Beatrix Busse, Nina Dumrukcic, Ingo Kleiber

Introduction

Abstract: In the wake of the COVID-19 pandemic, ICAME41, somewhat prophetically titled *Language and Linguistics in a Complex World*, was shifted from a physical to a virtual conference. In light of a rapidly changing world, ICAME41 aimed at challenging the future of (corpus) linguistics, its approaches, questions of transfer, and the intersection between various fields and areas of expertise. By exploring new formats of presenting, sharing, and discussing research, the conference also provided a glimpse into one of many possible futures for the field and academia as a whole. While this introduction is devoted to these questions, the articles in this volume focus on the complexity and diversity of language and on analyzing it with increasingly sophisticated methods and ever-larger datasets.

1 Questions and Concepts

The articles in this volume *Language and Linguistics in a Complex World* evolved from presentations at the 41st Conference of the *International Computer Archive of Modern and Medieval English* (ICAME41). This conference was initially planned to be held on-site at Heidelberg University in May 2020. However, it was then one of the first linguistics conferences which were realized online because of the COVID-19 pandemic. But let's not get ahead of ourselves and first outline the theme of both the conference and this volume. It will illustrate why – at least to some extent – the topic of the conference was chosen somewhat prophetically, especially given the deep crisis and massive changes affecting all of our human existence due to the COVID-19 pandemic, and that we write this introductory chapter with the experience of the last two years of being in a pandemic and now even with an atrocious war by the Russian aggressor against Ukraine. Our world has massively changed, and it is not easy to focus on the topics of the conference alone. However, we are fully convinced of the fact that the work we do, education and generating new research, have never been more important for this and the next generation, our society, and the planet – hence, for a peaceful, democratic, and sustainable world.

Next to discussing cutting edge research in, for example, the field of English (historical) corpus linguistics, the conference *Language and Linguistics in a Complex World: Data, Interdisciplinarity, Transfer, and the Next Generation* aimed to

Open Access. © 2023 the author(s), published by De Gruyter. CC BY This work is licensed under the Creative Commons Attribution 4.0 International License.
https://doi.org/10.1515/9783111017433-001

take (corpus) linguistics out of its comfort zone and discuss its (inter-)disciplinary and transfer potential in detail.

It aimed to determine the intersections between (corpus) linguistics and other academic fields such as sociology and psychology as well as marketing, law, politics, education, and art.

As a result of (hyper)globalization, digitization, gaining access to more and more information, and technological developments, we are faced with growing global and local complexity and interdependence of matter, lives, people, and things, which also includes language. Language continues to be the link between cultures, fields of study, and people. Furthermore, the means of analyzing, producing, and comprehending language is rapidly evolving through machine learning, artificial intelligence, and big data research, yet it remains at the core of the humanities.

We have not even begun to understand how all of these issues and concepts will interact (humanely, sensibly, peacefully, and sustainably) under the new circumstances of rapid change – nor have we yet considered what role linguistics and corpus linguistics may have to play in the solution of global challenges, a new world order, and how education and teaching will consequently have to be transformed. People communicate with one another and use language every day, yet a large section of the population is not familiar with the types of questions that are addressed in (corpus) linguistics. For example, how do scientists in this field look for patterns, and what can that tell us about human behaviour? Moreover, our aim was to scrutinize how contemporary, as well as upcoming methods and techniques which we have developed, might impact these questions in the future.

As corpus linguistics developed as a sub-branch of linguistics, a wealth of qualitative and quantitative research has been accumulated over the years, and highly innovative and ground-breaking tools have become accessible to linguists. We find that it is necessary to also share this knowledge with the public to be used for other purposes and to bridge the gap between academia and industry more than ever before. It is also of crucial importance to acknowledge software development and data itself as research outputs in their own right. This also extends to the way we present and publish findings, reviews, and analyses. While peer-reviewed journals, edited volumes, and monographs continue to add credibility and maintain a level of quality assurance, many scholars are also extending their outreach to include preprints, blogs, podcasts, video tutorials, code and data repositories such as GitHub, using forums to discuss important issues, and utilizing social media.

This has also been acknowledged and pointed out in the past. For example, in 2015, Mónica I. Feliú-Mójer wrote a blog post on the importance of effective communication, and how the more clearcut and comprehensible the message is, “science thrives.” Wu (2017) questions the use of complex terminology and jargon as well as compromising the ability to be an effective communicator when thinking of the general public as “other.” Making linguistic research accessible, Wagner et al. (2015) describe establishing a *Language Sciences Research Lab* within a science museum, that combines formal instruction with outreach and integration with the general public.

By transferring the methods and insights of (corpus) linguistics to society, we are not only increasing our impact as researchers, but also gaining further knowledge and input directly from stakeholders about their needs, which, given the current geopolitical circumstances, will become even more important. Furthermore, it has become essential to examine the issue of increasingly complex data and whether researchers have to acquire novel skills in areas outside of their current expertise. This also opens up the question of whether this needs to happen on an individual level or if there should be a more comprehensive collaboration between experts from various disciplines. Corpus linguistics is at a crossroads, and the time has come to evaluate and consider what the field will look like in the forthcoming years and how it will be shaped by young and emerging scholars as well as people from outside the traditional academic sphere.

The transition from workshops and conferences in physical presence to digital and hybrid spaces also means that participants from all over the world, who might otherwise be prevented from attending due to, for example, travel expenses and time constraints, are able to contribute to, and learn from, participating in discussions with their peers. The virtual conference format has resulted in a number of sociodemographic changes such as greater attendance by women, members of historically under-represented institutions, as well as graduate students and postdoctoral associates (Skiles et al. 2021). Equity and inclusivity on this scale are unprecedented, and although establishing the schedules sometimes is challenging due to various time zones, nonetheless, academics are able to partake in networking with fellow researchers wherever they happen to be in the world. Moreover, Skiles et al. (2021) show in their study, which compares remote and in-person conferences, that the former has positive environmental factors because the participants’ travel-related carbon footprint is greatly decreased. All these observations raise the question of what conferences will and should look like in the future.

2 Digital Conferences

The decision to host ICAME41 in digital space instead of physically at Heidelberg University due to the COVID-19 pandemic was not made lightly. This shift challenged all of the previously well-established and familiar methods of organizing and executing academic conferences. The lockdown that ensued was an unprecedented way of working for most, where we embraced the technology that we had at our disposal to make the best of the situation. As an important international corpus linguistics conference that has been taking place since 1979, the *International Computer Archive of Modern and Medieval* English (ICAME) conferences have been inextricably linked to a multitude of traditions, social events, and culture – many of which are closely tied to the spaces ICAME has been happening in and at. Hence, one of the biggest challenges was to recreate the sense of community and establish a platform and fitting formats for discussing and sharing innovative ideas in the digital space.

Aside from the effort that usually goes into organizing large events, there was a myriad of questions about how high-quality content and social interactions can be brought into the digital space. The original blueprint for the conference and ideas that the organizing committee deliberated had to be radically modified. In early 2020, society was engulfed in fear and doubt as the COVID-19 pandemic swept across the globe. Yet, this was also an opportunity for creative thinking and opened the door to questioning the previously well-established way of organizing and attending conferences. While virtual conferences and events existed prior to the COVID-19 pandemic, the number of conferences offered in virtual and hybrid formats has since skyrocketed. Our team took this challenge as an opportunity to brainstorm and think about not just how to recreate the familiar experience in the digital space, but what new and exciting prospects this could bring. Therefore, next to the established formats, ICAME41 featured, for example, a design thinking workshop with industry experts as well as a publicly streamed plenary discussion. Regarding the core academic program, the organizing committee combined synchronous and asynchronous contributions to balance the excitement of attending live plenaries while minimizing technical difficulties and stress by asking participants to upload pre-recorded talks and poster presentations. The keynote speakers and participants embraced these novelties and co-creatively, together with the organizers and other participants, created meaningful and interesting content. The process of moving ICAME41 into the digital space is discussed in further detail in Busse/Kleiber (2020), and the purpose of the paper is to share our experiences and best practices that serve as guidelines for future event organizers.

One of the pillars of contemporary academic research ought to be sharing knowledge and ideas with others. Closely related to this, we are seeing that more and more linguists are embracing open science and open education. We see an unprecedented amount of code, data, and (learning) resources being developed and made available publicly, collaboratively, and openly. By providing people both within and outside the academic community with the opportunity to be able to learn about research currently conducted at higher education institutions, not only is the information available to more people, but they are able to make their own contributions and replicate and verify the research being conducted by others. This process ensures that there is less discrimination towards people and institutions who may not have the resources to conduct similar studies but nonetheless have the intellect and creative thinking that we as a society would very much all benefit from.

3 Articles in this Volume

Upon successful completion of the conference, scholars who were interested in publishing their papers in these proceedings were invited to submit their work. This is a more in-depth look at some of the papers which were outlined in a more concise manner in the published Extended Book of Abstracts (Busse/Dumrukcic/Möhlig-Falke 2021). The papers underwent a double peer-review process by experienced and qualified experts in the field who kindly provided feedback to the contributors. The general trend we noticed over the course of the conference, and by reading the contributions was that there is a wide and inclusive perspective on language(s). Moreover, there is a continuation of the tendency to use increasingly sophisticated quantitative and qualitative methods. The adoption of more complex and sophisticated technology and methodology is also enhancing research as corpus linguists are finding new as well as faster, and more efficient ways of looking for language patterns in ever-increasing amounts of linguistic data.

In his paper on World Englishes, Axel Bohmann uses the Contrastive Usage Profiling (CUP) method in order to quantify relations among different varieties of English based on lexical co-occurrence. This method relies on word embeddings to represent word usage using online discourse data from the Corpus of Global Web-based English (GloWbE, Davies 2013). The author considers the profiles of individual words (i.e., ‘English’, ‘holy’, ‘chop’, ‘yard’, ‘football’, ‘boot’) in 20 varieties of English such as New Zealand, Irish, the United States, Canadian etc., and introduces a word embedding model that is constructed for each national subcorpus of GloWbe (Davies 2013). This procedure uncovers relationships among varieties, both in regard to individual words and in an aggregate view. The results show differentiation between countries in phase five according to Schneider (2007) and formerly colonized countries that are still in the process of postcolonial linguistic emancipation. Furthermore, most other varieties differ from British English and American English rather than being more drawn to either of them.

Axel Bohmann, Julia Müller, Mirka Honkanen, and Miriam Neuhausen present the findings of a large-scale, multivariate study of how passive alternation developed in 19th- and 20th-century American English. There has been an increase in the use of GET to form passive sentences in American English, and a decrease in frequency of the BE-passive construction. A Python script was written to extract all instances of lemma BE and GET + past participle from the Corpus of Historical American English (COHA, Davies 2010), totalling 2,318,251 tokens. Intervening adverbs and negators were also included. Diachronic change, informality, subject responsibility, adversativity, and non-neutrality were assessed in relation to the GET-passive along with a range of syntactic predictors. One of the strongest predictors in the logistic mixed-effects model was the publication year of the text. There was a general rise in GET both in absolute numbers and as a competitor to BE throughout the observed time period (1830–2000), confirming the informality hypothesis that it is more likely to be used in informal contexts. Other constraints such as subject responsibility have weakened over time. Findings for adversativity/non-neutrality were less conclusive, but there was no strong evidence for the significance of these suggested semantic characteristics of the GET-passive. The semantic group of the passivized verb shows a particularly strong effect size. The article concludes that there is strong lexical-semantic conditioning of the passive alternation.

Gavin Brookes examines discourses around social class in British press coverage of obesity and how language has the power to shape societal perspectives on health and illness. The author uses a broadly social constructionist view of discourse, and a corpus-based approach is used to conduct a critical discourse analysis (CDA). The data is taken from a 36-million-word corpus of obesity-related newspaper articles published between 2008 and 2017 (Brookes/Baker 2021). Normalized frequency analysis of the phrase *social class* as a sub-sample of the newspapers mentioning obesity showed that left-leaning broadsheets have a tendency to frame obesity and poor diet as consequences of social class with social inequalities construed as the cause not only of obesity but also of health inequalities more widely. On the other hand, the right-leaning newspapers, including both tabloids and broadsheets, offered discourses that mitigated the influence of social class on obesity, claiming that obesity affects people at all class levels and that lifestyle choices are more influential in the development of obesity.

Steven Coats examines corpora compiled from YouTube automatic speech recognition (ASR) transcripts from channels in the United States, Canada, and the British Isles to study regional language variation in spoken English. The method of data collection relies on web scraping and open-source software for the automatic identification and downloading of suitable channel content as well as dealing with the rate-limiting issues that arise thereby. Word frequency statistics are used to assess the accuracy of the downloaded transcripts. The ASR transcripts (approximately 500,000 words) are compared to manual transcripts of city council meetings in Philadelphia to determine word error rates. Moreover, word embeddings are used to create a language model from a subset of the corpus. A transcript classification task is undertaken using vector-based distributed representations of transcript content. Furthermore, the article concludes that although there is a certain degree of error, utilizing ASR transcripts in corpus linguistic research is useful for the study of regional language variation.

The following article is by María-Isabel González-Cruz, who explores the pragmatic roles and effects that Anglicisms seem to play in a corpus of headings taken from the Spanish regional digital newspaper *Canarias 7*. The corpus includes a total of 1,618 headings with Anglicisms collected between 2019 and 2020. Using a qualitative approach, the author differentiates between three categories of Anglicisms: 1) new Anglicisms – those which have not been registered yet in the *Diccionario de la Lengua Española* (DLE), the official dictionary published online by the Royal Academy of the Spanish Language; 2) registered Anglicisms and 3) proper nouns. The proper nouns are further divided into categories such as titles, names, toponyms, and acronyms. The author concludes that Anglicisms tend to be used for their brevity and precision, to indicate certain attitudes, such as giving a humorous touch (through word-play or by resorting to familiar phrases), to provide connotations of modernity as well as perform a euphemistic role.

Yoko Iyeiri and Mariko Fukunaga compiled the ABCFM Hawaii Corpus by assembling selected writing from the Hawaiian Mission Children's Society Library (HMCS Library) in Honolulu which holds a large collection of 19th-century journals, letters, and an autobiography written by members of the American Board of Commissioners for Foreign Missions (ABCFM) (cf. Forbes et al. 2018). The Hawaii Corpus, which encompasses approximately 653,100 words, represents the state of 19th-century American English, while at the same time providing material suitable for historical sociolinguistic analyses, showing the variability of English among different authors. The eight authors in the corpus were well-educated, and all belonged to the same community with shared missionary aims. Therefore, any individual deviations from the norm tend to be rather subtle. The style of one person showed a relatively informal trend when compared to other members. Although other authors also employed some features of negation, this particular person's deviation was always marked and consistent. The paper explores some variable aspects of negation in the data, with a focus on the use of the auxiliary do in negation. After considering the frequency of negation, findings show that while negative constructions are relatively stable in the 19th century, the use of 'do' in negation was not yet consistent.

Gerold Schneider uses context-aware language models to compare the reading performance of L1 to L2 language users. The main research questions addressed which features correlate to and predict reading time, variation between L1 and L2 readers, whether reading time can be predicted in L2 as well as for L1 readers, and if longer reading time shows which constructions are particularly difficult for L2 readers. Data from the Ghent Eye tracking Corpus (GECO, Cop et al. 2017) was used and restricted to only L1 English readers whose dataset was complete, and L2 readers who had less than 50% daily exposure to English. Key points of analysis include surprisal, recency in the discourse, word length, and punctuation to predict reading times in psycholinguistic experiments obtained by measuring eye tracking since research shows that frequency and expectation can affect what is easier to process (e.g., Conklin/Pellicer-Sánchez/Carrol 2018). The study showed strong correlations between reading times and surprisal, although considerably less for L2 readers.

This collection of papers demonstrates how research can thrive even in times of great unpredictability and concern. As Mahlberg/Brookes (2021: 442) mention in their recently published article on corpus linguistics and the COVID-19 pandemic, this is a "testament to the applied nature of corpus linguistics, as well as to the innovativeness of our research community to respond rapidly and creatively to the most urgent global challenges of our time." While ICAME41 was in some ways a deviation from the traditional conference experience, it nonetheless provided insight into new ways of carrying out, presenting, and sharing research with the ICAME community and beyond.

References

Brookes, Gavin/Baker, Paul (2021): *Obesity in the News: Language and Representation in the Press*. Cambridge: Cambridge University Press.

Busse, Beatrix/Kleiber, Ingo (2020): "Realizing an online conference: Organization, management, tools, communication, and co-creation." In: *International Journal of Corpus Linguistics* 25, 322–346.

Busse, Beatrix/Dumrukcic, Nina/Möhlig-Falke, Ruth (Eds.) (2021) *Language and Linguistics in a Complex World Data, Interdisciplinarity, Transfer, and the Next Generation. ICAME41 Extended*

Book of Abstracts. The International Computer Archive of Modern and Medieval English Annual Conference, May 20–23, 2020, Heidelberg University and University of Cologne, Germany.

Conklin, Kathy/Pellicer-Sánchez, Ana/Carrol, Gareth (2018): *Eye–Tracking. A Guide for Applied Linguistics Research*. Cambridge: Cambridge University Press.

Cop, Uschi/Dirix, Nicolas/Drieghe, Denis/Duyck, Wouter (2017): "Presenting GECO: An eye tracking corpus of monolingual and bilingual sentence reading." In: *Behavior Research Methods* 49, 602–615.

Davies, Mark (2010-): *The Corpus of Historical American English (COHA). 400 million words, 1810–2009*. Online at: http://corpus.byu.edu/coha/ <14 April, 2022>.

Davies, Mark (2013): *GloWbE. Global Web-based English*. Online at: https://www.english-corpora.org/glowbe/ <14 April, 2022>.

Feliú-Mójer, Mónica I. (2015): *Effective Communication, Better Science*. Online at: https://blogs.scientificamerican.com/guest-blog/effective-communication-better-science/ <14 April, 2022>.

Forbes, David W./Kam, Ralph Thomas/Woods, Thomas A. (2018): *A Biographical Encyclopedia of American Protestant Missionaries in Hawaii and their Hawaiian and Tahitian Colleagues, 1820–1900*. Honolulu: Hawaiian Mission Children's Society.

Mahlberg, Michaela/Brookes, Gavin (2021): "Language and Covid-19: Corpus linguistics and the social reality of the pandemic." In: *International Journal of Corpus Linguistics* 26, 441–443.

Schneider, Edgar W. (2007): *Postcolonial English. Varieties Around the World*. Cambridge: Cambridge University Press.

Skiles, Matthew/Yang, Euijin/Reshef, Orad/Muñoz1, Diego Robalino/Cintron, Diana/ Lind, Mary Laura/Rush, Alexander/Calleja, Patricia Perez/Nerenberg, Robert/Armani, Andrea/ Faust, Kasey M./Kumar, Manish (2021): Conference demographics and footprint changed by virtual platforms. *Nature Sustaianability*.

Wagner, Laura/Speer, Shari R./Moore, Leslie C./ McCullough, Elizabeth A./Ito, Kiwako/ Clopper, Cynthia G./Campbell-Kibler, Kathryn (2015): "Linguistics in a science museum: Integrating research, teaching, and outreach at the language sciences research lab." In: *Language & Linguistics Compass* 9, 420–431.

Wu, Katherine (2017): *Why can't scientists talk like regular humans*. Scientific American. Online at: https://blogs.scientificamerican.com/observations/why-cant-scientists-talk-like-regular-humans/ <14 April, 2022>.