergestützt-manuell durchgeführte Satzanalyse erläutert. Schließlich wird gezeigt, wie und in welchem Umfang die Daten der sprachwissenschaftlichen Analyse auf den Beschreibungsebenen Wort, Wortfügung und Satz in einer relationalen Datenbank abgebildet sind. Zugleich wird das langfristige Potenzial sichtbar, das die strukturierten Daten in sich bergen und das sich noch einmal immens weitet, wenn Verfahren einbezogen werden, die auf Crowdsourcing basieren, um damit eine kollaborative Forschungsplattform vorzubereiten. Sprache und Literatur des Alten Testaments bieten ideale Voraussetzungen, um Verfahren des kollaborativen Arbeitens, der heuristischen und systematischen Berechnung des Datenbestandes unter variablen Bedingungen, der statistischen Modellierung und Visualisierung sowie der Verknüpfung mit Datenbanken z.B. der Archäologie, Kunstgeschichte und Musikwissenschaft zu entwickeln und zu erproben.

Daphné Kerremans (LMU)

The NeoCrawler: Identifying English Neologisms on the Internet and Investigating their Diffusion in the Online Speech Community

Despite the ubiquity of natural language, empirically-oriented linguistic studies have faced many challenges in compiling a sufficiently large, authentic, representative and easy to handle language sample as required by good scientific practice. If the development and availability of large electronic corpora constituted the first wave of (r)evolution in 20th century linguistic methods a few decades ago, the more recent commercialisation of the Internet certainly marks the second wave. Unlike neatly organised and carefully balanced corpora however, using Web material for linguistic purposes often feels like looking for the needle in the digital haystack. The useable language nuggets tend to be entangled in a jumble of linguistically irrelevant data such as photos, videos and programming code. One of the possible solutions is offered by the webcrawling technique, i.e. automatically trawling the Internet and extracting the desired material (e.g. Fletcher's KWiCFinder, 2001 or WebCorp by RDUES, Birmingham City University, 1999).

The NeoCrawler, built at the LMU by Susanne Grandmontagne, Hans-Jörg Schmid and myself, is precisely such a webcrawler, designed to study the linguistic behaviour of English neologisms on the Internet (cf. Kerremans, Stegmayr and Schmid 2012). The NeoCrawler differs from the publicly available webcrawlers with regard to three important aspects. Firstly, it contains a neologism-identification module, the Discoverer. The Discoverer scans online texts, either randomly or URL-driven, for grapheme combinations not recognised as (parts of) English words by the Google N-gram Corpus (cf. Evert 2010) and the English Wikipedia domain. The second important difference to other crawlers pertains to the data storage. The Observer, the second module, not only extracts the relevant passages from the Web in weekly crawling rounds, but also retains the entire web page in its original form in the database. Thus, future replications become possible and the sudden disappearance of web pages, complicating quantitative evaluation, is compensated. Thirdly, the Observer also contains an in-built linguistic classification scheme, which provides a linguistic profile for the neologisms as it captures the social, semantic, stylistic and pragmatic features and facilitates further research. In my talk, I will not only present the NeoCrawler and its innovative components, but also illustrate its potential for diffusion studies in linguistics by focusing on some recent results from my PhD research.

Katharina Franko (LMU)

":D Ciao ciao Italia hallo Deutschland! <3"

Eine Analyse deutsch-italienischer Facebook-Statusmitteilungen

Seit Bestehen des Internets und vor allem in Zeiten des Web 2.0 ist man sich einig, dass es sich bei der computervermittelten Kommunikation (CvK) um einen speziellen Sprachtyp handelt, der, obwohl medial grafisch, viele Merkmale mit dem konzeptionell Mündlichen teilt (vgl. Dorlejin, 2009; Storrer, 2001; etc.). Gerade das sogenannte Microblogging in Social Networks, wie z.B. Facebook oder Twitter, erfüllt Kommunikationsbedingungen und folgt Versprachlichungsstrategien, wie sie nur dort zu finden sind (vgl. Shafie, 2013; Kneidinger, 2010; etc.). Diese neuen Formen der Kommunikation haben zur Folge, dass

theoretische Ansätze neu überdacht, angepasst und gegebenenfalls neu formuliert werden müssen.

In dem Vortrag soll ein Teil der Ergebnisse aus einem Dissertationsprojekts vorgestellt werden, dessen Ziel es ist, Code-Switching bei bilingualen Benutzern auf Social Network Plattformen zu untersuchen unter Einbeziehung der Besonderheiten, die sich aus den verwendeten Sprachen Deutsch und Italienisch ergeben. Die Arbeit wird zum einen dazu beitragen, neue Erkenntnisse in dem Bereich der computervermittelten Kommunikation zu erlangen, und wird zum anderen eine neue Sichtweise auf Sprachwechsel ermöglichen. Forschungsgegenstand ist ein computergestütztes, annotiertes Korpus, das Facebook Statusmitteilungen und Kommentare von Benutzer enthält, die sowohl auf deutsch als auch auf italienisch schreiben.

Vor diesem Hintergrund sollen vor allem die folgenden zwei Punkte Inhalt des Vortrags sein: Zum einen soll auf die Datenextraktion von Facebook-Statusmitteilunge und der daraus resultierenden Möglichkeiten und Einschränkungen eingegangen werden und zum anderen sollen sprachliche Besonderheiten dieses speziellen Kommunikationsmediums vorgestellt werden. Es sollen dabei linguistische, sowie extra-linguistische Merkmale, die die Statusmitteilungen der deutsch- und italienisch-sprachigen sowie die der bilingualen Sprecher kennzeichnen, präsentiert und analysiert werden. So ist beispielsweise im Italienischen ein übermäßiger Gebrauch des Gerundiums ("bevendo caffè con mia sorella argentina... portando le usanze italiane in argentina :D"; Ital314) zu beobachten. Außerdem sollen nach einem soziolinguistischen Ansatz die Motive und Beweggründe, die zum Schreiben einer Statusmitteilung geführt haben, beleuchtet werden. Auch die Wahl der Sprache oder der Wechsel von einer Sprache in die andere innerhalb eines einzigen Beitrags soll hinterfragt werden.

Literatur:

Dorleijn, M.; Nortier, J. (2009), Code-switching and the internet. In B.E Bullock, A.J Toribio (Eds.). Linguistic Code-switching. Cambridge (UK): Cambridge University Press.

Kneidinger, B. (2010): Facebook und Co. Eine soziologische Analyse von Interaktionsformen in Online Social Networks. Wiesbaden: VS Verlag für Sozialwissenschaften.

Shafie, L. A.; Nayan, S. (2013): Languages, Code-Switching Practice and Primary Functions of Facebook among University Students. In Study in English Language Teaching 2013 (Vol. 1, No. 1), pp. 187–199, letzter Aufruf 21.08.2013.

Storrer, A. (2001). Sprachliche Besonderheiten getippter Gespräche; Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. Stuttgart: ibidem-Verlag.

Desislava Zhekova (LMU) & Thomas Krefeld (LMU)

Quantitative Typology and Dialect Dynamism: Analysis of Data from Contemporary Media

This project will apply quantitative typology to a novel type of dialectal data in order to exhibit and present the internal typological variation within one language. The research will be based on Italian and its dialects, because the most important typological isoglosses of Romance languages are crossing the Italo-Romance area with northern dialects which are close to the type that the French language is attributed to on the one hand and central and

southern dialects which are much closer to the Ibero-Romance or the Balkano-Romance type. The northern and southern zones will be exemplarily represented by Piedmontese and Sicilian.

The present project which is highly corpus-driven will focus on two main dimensions of linguistic variation. The first dimension is dialectal divergence as a result of diachronic processes; the linguistic distance is formulated in terms of morphosyntactic syntheticity/analyticity and rules of linearization or word order. The second dimension is synchronic convergence of the two dialects induced by their increased written use after the revolution of the New Media (web 2.0).

We plan to make use of novel datasets that have been formed by the recently introduced use of written dialects in new media, such as Wikipedia and Twitter. This contemporary type of data will be confronted with basilectal data extracted from transcriptions of oral recordings of the same dialect. With the latter we aim to show effects of Italianization which emerge ineluctably in writing because the writer is strongly influenced by the written standard variety: from a typological point of view the Italianization means functional convergence. In our study, we also aim to reevaluate the five metrics used in state-of-the-art literature to measure syntactic distance on this new type of data which is also marked by a significantly larger size. Based on the large amount of data, we will aim to show that there are statistically significant differences between dialects that currently other methods are not able to find. Additionally, the project will create a large scale morphosyntactically annotated dialect corpus, which does not yet exist, and will evaluate the applicability of collaborative content for syntactic quantitative typology.