Hamburg/Indiana Linguistics Workshop

12-13 December 2019, Universität Hamburg

Digital language research: 'computation' meets 'interaction'

Programme

Thursday, 12 December	
10.00—10.30	Greetings & Introduction
10.30—12.00	**Small stories research meets technography: The role of digital corpora and data mining in tracking** Alex Georgakopoulou & Anda Drasovean (King's College London)
	Integrating Discourse Perspectives and Computational Approaches Holly Lopez Long (Indiana University)
12.00—13.30	Lunch
13.00–15.00	**What 300-dimensional Fridges can Tell Us about Language** Dirk Hovy (Bocconi University)
	When I was in Paris last month… – Exploring a Corpus of Food Blogs Melanie Andresen & Heike Zinsmeister (Universität Hamburg)
15.00—15.30	Break
15.30—17.00	**A mixed methods approach to digital punctuation: Social indexicalities in frequencies and interactional practices** Florian Busch (Universität Hamburg)
	Computational Approaches to Computer-Mediated Discourse Analysis: From Text to Graphics Susan Herring (Indiana University) – *via video conference*
Friday, 13 December	
9.30	Arrivals & Coffee
09.45—12.00	**When the Minority Matters: Sentiment Analysis and Abusive Language Detection in Social Media with Imbalanced Data** Sandra Kuebler (Indiana University)
	Toward learning representations of social meaning Dong Nguyen (Utrecht University)
	Investigating device effects on language: a mixed-methods approach with different-sized corpora Jenia Yudytska & Jannis Androutsopoulos (Universität Hamburg)
12.00—13.00	Lunch
13.00—13.45	**Four common misconceptions of linguistic variation in computational social science** Cornelius Puschmann (University of Bremen)
13.45—14.30	Concluding Discussion
14.30	Coffee & Farewell

ABSTRACTS

Small stories research meets technography: The role of digital corpora and data mining in tracking

Alex Georgakopoulou & Anda Drasovean (King's College London)

Small stories research, a social interactional paradigm for the analysis of everyday life stories and identities in both face-to-face conversations and digital environments, has recently been extended in Georgakopoulou's work (www.ego-media.org), serving, amongst others, as a paradigm for critically interrogating the current 'engineering' (designing) of stories on social media platforms. Since its inception, the main methods of data collection and analysis in small stories research have been located in ethnographic, contextualized, micro-analytical modes. It is often the case that any 'quantitative' dimension in inherently 'qualitative' approaches tends to be employed as a way of 'scaling up'. In this talk, however, based on the exigencies of communication on social media environments and on what constitutes 'data' in them, we want to claim a synergy of 'interactional' with 'computational' methods, on the basis of its contribution to what we call a technographic approach to stories: an approach that uses genealogical perspectives on platforms, tracking of evolving affordances, and corpus-assisted critical discourse analyses. We will show how combining, broadly speaking, qualitative and quantitative perspectives has involved an iterative process of back & forth and of 'bracketing'. We also want to talk about the challenges we have faced in essentially amalgamating approaches with different priorities: from technical limitations of search engines and issues of 'translating' analytical queries across modalities to more mundane yet pivotal issues for the analysis, such as figuring out 'local posting times'.

Integrating Discourse Perspectives and Computational Approaches

Holly Lopez Long, Indiana University

The growing availability of textual data from social media and technology-mediated environments has engendered interest in methods for processing and analyzing big data. Some scholars have been turning to discourse phenomena to find solutions for more challenging problems such as, inferring the relationships between sentences (Pan et al. 2019) or detecting politeness (Danescu-Niculescu-Mizil et al. 2013). Thus, the importance of discourse analytic perspectives and the examination of discourse phenomena persists, especially as scholars endeavor to help machines understand and mimic the complex maneuvering of human interactions. The purpose of this session is to provide a path for integrating a discourse analytic perspective while using computational approaches. To do this, I will discuss the evolution of my research examining politeness strategies on mobile dating interactions – from its start as a discourse-focused project and its trajectory, as I transition to more computational approaches. This work will serve as a jumping-off point for discussing the development of my research questions and their relationship to the approaches employed, as well as the skills that a discourse analyst must acquire to incorporate computational approaches to their work. It will also serve as an illustration for the limitations of both discourse analysis and computational approaches.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. arXiv preprint arXiv:1306.6078

Pan, B., Yang, Y., Zhao, Z., Zhuang, Y., Cai, D., & He, X. (2019). Discourse marker augmented network with reinforcement learning for natural language inference. arXiv preprint arXiv:1907.09692.

What 300-dimensional Fridges can Tell Us about Language

Dirk Hovy

Recent advances in machine learning have produced a new way of representing words in a multi-dimensional vector space. Each word is represented as a point in this space, and its position is determined by the contextual similarity to all other words. This is not unlike arranging word magnets on a fridge. The individual dimensions do not correspond to any meaningful interpretation (e.g., dimension 5 is for vowels), but the overall position has to be interpreted holistically.

These vector representations, or embeddings, have turned out to capture a variety of latent factors, from lexical semantics to syntax to socio-demographic aspects to societal attitudes.

In this talk, I will give a brief introduction of the method (a direct implementation of the distributional hypothesis by Firth), and then show several applications these embeddings enable: they capture regional variation at a intra- and interlingual level, they distinguish varieties and linguistic resources, and they allow for the assessment of changing societal norms and associations. The ease of use and the range of applications make embeddings a valuable tool for further research in (computational) sociolinguistics.

***When I was in Paris last month…* – Exploring a Corpus of Food Blogs**

Melanie Andresen & Heike Zinsmeister (Universität Hamburg)

The genre of food blogs is currently very popular and linguistically interesting as a merge of traditional recipe books and new ways of communicating in the blogosphere. The communicative functions of food blogs clearly go beyond the presentation of recipes, e. g. serving as “tools for identity building” (Lofgren 2013). Our object of study is a corpus of 150 posts from 15 German food blogs. We approach this corpus in an explorative way, looking for noteworthy features of the corpus that can be identified in a data-driven way partly by comparing the food blog data with standard newspaper data.

For this purpose, we present a number of common quantitative methods for corpus exploration like keywords, surface-based and syntactic collocations (Andresen & Zinsmeister 2017), principal components analysis, and topic modeling. The aim of the presentation is to explore – together with the audience – what we can learn about food blogs this way and how these computational methods can be profitably combined with interactional approaches.

Andresen, Melanie and Heike Zinsmeister. 2017. The Benefit of Syntactic vs. Linear N-grams for Linguistic Description. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), S. 4-14. Pisa, Italy, September 18-20 2017.

Lofgren, Jennifer. 2013. Food Blogging and Food-related Media Convergence. M/C Journal 16(3). http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/638.

A mixed methods approach to digital punctuation: Social indexicalities in frequencies and interactional practices

Florian Busch, Universität Hamburg

Drawing on data and methods from my recently completed doctoral dissertation on registers of digital writing among German adolescents, I discuss how digital punctuation as a syntactic, interactional and social resource can be investigated by integrating quantitative and qualitative methods. In particular, three integrated levels of analysis are reviewed: (1) By using a Python script for extracting punctuation frequencies, quantitative usage patterns in a sample of 47 WhatsApp chatlogs (151,970 word tokens) are investigated. (2) Sequential analysis of chat threads provides further interactional interpretations of these frequencies and shows how usage patterns accumulate from recurring

functions of punctuation. (3) Ethnographic data from seven semi-structured interviews offer insight on how digital writers metapragmatically reflect on various styles of digital punctuation and the social values and situated identities that they are enregistered with. The paper argues that these three methodological layers (re-)construct different indexical orders of punctuation. The integration of computational, interactional and metapragmatic methods reveals characteristics of digital punctuation as a 'communicative practice' in a holistic way.

Computational Approaches to Computer-Mediated Discourse Analysis: From Text to Graphics

Susan C. Herring herring@indiana.edu

The internet is a vast repository of pre-digitized communication and social engagement (Gernsbacher, 2014). For discourse linguists, automated analysis of computer-mediated corpora has the potential to 1) address classic questions on a larger scale and 2) identify new insights using techniques that did not previously exist (Jones & Dye, 2017). However, the second potential is more realized than the first: Significant gaps exist between the traditional interests of discourse analysts and the kinds of analyses that automated tools and methods such as Natural Language Processing (NLP) facilitate. Further, NLP is focused on text mining, while online discourse increasingly incorporates graphical elements (*graphicons*) such as emoji, stickers, GIFs, and images. Little systematic attention has yet been paid to automating analysis of graphical computer-mediated discourse (CMD). In this talk, I discuss these lacunae and describe attempts to address them with reference to the Computer-Mediated Discourse Analysis (CMDA) research paradigm (Herring, 2004).

CMDA differs from other discourse-focused approaches in that its data are online interactions, and, as a "coding and counting" approach (Herring, 2004), it lends itself well to quantification. Moreover, several methods in the CMDA "toolkit" can be automated, especially at the levels of Participation and Structure. Traditionally, pragmatics and social phenomena have been identified as most challenging to NLP because of their dependence on larger textual units and surrounding context, but as I will illustrate, creative work is being done using machine learning and other computational methods to analyze pragmatic and social phenomena in CMD. Where automation is most lacking is for analysis of interaction management (turn-taking, topic development, repair, etc.); reasons for this and possible solutions will be suggested.

With regard to computational approaches to graphical CMD, some isolated automated analyses have been done. For example, machine learning and big data techniques have been employed to 1) use emoji and emoticons to improve sentiment analysis of text, 2) use surrounding text to analyze the intended meaning of graphicons, 3) use computer vision to determine the contents of graphicons, and 4) predict and suggest appropriate graphicons in the context of a CMD conversation (Dainas, 2019). These promising studies touch on core discourse issues (structure, meaning, and interaction), while leaving other gaps. In concluding, I describe efforts on the part of my research team to systematize a CMDA-based methodological paradigm for graphical CMDA and suggest ways that this paradigm might incorporate automated methods into its "toolkit."

Dainas, A. R. (2019, April). Communicative uses of graphicons in graphical computer-mediated communication. Ph.D. Qualifying Paper, Department of Information and Library Science, Indiana University, Bloomington.

Gernsbacher, M. A. (2014). Internet-based communication. *Discourse Processes, 51*, 359-373.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338-376). New York: Cambridge University Press.

Jones, M. N., & Dye, M. W. (2017). Big data approaches to studying discourse processes. In M. F. Schober, D. N. Rapp, M. A. Britt (Eds.), *The Routledge handbook of discourse processes, 2nd edition* (pp. 117-124). London: Routledge.

When the Minority Matters: Sentiment Analysis and Abusive Language Detection in Social Media with Imbalanced Data

Sandra Kuebler, Indiana University

Data sets for sentiment analysis are typically balanced between positive and negative opinions, which is often a simplification of the situation. In contrast, I am interested in working with highly imbalanced data sets and investigating methods to improve classification results for cases when we are interested in the minority cases, with a focus on feature selection and sampling methods. I will present research on cooking reviews as well as on abusive language detection. We will also venture into multilingual abusive language detection.

Toward learning representations of social meaning

Dong Nguyen (Utrecht University)

The way language is represented in Natural Language Processing has changed radically with the emergence of neural network approaches that learn to represent words, sentences, and other linguistic units as dense real-valued vectors. Word vector representations have not only become a standard component in modern NLP systems, they are also increasingly used as first-class research objects to explore social and linguistic questions. The large body of work has focused almost exclusively on the semantic and syntactic aspects of words and how these are encoded in word representations. However, word representations are automatically learned from large corpora based on the contexts in which words occur and therefore have the potential to encode various aspects—not only syntactic and semantic ones but also social aspects.

In this talk, I will present an analysis of word representations learned from Twitter and Reddit data. The analysis focuses on representations of spelling variants, as different spelling variants can carry different social meanings. I will then share some thoughts on how approaches for learning word representations could be improved by rethinking what we mean by context.

Investigating device effects on language: a mixed-methods approach with different-sized corpora

Jenia Yudytska & Jannis Androutsopoulos, Universität Hamburg

The notion that the choice of communication device may impact on the form of digital language messages is part of contemporary language-and-media ideologies, as evidenced when people formulate assumptions about production device to explain e.g. the length of a message, the use (or omission) of punctuation, and other aspects of usage. In linguistic research on computer-mediated communication, by contrast, language produced via personal computer and mobile devices has generally been treated as broadly similar. Empirical research findings on how the choice of communication device affects language production are as yet inconclusive and often incidental to other research aims.

We present the design of a new project that centres a potential ‘communication device effect’ (CDE) on language. The research will ask whether different digital communication hardware (personal computer vs. mobile device) correlates with micro-linguistic features (e.g. punctuation, emoticons) and/or discourse features (e.g. message length, topic choice),and what role contextual and social factors may play in this context. To explore these questions empirically, a novel corpus using data from the platforms Discord (a popular new chat server platform) and Twitter is being compiled. We reflect on the design of a corpus that enables both quantitative and qualitative analysis. While robust quantitative analysis calls for a large-scale corpus, a qualitative, fine-grained approach is typically

only possible with a smaller, more manageable corpus. We consider the possibility of constructing two complementary corpora: a smaller, traceable corpus of messages collected from a select set of participants across two platforms (Discord and Twitter), and a large-scale, untraceable corpus collected from Twitter. We discuss how such corpora can complement each other's weaknesses and offer an in-depth perspective on the research question.

Four common misconceptions of linguistic variation in computational social science
Cornelius Puschmann, Bremen

The analysis of language data plays a pivotal role in the nascent field of computational social science. Methods such as topic modeling and sentiment analysis are widely used by researchers not necessarily trained in linguistics and generally interested in issues not fundamentally related to language per se, but to the social processes it is presumed to index. While a large body of linguistic knowledge exists that has limited relevance to social science applications of computational content analysis, certain assumptions about the patterned and systematic nature of linguistic variation have substantial impact on the scope and validity of CSS analyses.
My talk will cover four common misconceptions concerning language and linguistic variation common in (computational) social science:

1. a normative (or dismissive) approach to non-standard language (i.e. the assumption that certain units of language are per se meaningful, while others are 'meaningless'),
2. an understanding of linguistic meaning that is restricted to 'obvious' denotational dimensions (for example, to the topic of a newspaper article or the emotions expressed in a social media posting), largely omitting linguistic variation that is socially or stylistically conditioned,
3. an aversion to working with syntactic data (or anything on the level of phrase, clause or sentence beyond the bag-of-words approach),
4. a graphemic understanding of what constitutes a word (particularly dangerous when stemming or lemmatization are used).

Having pointed out how these misconceptions influence CSS analyses, examples of ways in which the field is progressing to strategies that better incorporate linguistic knowledge will be discussed.

PARTICIPANTS

Melanie Andresen
Research associate, Corpus Linguistics
Universität Hamburg
https://www.slm.uni-hamburg.de/germanistik/personen/andresen.html

Jannis Androutsopoulos
Professor of German Linguistics and Media Linguistics
Universität Hamburg
https://www.slm.uni-hamburg.de/germanistik/personen/androutsopoulos.html

Florian Busch
Research associate, German Linguistics and Media Linguistics
Universität Hamburg
https://www.slm.uni-hamburg.de/imk/personen/busch.html

Anda Drasovean
Research associate

King's College London

Alexandra Georgakopoulou
Professor of Discourse Analysis and Sociolinguistics
King's College London
https://www.kcl.ac.uk/people/alexandra-georgakopoulou

Susan C. Herring
Professor of Information Science, Adjunct Professor of Linguistics,
Director of the Center for Computer-Mediated Communication, Indiana University
herring@indiana.edu
https://info.sice.indiana.edu/~herring/
https://ccmc.ils.indiana.edu
https://www.languageatinternet.org

Dirk Hovy
Associate Professor of Computer Science
Bocconi University
www.dirkhovy.com
@dirk_hovy

Sandra Kuebler
Professor of Computational Linguistics, adjunct appointments in Cognitive Science, Computer Science, Germanic Studies
Indiana University
cl.indiana.edu/~skuebler

Holly Lopez Long
Ph.D. student in Information Science
Indiana University
holly.lopez.long@gmail.com

Dong Nguyen
Assistant professor, Department of Information and Computing Sciences, Utrecht University
Research Fellow, Alan Turing Institute, London
https://dongnguyen.nl/

Cornelius Puschmann
Professor of Media and Communication, ZeMKI, University of Bremen
Affiliate researcher, Leibniz Institute for Media Research
cbpuschmann.net
https://twitter.com/cbpuschmann

Jenia Yudytska
Doctoral Candidate in General Linguistics
Universität Hamburg
j.yudytska@gmail.com

Heike Zinsmeister
Professor of German Linguistics and Corpus Linguistics
Universität Hamburg
https://www.slm.uni-hamburg.de/germanistik/personen/zinsmeister.html