First circular

ICAME47 - A CONFLUENCE OF CORPUS RESEARCH IN THE AGE OF AI

WELCOME!

Dear ICAME participants,

We are pleased to announce that **ICAME47** will take place in **Koblenz, Germany, from 26–30 May 2026**. ICAME (International Computer Archive of Modern and Medieval English) is an annual international conference and one of the longest-standing organisations bringing together linguists and data scientists working with English language corpora. The conference theme is **"A Confluence of Corpus Research in the Age of AI."**

We received a large number of submissions from over 25 countries across all continents, and we are delighted to announce that the programme will feature **four plenary talks, five pre-conference workshops**, as well as a wide range of **full papers, work-in-progress reports, and software demonstrations**.

As a small and very young university, we are immensely proud to host one of the longest-running conferences in corpus linguistics. In this first circular, you will find further information about the conference venue, the conference warming, the boat trip, the conference dinner and disco, expected fees, travel and accommodation, Koblenz and its surroundings, plenary speakers, and the pre-conference workshops.

We very much look forward to welcoming you to Koblenz in May. In the spirit of the conference theme - **let's flow together!**

With all very best,
The ICAME47 Organising Committee

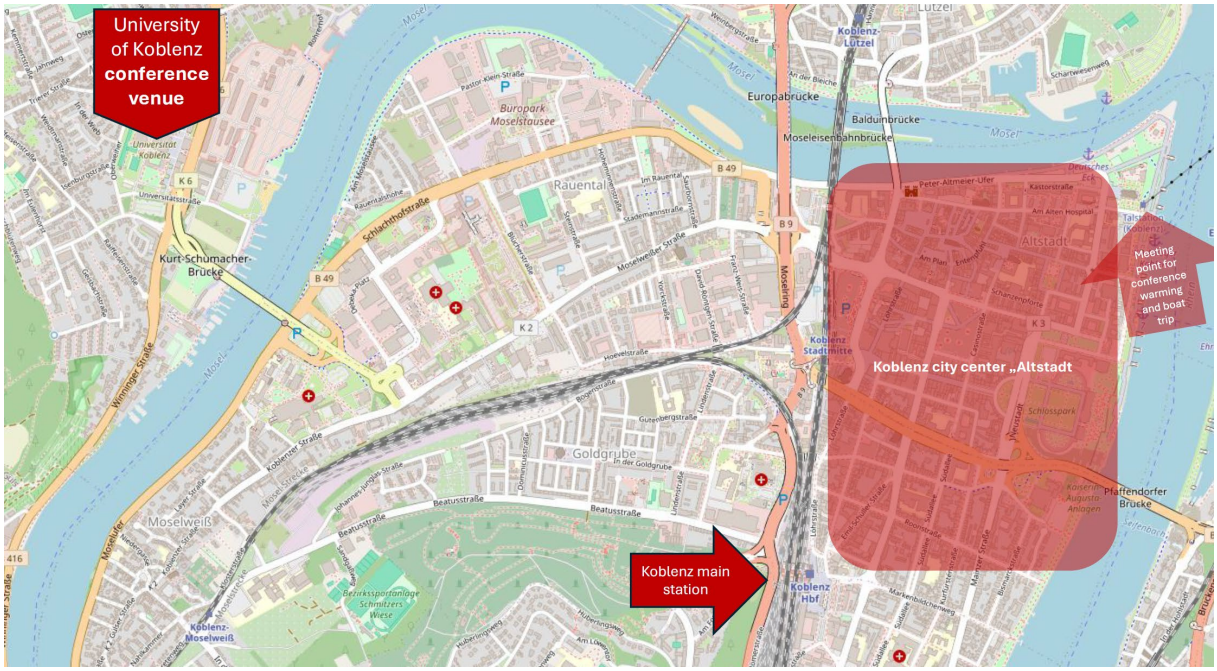
CONFERENCE VENUE

ICAME47 is going to take place on the Metternich campus of the **University of Koblenz** – Germany's youngest university, located just steps away from the scenic banks of the Moselle River.

The exact address is

Universitätsstraße 1
56070 Koblenz
Germany

The campus is located in **Metternich**, on the north-western side of the Moselle River. It is well connected to the city centre by regular bus services. Please note, however, that two major rivers - the Rhine and the Moselle - run through the city. Depending on which side of the river you are staying on, this may affect your daily commute. Here is a map for an overview:



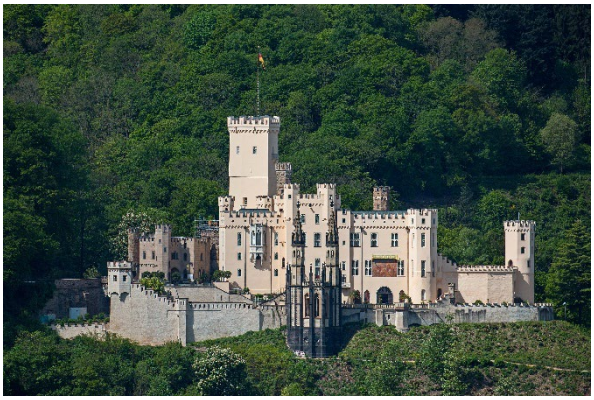
CONFERENCE WARMING - WEDNESDAY (27 May)

The conference warming will take place at the top of **Festung Ehrenbreitstein**, one of the most iconic landmarks of the region. With a history spanning more than a thousand years, the fortress offers a unique historical setting for the opening of the conference. Situated high above the city, it provides a spectacular panoramic view over Koblenz and the confluence of the Rhine and Moselle rivers. The combination of impressive architecture, rich history, and breathtaking scenery makes **Festung Ehrenbreitstein** an unforgettable venue and an ideal location to welcome participants and set the tone for the conference.

Transportation to and from the venue will be organised, so participants do not need to worry about logistics. We also hope to use the famous **Koblenz Cable Car** to reach the fortress. All further details will be communicated at a later date.

BOAT TRIP - THURSDAY (28 May)

As is ICAME tradition, we will, of course, also host our conference **boat trip**. This much-anticipated social event offers participants a relaxed opportunity to network, continue discussions in an informal setting, and enjoy the scenic landscape along the rivers surrounding Koblenz. During the trip, participants will experience the world-famous Rhine romanticism – *Rheinromantik* – whose dramatic landscapes and castle-lined riverbanks famously inspired Romantic writers, including Lord Byron. The route will take us through parts of the **Kulturlandschaft Oberes Mittelrheintal**, a UNESCO World Heritage Site renowned for its exceptional cultural landscape. More details regarding the route, schedule, and practical arrangements will be announced in due course.



CONFERENCE DINNER AND DISCO – FRIDAY (29 May)

Another ICAME tradition is the **conference dinner and disco**, which will take place at the elegant restaurant and bar **adaccio**, located in the city centre of Koblenz. **adaccio** is situated next to Görresplatz, one of the city's central squares, and offers a stylish setting for an enjoyable evening together. Further information about the venue is available on the restaurant's website: https://www.adaccio.de/

The exact address is

Firmungstraße 2,

56068 Koblenz
Germany

The evening will feature a gala dinner with free drinks until 10:00 pm, followed by a DJ and disco until midnight. This event provides an excellent opportunity to socialise with colleagues, continue conference discussions in a relaxed atmosphere, and celebrate ICAME47 together.

The price for the conference dinner and disco is **€80** per person. You are, of course, welcome to book places for accompanying guests. Please inform the organising committee as soon as possible if you wish to do so, as availability is limited and advance planning is required.

EXCPECTED FEES AND REGISTRATION

While we are also applying for further external funding, the conference fees (including pre-conference workshops, lunch and coffee breaks on each conference day, the boat trip, and the conference warming) are expected to be as follows:

Early-bird regular registration: 369€

Early-bird PhD student registration: 269€

Late regular registration: 419€

Late PhD student registration: 319€

As stated above, the **conference dinner** on Friday (including the gala dinner, free drinks until 10:00 pm, and a DJ and disco until midnight) is priced at **80€** per person.

The registration phase should start on **15 February 2026**. The registration periods are as follows:

- **Early-bird registration:** 15 February – 31 March 2026
- **Late registration:** 1 April – 7 May 2026

The registration process will be conducted via our Conftool platform: **https://www.conftool.org/icame47/**

More information about the registration process will follow shortly.

TRAVEL

As individual travel schedules may vary considerably, participants are expected to make their own travel and accommodation arrangements. If you are travelling from outside continental Europe, the most convenient airports are **Frankfurt** (approx. 1.25 hours away), **Cologne/Bonn** (approx. 1 hour away), and **Düsseldorf** (approx. 1.45 hours away). All three airports offer direct train connections to **Koblenz Hauptbahnhof** (main station).

Train connections can be checked on the Deutsche Bahn website: **https://int.bahn.de/**

Please note that train services in Germany are frequently subject to delays, so we strongly recommend allowing sufficient buffer time when planning your journey.

Koblenz can also be reached by car (e.g. via the A3 → A48 or the A61) and even by boat via the **Rhine** or **Moselle** rivers.

HOTELS AND ACCOMMODATION

As stated before, the academic programme of ICAME47 is going to take place at the University of Koblenz which is located in Koblenz-Metternich [Address: Universitätsstraße 1, 56070 Koblenz, GERMANY]. The social programme of ICAME47 will take place in the city centre of Koblenz (Altstadt). Please note that two major rivers – the Rhine and the Moselle – run through the city. Depending on which side of the river you are staying on, this may affect daily commuting. However, there are regular bus connections from the city centre to the university campus.

A list of hotels with rooms reserved for conference participants is available here:

https://wp.uni-koblenz.de/icame47/wp-content/uploads/sites/211/2025/12/Hotel_Info_ICAME47.pdf

KOBLENZ AND ITS SURROUNDINGS

Situated in one of Germany's most renowned wine regions, **Koblenz** is a popular tourist destination and a wonderful place to explore. Further information and inspiration for sightseeing and leisure activities can be found here:

https://www.visit-koblenz.de/en/

https://www.visit-koblenz.de/en/sights

https://www.koblenz-touristik.de/en/departments/touristik

https://www.visit-koblenz.de/en/tourist-information

POSTER LIGHTNING TALKS AND DISCUSSIONS

We are delighted to announce that we will have **four poster presentations** at the conference. In order to highlight these contributions, we will hold **poster lightning talks** on **Thursday, 28 May**, directly following the morning plenary talk.

During these lightning talks, poster presenters will have **4 minutes** to briefly present a digital version of their poster and introduce their work to the audience in the lecture hall. This will be followed by a coffee break in the foyer, where the physical posters will be displayed and discussed in more detail. There may also be an opportunity for further informal discussion during this time.

QUIET ROOM / WORKING ROOM

A quiet room will be available throughout the conference for participants who wish to work, read, or take a break from the conference programme. This space can be used for individual work or moments of rest and reflection. We hope this room will contribute to a comfortable and inclusive conference experience for all participants.

PLENARY SPEAKERS

Laurence Anthony

Laurence Anthony

Laurence Anthony is Professor of Applied Linguistics at the Faculty of Science and Engineering, Waseda University, Japan, where he serves as a founding member of the Center for English Language Education in Science and Engineering (CELESE). His main research interests are in language data science, AI, corpus linguistics, educational technology, and English for Specific Purposes (ESP) program design and teaching methodologies. He received the National Prize of the Japan Association for English Corpus Studies (JAECS) in 2012 for his work in corpus software tools design, including the creation of AntConc.

Natalia Levshina

Natalia Levshina

Natalia Levshina is an assistant professor of communication and computational methods at Radboud University. Her main research interests are linguistic typology, corpora, cognitive and functional linguistics. She also teaches courses on different topics around AI, including Deep Learning and Large Language Models, chatbots and algorithmic bias. After obtaining her PhD at the University of Leuven in 2011, she worked in Jena, Marburg, Louvain-la-Neuve, Leipzig, where she got her habilitation qualification in 2019, and at the Max Planck Institute for Psycholinguistics in Nijmegen. She has published a book "Communicative Efficiency: Language structure and use" (Cambridge University Press, 2022), in which she formulates the main principles of communicatively efficient linguistic behaviour and shows how these principles can explain why human languages are the way they are. Natalia is also the author of a best-selling statistical manual "How to Do Linguistics with R" (Benjamins, 2015).

Jonathan Culpeper

Jonathan Culpeper is Professor of English Language and Linguistics at Lancaster University, UK. His research spans pragmatics, stylistics and the history of English, all of which he has pursued at some point through corpus methods. A major publication in historical corpus linguistics is *Early Modern English Dialogues: Spoken Interaction as Writing* (2010, CUP; with Merja Kytö). He is currently leading the corpus-based £1 million AHRC-funded *Encyclopaedia of Shakespeare's Language* project, which will provide evidence-based and contextualised accounts of Shakespeare's language.

Jonathan Culpeper

Jane Stuart-Smith

Jane Stuart-Smith

Jane Stuart-Smith has been Professor of Phonetics and Sociolinguistics at the University of Glasgow since 2013, first joining the University in 1997 as Lecturer in English Language, where she has worked with colleagues to develop the Glasgow University Laboratory of Phonetics (GULP). With members of GULP and collaborators outwith Glasgow, she considers the many relationships between speech and society, taking the rich linguistic variation in Scotland as the basis for her work (e.g. **Sounds of the City**). More recently, she has been using corpus phonetic techniques to consider phonetic and phonological variation over space and time in the Englishes of the British Isles and North America (**SPeech Across Dialects of English - SPADE**), and is currently examining **Variability in Child Speech**, in a large longitudinal and cross-sectional cohort of typically-developing children in Scotland. Jane also works closely with colleagues in Scotland to promote the public understanding of phonetics by developing accessible web resources for speech and accents (e.g. **Seeing Speech**; **Dynamic Dialects**; **STAR**).

PRE-CONFERENCE WORKSHOPS – TUESDAY (26 May)

We are delighted to announce five exciting **pre-conference workshops**, all of which will take place on **Tuesday, 26 May**.

This is an overview of the workshop titles and conveners. You can find further information about the workshops and confirmed/expected speakers on the other pages below. Please note that the workshop schedules may be subject to change.

A Confluence of Languages through Corpus-based Contrastive Research in the Age of AI Signe O. Ebeling, Hilde Hasselgård, Marie-Pauline Krielke, Isabell Landwehr, & Sylvi Rørvik contrastive-icame47@ilos.uio.no	**Data management, corpora and AI** Sabine Bartsch (sabine.bartsch@tu-darmstadt.de), Ilka Mindt (mindt@mail.uni-paderborn.de)	**Corpus-Linguistic Approaches to Reddit** Daria Dayter (daria.dayter@tuni.fi), Thomas C. Messerli (thomas.messerli@unibas.ch) & Sven Leuckert (sven.leuckert@tu-dresden.de)
ICE Corpora in the Age of AI Ulrike Gut (gut@uni-muenster.de), Stella Neumann (stella.neumann@ifaar.rwth-aachen.de) & Gerold Schneider (gschneid@cl.uzh.ch)	**Corpus and Computational Linguistics meet Fake News, Mis- and Disinformation and LLMs** Silje Susanne Alvestad (s.s.alvestad@ilos.uio.no) & Nele Põldvere (nele.poldvere@ilos.uio.no)	

WORKSHOP: A Confluence of Languages through Corpus-based Contrastive Research in the Age of AI

Signe O. Ebeling, Hilde Hasselgård, Marie-Pauline Krielke, Isabell Landwehr, & Sylvi Rørvik
(contrastive-icame47@ilos.uio.no)

The workshop takes the general conference theme as its starting point, as reflected in its title, and thus combines both the longstanding tradition of having a contrastive workshop at ICAME conferences and the start of a new era for corpus linguistics represented by the emergence of AI. As can be seen below, the proposed contributions also bridge the potential gap of traditional corpus linguistics and the influx of AI-inspired topics and methods: Eight of the proposed papers either explore AI-generated datasets or employ AI to supplement traditional corpus methods, while the other four papers represent more traditional contrastive corpus-based research. In this way, this proposed 15th iteration of the ICAME contrastive pre-conference workshop is a mirror image of the current context for corpus linguistics.

This is the workshop programme:

Time	Title
9.30-9.45	Welcome and opening
9.45-10.15	**Sylvi Rørvik**: Noun-phrase complexity in English and Norwegian research articles: A comparison of AI-generated and human-authored texts.
10.15-10.30	**Markéta Malá**: Translating the information structure: Indefinite subjects in human and machine translations from English to Czech.
10.30-11.00	Coffee break
11.00-11.30	**Signe Oksefjell Ebeling & Jarle Ebeling**: Making sense of GenAI for contrastive purposes: Verbs of sensing in English and Norwegian.
11.30-12.00	**Hilde Hasselgård**: Attitudinal stance adverbials in English and Norwegian.
12.00-12.30	**Lobke Ghesquière & Faye Troughton**: Artificial Intelligence or Intelligent Analysis: Using ChatGPT to contrast exclamative constructions.
12.30-14.00	Lunch
14.00-14.30	**Isabell Landwehr, Marie-Pauline Krielke & Sergei Bagdasarov**: Echoes of AI: English, German and Spanish news language before and after ChatGPT.
14.30-15.00	**Noelia Ramón & Belén Labrador**. Exploring the use of AI-enhanced tools for corpus-based contrastive research: A case study on promotional texts in English and Spanish.
15.00.15.30	**Jolanta Šinkūnienė**: Moves, Stance and Engagement in the concluding sections of Research Articles in Lithuanian and English in two disciplines.
15.30-16.00	Coffee break

16.00-16.30	**Øyvind Thormodsæter & Signe Laake**: "And the reason is…": A cross-linguistic investigation of the English patterns *The reason is because / The reason is that* and the Norwegian patterns *Grunnen er fordi / Grunnen er at*.
16.30-17.00	**Ronel Wasserman, Bertus van Rooy & Megan van Winsen**: African English and Afrikaans past time reference: contrasts and competition in contact since the 19th century.
17.00-17.15	Closing of workshop

In total, the proposed contributions comprise comparisons of English with six other languages (Afrikaans, Czech, German, French, Lithuanian and Norwegian), and material from a range of registers, including news, promotional discourse, academic prose, and business communication. Considering the variation in approaches and in language pairs investigated, we believe the proposed workshop would constitute a worthy addition to the long line of previously organized successful contrastive ICAME workshops, as attested by the below list of publications.

Previous workshop publications:

Aijmer, K. & Altenberg, B. (eds) (2013). *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson*. Amsterdam: John Benjamins. (ICAME32, Oslo 2011).

Aijmer, K. & Hasselgård, H. (eds) (2015). *Cross-linguistic Studies at the Interface Between Lexis and Grammar*. Special issue of Nordic Journal of English Studies, (Vol 15:1). (ICAME34, Santiago de Compostela 2013).

Altenberg, B. & Aijmer, K. (eds) (2013b). *Text-based Contrastive Linguistics*. Special issue of Languages in Contrast (Vol. 13:2). (ICAME33, Leuven 2012).

Brems, L., Ghesquière, L. & Vanderbauwhede, G. (eds) (forthcoming). Special issue of *English Text Construction* (ICAME45, Vilnius 2025).

Čermáková, A., Ebeling, S.O., Levin, M. & Ström Herold, J. (eds) (2021). *Crossing The Borders:Analysing Complex Contrastive Data*. Bergen Language and Linguistics Studies (BeLLS), (Vol 11: 1). (ICAME41, Heidelberg 2020).

Čermáková, A., Egan, T., Hasselgård, H. & Rørvik, S. (eds) (2021). *Time in Languages, Languages in Time.* Amsterdam: John Benjamins. (ICAME40, Neuchâtel 2019).

Čermáková, A., Hasselgård, H., Malá, M. & Šebestová, D. (2024). *Contrastive Corpus Linguistics: Patterns in Lexicogrammar and Discourse*. Bloomsbury (ICAME42, Dortmund 2021).

Ebeling S.O. & Hasselgård, H. (eds) (2015). *Cross-linguistic Perspectives on Verb Constructions*. Newcastle: Cambridge Scholars Publishing. (ICAME35, Nottingham 2014).

Ebeling S.O. & Hasselgård, H. (eds) (2018). *Corpora et Comparatio Linguarum: Textual and Contextual Perspectives.* Bergen Language and Linguistics Studies (BeLLS) (Vol 9: 1). (ICAME38, Prague 2017).

Ebeling S.O. & Hasselgård, H. (eds) (2024). *English in Contrast: Corpus-based Approaches*. Special issue of Nordic Journal of English Studies, (Vol 23:2). (ICAME44, Vanderbijlpark 2023).

Egan, T. & Dirdal, H. (eds) (2017). *Cross-linguistic Correspondences.* Amsterdam: John Benjamins. (ICAME36, Trier 2015).

Janebová, M., Lapshinova-Koltunski, E. & Martinková, M. (eds) (2017). *Contrasting English and other Languages through Corpora*. Newcastle: Cambridge Scholars Publishing. (ICAME37, Hong Kong 2016).

Levin, M. & Ström Herold, J. (eds) (2024) Special issue of *Languages in Contrast*. (ICAME43, Cambridge 2022).

Rørvik, S. & Izquierdo, M. (eds) (Forthcoming) *Cross-linguistic Register Variation.* Amsterdam: John Benjamins. (ICAME45, Vigo 2024).

WORKSHOP: ICE CORPORA IN THE AGE OF AI

ULRIKE GUT, STELLA NEUMANN & GEROLD SCHNEIDER

gut@uni-muenster.de, stella.neumann@ifaar.rwth-aachen.de, gschneid@cl.uzh.ch

The International Copus of English (ICE) project, founded in 1990 (Greenbaum 1996, Greenbaum & Nelson 1996), has been a tremendous success: 15 ICE corpora of Englishes around the world have been completed to date and numerous articles on the use and structures of varieties of English have been published on findings generated by these corpora. What role do the ICE corpora play today, 35 years after their conception? On the one hand, 35 years after their conception, ICE corpora are increasingly being criticised as being too small, using an outdated data format and being difficult to handle for automatic analyses. Moreover, those ICE corpora collected
20 years ago may be considered outdated or at least not representing the contemporary use of the respective variety of English anymore (e.g. Botha & Bernaisch 2025). More generally, the original design, to some extent modeled on contexts of use applying to Great Britain, turned out not to match the reality of all varieties of English. On the other hand, ICE corpora have been shown to still constitute excellent sources for research on varieties of English even in comparison with big data (Loureiro-Porto 2017). In particular, coverage of a range of spoken and written registers sets the ICE corpora apart from more recent corpora. By the same token, some ICE corpora have been updated and extended to include modern data formats as well as new annotations (e.g., Conrad et al. 2025, Gut & Fuchs 2017, Schützler et al. 2017, Kallen & Kirk 2012, Wong et al. 2011, Wunder et al. 2010).

This workshop addresses the question of what role ICE corpora can play in the age of AI: Do they still constitute a good source for research, especially compared to the mega corpora of online data? How can they be updated for easier processing? How can new ICE cropora be collected using AI methods such as Whisper for the automatic transcription of spoken data? Should they be increased in size? Should new text categories be added and/or the original corpus design be revised? How can we be sure not to include texts generated by AI?

The workshop is intended to bring together people involved in the (i) compilation, (ii) computational handling and (iii) use of the International Corpus of English.

References:

Botha, W., & Bernaisch, T. (2025). World Englishes and sociolinguistic variation. *World Englishes, 44*, 2–11. DOI: 10.1111/weng.12695

Conrad, S., Neumann, S., Frenken, F., & Schneider, G. (2025). Updating the international corpus of English for the 21st century: Towards a standardized XML-compliant markup. *Corpus Linguistics 2025 Book of Abstracts*, 74.

Gut, U., & Fuchs, R. (2017). Exploring speaker fluency with phonologically annotated ICE corpora. *World Englishes, 36*, 387–403.

Greenbaum, S. (1996). *Comparing English Worldwide*. Oxford University Press.

Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) project. *World Englishes*, *15*, 3–15.

Kallen, J. & Kirk, J. (2012). SPICE-Ireland: A User's Guide. https://johnmkirk.etinu.net/johnmkirk/documents/003648.pdf

Loureiro-Porto, L. (2017). ICE vs GloWbE: Big data and corpus compilation. *World Englishes, 36*, 448–70. DOI: 10.1111/weng.12281

Schützler, O., Gut, U., & Fuchs, R. (2017). New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In S. Hancil & J. Beal (Eds.), *Perspectives on Northern Englishes* (pp. 273–301). De Gruyter Mouton.

Wong, D., Cassidy, S., & Peters, P. (2011). Updating the ICE annotation system: Tagging, parsing and validation. *Corpora*, *6*(2), 115–144. https://doi.org/10.3366/cor.2011.0009

Wunder, E.-M., Voormann, H., & Gut, U. (2010). The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal, 34*, 78–88.

PRELIMINARY WORKSHOP PROGRAMME

Time	Presenter	Title
14.00-14.20	Philipp Meer and Polina Kashkarova (University of Münster)	From annotation to automation: Leveraging AI to build phonological corpora of West African Englishes
14.20-14.50	Henning Schreiber and Ismail Afolabi (Hamburg University)	ICE Transformers
14.50-15.20	John Kirk (University of Vienna)	ICE-Corpora: Reconsideration in the age of AI
15.30-16.00	*Coffee break*	
16.00-16.20	Aishath Suad (Maldives National University), Tobias Bernaisch (Justus Liebig University Giessen), Julia	AI-enhanced and smartphone-based compilation of spontaneous speech: The creation of the Corpus of Spoken Maldivian English

	Degenhardt (Augsburg University), Barbara Güldenring (Justus Liebig University Giessen), Eliane Lorenz (Justus Liebig University Giessen)	(CoSpoMalE)
16.20-16.40	Stella Neumann (RWTH Aachen University)	ICE21: Suggestions for an updated high-quality corpus capturing the diversity of English
16.40-17.00	Robert Fuchs (University of Bonn)	Balancing backward comparability and register adequacy: Towards an ICE 2.0 text category design in the age of AI
17.10-17.30	Final discussion	

WORKSHOP: Corpus-Linguistic Approaches to Reddit

Daria Dayter, daria.dayter@tuni.fi
Thomas C. Messerli, thomas.messerli@unibas.ch
Sven Leuckert, sven.leuckert@tu-dresden.de

Reddit is a prominent social platform on which millions of international users gather every day to discuss an enormous range of topics in over a hundred thousand different communities called 'subreddits' (Proferes et al. 2021). As an active site of digital language use, Reddit has also emerged as a popular resource for linguistic research, including work from pragmatic (e.g., Dayter & Messerli 2022; Androutsopoulos 2023), sociolinguistic (e.g., Dynel & Poppi 2019; Leuckert & Leuckert 2020), and register-based (e.g., Biri 2022; Liimatta 2022) perspectives. It serves both as a specific site of situated language use and as a sandbox for the exploration of digital discourses more generally.

The goals of the proposed workshop are (1) to showcase the potential of language use on Reddit specifically for corpus-linguistic approaches; (2) to highlight the multifaceted nature of communicative practices on the platform; and (3) to open up avenues for further collaboration. Contributors to the workshop present case studies of their research on Reddit, emphasising methodological challenges and questions. These include, among others, increasingly AI-produced contributions on Reddit, changes in how data may be accessed and reproduced, and research ethics. Overall, the workshop serves to bring together corpus linguists working with Reddit data and to advance social media research in linguistics.

The panel features talks that exemplify the diversity of corpus-linguistic approaches to Reddit discourse. Together, these contributions discuss the richness of Reddit as a site for corpus-based linguistic inquiry.

References:

Androutsopoulos, J. (2023). Punctuating the other: Graphic cues, voice, and positioning in digital discourse. Language & Communication, 88, 141–152. https://doi.org/10.1016/j.langcom.2022.11.004

Biri, Y. (2022). Epistemic stance in the climate change debate: A comparison of proponents and sceptics on Twitter and Reddit. *Register Studies*, *4(2)*, 232–262.

Dayter, D. & Messerli, T. C. Messerli (2022). Persuasive language and features of formality on the r/ChangeMyView subreddit. *Internet Pragmatics*, *5(1)*, 165–195.

Dynel, M., & Poppi, F. I. M. (2019). Risum teneatis, amici?☆: The socio-pragmatics of RoastMe humour. Journal of Pragmatics, 139, 1–21. https://doi.org/10.1016/j.pragma.2018.10.010

Leuckert, S. & Leuckert, M. (2020). Towards a digital sociolinguistics: Communities of practice on Reddit. In S. Rüdiger & D. Dayter (Eds.), *Corpus Approaches to Social Media* (pp. 15–40). John Benjamins.

Liimatta, A. (2022). Do registers have different functions for text length?: A case study of Reddit. Register Studies, 4(2), 263–287. https://doi.org/10.1075/rs.22007.lii

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, *7(2)*, 1–14.

PRELIMINARY WORKSHOP PROGRAMME

08:00-09:00	**Registration and workshop prep**
09:00-10:30	• Workshop opening • "thatsthejoke.jpg – Affordances and referentiality on Reddit", Max Hoferichter (University of Greifswald) (**WIP**) • "dOnT uSe ThAt WeIrD sPoNgEbOb MeMe: Alternating caps as a strategy for voice management in digital discourse", Axel Bohmann (University of Cologne)
10:30-11:00	**Coffee break**
11:00-12:30	• "'He's like a god damned laser guided missile': A corpus-assisted analysis of metaphor use in the R/Dementia subreddit", Gavin Brookes (Lancaster University) • "Ban-happy mods and shadow bans: A corpus-pragmatic analysis of speech acts on Reddit", Claudia Lange (Technische Universität Dresden) & Sven Leuckert (Technische Universität Dresden) • "Pragmatic strategies and audience response: Advice-giving and upvotes on Reddit", Rickey Lu (The Hong Kong Polytechnic University)
12:30-14:00	**Lunch**
14:00-15:30	• "Comment type sequences on Reddit: A data-driven functional-pragmatic approach", Ylva Biri (University of Helsinki) & Aatu Liimatta (University of Helsinki) • "Evaluative practices and persuasion in online debates: a corpus-based study of /r/changemyview", Daria Dayter (University of Tampere) & Thomas C. Messerli (University of Basel)

WORKSHOP: Data management, corpora and AI

Sabine Bartsch (sabine.bartsch@tu-darmstadt.de), Ilka Mindt (mindt@mail.uni-paderborn.de)

Corpus linguistics has a long tradition of empirical studies and has been at the forefront of implementing digital technology in research and teaching. The most recent empirical and digital turn raises a number of methodological questions concerning the suitability, representativeness, access, storage and archiving of our data. Furthermore, there are concomitant questions concerning the research processes and expressivity of our findings. Related to this are questions concerning our own accountability in the research process including transparency and reproducibility of the research process as well as data accessibility leading to questions of research data management and the role of data publication. And lastly, there are questions surrounding our agreement on the accepted methodology within our research community as well as the development of the required data literacy among researchers who are, more often than not, simultaneously teachers passing on methodological competencies and values within the community. These questions have become even more acute in the context of growing corpus sizes often collected from less well curated sources and whose sheer size necessitates automated and quantitative approaches for both annotation and analysis that offer a wealth of new insights, but at the same time pose a challenge for comprehensive analysis and close inspection of language samples. The call for linguistics to critically engage with handling a wider spectrum of ever larger corpora while ensuring transparency and reproducibility of research processes and findings is all the more pressing in the context of large language models and the potential and challenges of artificial intelligence applications in corpus linguistics.

In order to further the discussion of these issues in the research community, we propose a half-day workshop that brings together people working in the areas of corpus analysis, corpus compilation, data collection, data management, archive hosting etc. We would like to turn this workshop into a forum for discussion and exchange on desiderates, best-practices, accessibility options and exchange ideas on how data management, methodology and AI use may shape the future.

PRELIMINARY WORKSHOP PROGRAMME

Time	Presenter	Title
14.00-14.30	Martin Wynne (University of Oxford)	Fifty years of data management at the Oxford Text Archive
14.30-15.00	Christoph Draxler (LMU Munich)	Automatic Speech Recognition: 3+1 Stairways to Heaven

15.00-15.30	Peter Uhrig (FAU Erlangen-Nürnberg)	Corpus Management and Annotation for Large Multimodal Corpora
15.30-16.00	*Coffee break*	
16.00-16.30	Mark Davies (Brigham Young University)	Integrating insights from AI/LLMs into English-Corpora.org

WORKSHOP: Corpus and Computational Linguistics Meet Fake News, Mis- and Disinformation and Large Language Models

Silje Susanne Alvestad & Nele Põldvere,
s.s.alvestad@ilos.uio.no, nele.poldvere@ilos.uio.no

This workshop will take a corpus- and computational-linguistics perspective on *fake news* and related phenomena, where *fake news* is defined along the axes of veracity and honesty, giving rise to three types: 1) false but honest news, such as errors, which corresponds to *misinformation*; 2) false and dishonest news, such as lies; and 3) true but dishonest news, in which crucial pieces of information may be omitted (so as to fit a certain narrative, as seen, arguably, in *propaganda*), or in which true information may be taken out of context. Fake news types 2) and 3) involve an intention to deceive and so overlap with typical definitions of *disinformation* (see Grieve & Woodfield, 2023).

Fake news and related information disorders can be harmful to our societies. Specifically, when we change our beliefs and subsequent behaviour based on false or misleading information it can harm our health and lives, sow distrust (Funk et al., 2023), and disrupt election processes (Jamieson 2018). Now, the societal challenge posed by information disorders is amplified by the rapid development within generative AI, exemplified by Large Language Models (LLMs), with the launch of OpenAI's ChatGPT in November 2022 as a significant milestone. The output of LLMs depends on their training data, which can contain inaccuracies and biases. As a result, these models may unintentionally spread mis- or disinformation (Brandtzæg et al., 2023). They can also produce "hallucinations"—convincing but false statements (Spitale et al., 2023)—or partly incorrect content due to unreliable sources (Chen et al., 2023). This blend of fabricated and biased information makes it difficult to ensure the accuracy of online content (Buchanan et al., 2021). Moreover, LLMs hold the potential to generate misleading or false information at scale and at a quality that makes it indistinguishable from similar content authored by humans. Controlled experiments show that LLM-generated messages can change policy attitudes, at times matching or surpassing human levels of persuasiveness (Bai et al., 2025; Salvi et al., 2025). Research has shown that people find it more difficult to identify disinformation produced by AI than similar content produced by humans (Zhou et al., 2023), and in simulated news recommendation systems, researchers have found a new phenomenon referred to as "truth decay", by which genuine news increasingly falls behind LLM-generated mis- and disinformation in visibility and ranking. This shift happens because LLM-generated content typically shows lower perplexity, making it appear more fluent and familiar. As a result, such content often receives higher recommendation scores and greater visibility (Hu et al., 2025). This dynamic has serious implications for the spread of mis- and disinformation, since increased exposure can boost perceived credibility through the illusory truth effect. All of this highlights the need for effective identification and

verification systems. We believe that especially corpus and computational linguists should recognize the urgency of the moment and hereby be invited to act.

Against this background, our workshop will shed light on the rising societal challenge posed by information disorders from a corpus- and computational-linguistics perspective. We ask questions including, but not limited to, what the linguistic features are of such information disorders, whether the disorders can be identified based on such features, whether the features have changed, and are changing, over time, what the capabilities and limitations of various LLMs are in the context of producing and disseminating misleading information, whether the LLMs have any fingerprint in the context of mis- and disinformation, and how to develop a best practice for linguistic investigations of LLM output.

References:

Bai, H., Voelkel, J., Muldowney, S., Eichstaedt, J., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications, 16*, Article 61345. https://doi.org/10.1038/s41467-025-61345-5

Brandtzaeg, P. B. (2023). “Good” and “Bad” Machine Agency in the Context of Human-AI Communication: The Case of ChatGPT. In *International Conference on Human-Computer Interaction*, (pp. 3–23). Springer Nature Switzerland.

Buchanan, B., Lohn, A., Musser, M. & Sedova, K. (2021). Truth, lies and automation. How language models can change disinformation. Center for Security and Emerging Technology, May 2021. https://doi.org/10.51593/2021CA003

Chen, C. & Shu, K. (2023). Combating misinformation in the age of LLMs: Opportunities and challenges. https://doi.org/10.48550/arXiv.2311.05656

Funk, A., Shahbaz, A., & Vesteinsson, K. (2023). *Freedom on the Net 2023. The repressive power of artificial intelligence*. Freedom House report. https://freedomhouse.org/sites/default/files/2024-10/FOTN2023Final24.pdf

Grieve, J., & Woodfield, H. (2023). *The Language of Fake News*. Cambridge Elements in Forensic Linguistics. Cambridge University Press

Hu, B., Sheng, Q., Cao, J., Li, Y., & Wang, D. (2025). LLM-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 435–445). Association for Computing Machinery. https://doi.org/10.1145/3726302.3730027

Jamieson, K. H. (2018). *Cyberwar: How Russian hackers and trolls helped elect a president. What we don't, can't, and do know.* Oxford University Press.

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature human behaviour*, *9*(8), (pp. 1645–1653). https://doi.org/10.1038/s41562-025-02194-6

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. doi: 10.1126/sciadv.adh1850.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & Choudhury, M. D. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581318

PRELIMINARY WORKSHOP PROGRAMME

08:00-09:00	**Registration and workshop prep**
09:00-10:30	• Silje Susanne Alvestad & Nele Põldvere (University of Oslo): Workshop opening and talk "Corpus and Computational Linguistics Meet Fake News, Mis- and Disinformation and Large Language Models" • Sophie Llewellyn (University of Oslo): *Stance expressions in AI-generated vs. human-written fake news* • Marina Ernst & Frank Hopfgartner (University of Koblenz): *Can LLMs be deceived into trusting disinformation?*
10:30-11:00	Coffee break
11:00-12:30	• Stefano Sbalchiero, Alessandro Meneghini, & Arjuna Tuzzi (University of Padova): *Fake and real, beyond binary classification in fake news detection. Challenges and solutions with topic modelling techniques* • Fabio Carrella (University of Campinas): *Testing community-level interventions in online echo chambers using LLM agents* • Workshop Wrap-up