BOOK OF ABSTRACTS

Digital Research Data and Human Sciences (DRD Hum)

2024

University of Eastern Finland

Joensuu, Finland, 10-12 December

BOOK OF ABSTRACTS

DRD HUM Joensuu 10th – 12th of December 2024

This booklet is divided into three parts:

1. Plenaries
2. Paper Presentations
3. Posters

The abstracts are presented in alphabetical order, based on the lead-author's surnames.

NB: While we have striven to obtain uniformity, not all contributors were able to submit their abstracts in line with the template we use. Unfortunately, we did not have the time and resources to change every submission.

Abstracts:

Plenary Speakers

How I went from mining to constructing to making literary data

Bode, Katherine
Australian National University, Australia
katherine.bode@anu.edu.au

Plenary Abstract

This paper describes how, in my data work, I went from being certain, to being uncertain, to understanding that un/certainties have nothing to do with the messy realities of textual practices. Alternatively, this paper tells the story of the past fifteen years of my work with data, and how I went from presuming its objectivity, to proclaiming its subjectivity, to exploring its performativity. In the process, I raise some conceptual and practical problems with the representational logic that organizes our thinking about (and with) data, and I suggest making as a paradigm for engaging with digital research data in the age of artificial intelligence. In considering the value/s of data work for the human sciences, I ask what we give up and what we gain if we refuse samples, biases, accuracy, transparency, and results, and instead approaching data making as a practice of making that values attention to detail, responsiveness, creativity, responsibility, trust, and care.

Keywords: Data, Performativity, Representativeness, Subjectivity, Transparency.

Tracing the Bias Loop: Data, Culture, and Society in the era of Artificial Intelligence

Foka, Anna
Uppsala University, Sweden
anna.foka@abm.uu.se

Plenary Abstract

Artificial Intelligence (AI) is expanding beyond the realm of computer science, significantly impacting various industries and facets of society. AI and Machine Learning are revolutionizing the fields of computer science and engineering, making professionals in these areas highly sought after. The future of AI is dynamic, already profoundly influencing education, professions, and society. However, there is a growing awareness of the dangers of over-relying on AI for every challenge. In this context, I advocate for embracing AI with a societal and humanistic sensibility. My focus is on the Archives, Libraries, and Museums (ALM) sector. Traditionally, professional decisions in ALM have been made by educated humans using their best judgment. However, machines are becoming increasingly influential in performing these tasks. The quality of machine decision-making is closely tied to the quality of the data and the parameters selected for classification. Using cultural heritage collections as an example, I discuss how AI offers new opportunities for access and engagement but also risks perpetuating historical biases embedded in these collections. I emphasize the importance of interdisciplinary collaboration among cultural heritage professionals, data scientists, and social scientists to identify and mitigate bias. We need to think together with machines to adopt a holistic approach to bias mitigation, integrating both technical and non-technical solutions to ensure that AI technologies contribute to a more inclusive society. I conclude with recommendations for future education, research, and practice, highlighting the need for ongoing monitoring and interdisciplinary collaboration to address the complexities of bias in AI applications within the cultural and creative industries, as well as in the training of professionals in the ALM sector.

Keywords: AI, ALM (Archives, Museums and Libraries), Classification, Cultural Heritage, Holistic Approach.

Corpus linguistics for the digital age – Making sense of the world through language and stories

Michaela Mahlberg

Friedrich-Alexander-Universität Erlangen-Nürnberg

Paper Abstract

In today's digital age, many areas of the humanities and social sciences have seen digital turns and there is plenty of innovation in methods and tools. Corpus linguistics is a discipline that aims to describe language on the basis of large sets of linguistic data. The study of large amounts of language data makes it possible to see patterns in language use. Treating language as data and identifying patterns brings linguistics increasingly closer to data sciences. At the same time, technological advancements and especially developments in AI raise questions about the relationship between patterns in data and human creativity. In this talk, I want to argue that – in the digital age – the focus of corpus linguistics needs to shift from developing digitally supported methods to developing theoretical foundations for language analysis. People use language to act, to do things and to interact with one another. Language is used to tell the stories and create the narratives that shape our society, our culture and our reality. It is time to consider how insights from corpus linguistics can shed light on the relationship between patterns and creativity in language use. To address this question, it is also crucial to look at how fictional stories and real-world narratives are interconnected. A better understanding of how people use language and stories to make sense of the world will be vital to tackle the challenges that humanity faces at the present time.

How Corpus Data is Interpreted and Corpus Analyses are Contextualised

Tony McEnery

a.mcenery@lancaster.ac.uk

Plenary Abstract

What does it mean to interpret corpus data? In this talk I will look at how corpus analyses are contextualised. Beginning from the position that a corpus is self contained, I will proceed, through a series of case studies, to show how the inclusion of context, in various forms, is essential to our understanding of both what a corpus is and what we understand when we analyse it. While the case studies will all be based on English, they will come from different centuries and social contexts and will look at a range of modes of production and registers. The degree to which the observations flex as we move between different levels of linguistic interpretation will be considered also. The talk will conclude by urging all corpus users to consider as broad a context as possible in their analyses.

Abstracts:

Paper Presentations

Regulation of AI? Comparing Czech and Portuguese Media Imaginaries with CADS

Raquel Amaro, Tibor Vocásek
CLUNL NOVA Lisbon, Portugal; ISS – FSV, CUNI Prague, Czechia
raquelamaro@fcsh.unl.pt; tibor.vocasek@fsv.cuni.cz

Paper Abstract

The recent buzz around Artificial Intelligence (AI) has raised significant debate in the EU on how to regulate it. Like any emerging technology, AI development depends on media-shaped public perception (Chuan et al., 2019). This research aims to contribute to the study of this perception through an interdisciplinary comparative perspective. It analyses AI regulation representations in the Czech and Portuguese online mainstream media, using Corpus Approaches to Discourse Studies (CADS) (Baker et al., 2008; Partington et al., 2013).

Czechia and Portugal are similar in area size, population, or GDP (Eurostat, 2024) but differ significantly regarding their tech sectors. While news headlines labelled Czechia the "sick man of Europe" due to its stuttering industrial economy (Willoughby, 2023), Portugal, with its vivid tech scene, was referred to as "Europe's Silicon Valley" (Böhnisch, 2021). Focusing on the EU debate, these countries' membership differs by almost two decades.

CADS combines corpus linguistics with traditional CDA while reflecting on its critiques (Orpin, 2005). It mitigates issues of data representativeness and interpretative transparency. CADS allows the investigation of the aggregate effects of language, highlighting typical discursive patterns. Conceptually, this study approaches AI regulation through "sociotechnical imaginaries", understood as "collectively held, institutionally stabilised, and publicly performed visions of desirable futures" (Jasanoff and Kim, 2015, p. 4).

Current studies using imaginaries are plagued by conceptual ambiguity (Richter et al., 2023). This research overcomes the issue by adopting a three-level imaginary concept (Sau, 2021) operationalised with CADS. It structures imaginary analysis by asking for representations of (1) social commentary, (2) vision of the future, and (3) means to achieve it. The analysis covers the period of discussions about the EU's "AI Act" regulation (3/2018-12/2023).

Comparable corpora, collected from digitally available national media in each country in this period, are compiled and explored using Sketch Engine (Kilgarriff et al., 2014) in two steps. Firstly, distinct imaginary layers are investigated via collocations of the topic-related keywords and analysed via concordances in each corpus. Secondly, the analysis focuses on keywords of the sub-corpus after the ChatGPT onset to reveal specifics of the "ChatGPT-moment". The results of both steps are then compared against each other.

Although not groundbreaking in terms of methodology, such research provides innovative, empirically rooted comparative insights into the current media debate on AI. It exhibits the usefulness of the imaginary concept for the CADS while providing a clearer perspective of sociotechnical imaginaries by grounding these to objective linguistic cues.

Keywords: Artificial Intelligence, Corpus Approaches to Discourse Studies, Media Discourse, Regulation of Digital Technologies, Sociotechnical Imaginaries

REFERENCES

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273–306. https://doi.org/10.1177/0957926508088962.

Böhnisch, M. (2021). *Portugal: Europe's Silicon Valley?* Deutsche Welle. https://www.dw.com/en/portugal-europes-silicon-valley/video-57359402.

Chuan, C.-H., Tsai, W.-H. S., & Cho, S. Y. (2019). Framing Artificial Intelligence in American newspapers. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1145/3306618.3314285.

Eurostat. (2024). *Digital Economy and Society*. https://ec.europa.eu/eurostat/web/digital-economy-and-society.

Jasanoff, S., & Kim, S.-H. (2016). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. The University of Chicago Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (n.d.). *The Sketch Engine: Ten Years on*. Lexicography. https://journal.equinoxpub.com/lexi/article/view/17961.

Orpin, D. (2005). Corpus linguistics and critical discourse analysis. *International Journal of Corpus Linguistics*, *10*(1), 37–61. https://doi.org/10.1075/ijcl.10.1.03orp.

Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, *5*(2), 83–108. https://doi.org/10.3366/cor.2010.0101.

Richter, V., Katzenbach, C., & Mike, S. (2023). Imaginaries of AI. In Simon Lindgren (Eds.), *Handbook of Critical Studies of Artificial Intelligence*. Preprint. Edward Elgar Publishing Ltd. https://doi.org/10.26092/elib/2190.

Sau, A. (2021). On cultural political economy: A Defence and constructive critique. *New Political Economy*, *26*(6), 1015–1029. https://doi.org/10.1080/13563467.2021.1879758.

Willoughby, I. (2023, December 11). *Economist Tomáš Dvořák on why Czechia is "Sick man of europe" – and how to move forward*. Radio Prague International. https://english.radio.cz/economist-tomas-dvorak-why-czechia-sick-man-europe-and-how-move-forward-8802580.

Acknowledgements

This study was supported by the Charles University, project GA UK No. 294723.

Part of this research is supported by the Portuguese national funding through the FCT – Portuguese Foundation for Science and Technology, I.P. as part of the project UIDB/LIN/03213/2020; 10.54499/UIDB/03213/2020 and UIDP/LIN/03213/2020; 10.54499/UIDP/03213/2020 – Linguistics Research Centre of NOVA University Lisbon (CLUNL).

Assessing the Linguistic Characteristics of AI-Generated Texts Across Different Registers

Tony Berber Sardinha

Paper Abstract

Recent studies that compare AI-generated texts to those authored by humans predominantly focus on lexical characteristics. This has resulted in a limited understanding of the ability of AI to mimic human writing from a lexicogrammatical perspective. Furthermore, this body of research often overlooks the role of register variation, neglecting to examine the degree to which AI-generated texts reflect the register-specific features found in human-authored texts. The premise of this study is that an evaluation of AI-generated text quality necessitates consideration of register. Prior corpus-based analyses of human-authored texts have convincingly shown that register significantly influences linguistic variation (Biber, 2012). Hence, for assessments aiming to determine equivalency between human-authored and AI-generated texts, register must be taken into account. It is postulated that the majority of training data for Large Language Models lacks explicit register labels, leading to a predominance of inferred over directly learned register distinctions by AI, which raises concerns about the precision and dependability of register knowledge in AI models. In this paper, we employ Multi-Dimensional (MD) Analysis (Biber, 1988, 1995; Berber Sardinha & Veirano Pinto, 2014, 2019) to assess the similarity between AI-generated and human-authored texts. This involves a detailed MD analysis of a corpus comprising register-specific texts produced by humans in natural settings and texts generated by ChatGPT 3.5. The comparison is grounded in the five principal dimensions of register variation identified by Biber (1988), which are determined by sets of co-occurring lexicogrammatical features. Both AI and human subcorpora include four distinct registers: news reports, research articles (in Chemistry and Applied Linguistics), student compositions, and conversations, with each category containing 100 texts, for a total of 800 texts (546,568 words). The human-authored subcorpus was compiled from verified sources that predate the public availability of AI to avoid any AI-generated content. The MD analysis indicated notable differences between AI-generated and human-authored texts across the individual registers and the five dimensions, with AI-generated texts generally not mirroring their human counterparts accurately. Additionally, a linear discriminant analysis, conducted to evaluate the capability of dimension scores to predict text authorship, showed that AI-generated texts could be distinguished with relative ease based on their multidimensional profiles. The findings highlight the existing challenges AI faces in replicating natural human communication effectively. The specifics of the register-based comparisons will be elaborated in the full paper.

Keywords: Multi-Dimensional Analysis, Register Variation, Artificial Intelligence

Multi-Dimensional Collocational Analysis of Discourses around COVID-19 Therapies

Tony Berber Sardinha, Maria Claudia Nunes Delfino, Ana Bocorny, Deise Prina Dutra, Simone Sarmento & Paula Tavares Pinto
Institution, country
email@email.com

Paper Abstract

The goal of this study is to describe the discourses related to the endorsement or opposition of alternative COVID-19 treatments during the pandemic. This was achieved by compiling a corpus of academic articles that either advocated for or criticized the use of treatments like hydroxychloroquine and azithromycin, in accordance with recommendations from the WHO and other health organizations. The dataset was selected to represent both perspectives equally. Our methodology employed Lexical Multi-Dimensional Analysis (LMDA; Berber Sardinha & Fitzsimmons-Doolan, 2024), an offshoot of Multi-Dimensional Analysis (Biber, 1988, 1995; Berber Sardinha & Veirano Pinto, 2014, 2019). This approach focuses on lexical features (e.g., lemmas) and applies multivariate statistical techniques, such as Factor Analysis, to identify correlated lexical features across the texts. Specifically, we examined the keyword collocations within two subgroups: pro-alternative treatment (PAT) and against-alternative treatment (AAT). Keywords for each subgroup were identified by using the other subgroup as a reference corpus. We then determined the collocations for each keyword within a four-word window on either side, comparing these within both the PAT and AAT datasets. For each keyword, this process yielded two sets of collocates, one for PAT and one for AAT. To manage differences in sample size, we selected the 500 most significant collocates from each subgroup based on their logDice scores. This enabled a direct comparison of how keyword collocation changes with differing treatment perspectives. For example, the colloca$on shi] for "hydroxychloroquine" includes its associa$on with 'treatment' in PAT texts, highlighting potential benefits and supportive guidelines, whereas in AAT texts, it is linked to discussions of mixed results concerning recovery times and side effects. Another example of collocation shift refers to ‘patients’: in the PAT texts, its collocates suggest a focus on high-risk individuals needing urgent care or facing greater health risks; conversely, in the AAT texts, 'patients' are depicted as participants in research, emphasizing a scientific evaluation of treatments' effects, efficacy, and safety. Through LMDA, we identified the major dimensions in keyword collocate use across the corpus. Initial results include a dimension contrasting discourses: one promotes the extensive use of repurposed drugs, focusing on potential benefits and minimizing risks, despite uncertain evidence of their efficacy and safety; the other advocates for a cautious evaluation of outcomes like mortality and clinical improvement, highlighting the importance of transparency and ethical considerations in research. The paper will introduce, discuss, illustrate, and compare the dimensions based on treatment stance.

Keywords: Multi-Dimensional Analysis, Register Variation, Artificial Intelligence

REFERENCES

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation - A Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

The forgotten 33 %. Finland-Swedish literature from a database perspective

Anna Biström
Department of Finnish, Finno-Ugrian and Scandinavian Studies, University of Helsinki, Finland
anna.bistrom@helsinki.fi

Abstract

In the digital era, one could say, that literary history is partly written and shaped in databases and digital archives, via metadata about authors and literature. What is the impact of such resources on the visibility – or invisibility – of authors of literary work? And to what extent is the Finland-Swedish minority literature "forgotten" seen from a database perspective?

In my paper, I will approach these questions with starting point in my research on the amount of database references, which in turn indicates the amounts of secondary literature about the group of Finland-Swedish authors writing in 1830–1930 (Biström 2021). The study builds on data compiled especially from the database Finna (finna.fi), and bibliographies. I have analyzed the amount of database references about authors, with the use of Excel, in comparison with data such as the authors gender and year of birth.

The database or archive on which quantitative studies are based, is however not complete, as has been pointed out by Katherine Bode (2014: 7–25) in her critique of Franco Moretti's claims to accuracy and objectivity. I have approached this issue with the concept of "database visibility" – which represents not only the actual amount of literature about an author, but rather the visibility and accessibility of this literature. In my paper, I will also develop this concept a bit further against the background of my data, as different databases give different perspectives on the visibility of authors.

My ongoing research focuses "invisible authors" – those with no relevant database references in Finna-searches with the authors' name as subject, exploring the question what the Finland-Swedish literary field looks like from the point of view of invisible authors, with a theoretical starting point in the concept of cultural memory (Assmann 2010). Among other things, my data however indicates that these forgotten authors represent – not 99% (Moretti 2013)- but around 33% of all the authors, which supports a point made by Kristina Malmio (2021) about Finland-Swedish literary history and the year of modernist debutants 1916: As my results also indicate, the minority literature (in this case to some extent a privileged one) – is not always forgotten, but on the contrary made more visible due to its importance for the identity of the minority.

Keywords: Finland-Swedish literature, databases, digital archives, cultural memory, minorities

REFERENCES

Assmann, Aleida (2010). “Canon and Archive”. In Astrid Erll & Ansgar Nünning (Eds.), *A Companion to Cultural Memory Studies* (97–107). De Gruyter. https://doi.org/10.1515/9783110207262

Biström, Anna (2021). ”Forskarnas favoriter och det stora outforskade. En grovgenom gång av finländska skönlitterära författare på svenska 1830–1930 och deras synlighet i databasreferenser och sekundärlittertur”. *Samlaren. Tidskrift för svenska och annan nordisk litteratur,* 142, 188–239.

Bode, Katherine (2014). *Reading by Numbers. Recalibrating the Literary Field*, Anthem Press.

Malmio, Kristina (2022). ”99%? En kvantitativ studie av litteratur publicerad på svenska i Finland året 1916”. In J. Bradley (Ed), *Tonavan Laakso: Eine Festschrift für Johanna Laakso*, Central European Uralic Studies 2 (538–567). Praesens Verlag.

Moretti, Franco (2013 [2000]). ”The Slaughterhouse of Literature”, In Franco Moretti, *Distant Reading*. Verso.

A Comparative Corpus-based Discursive News Values Analysis of Liz Truss' and Rishi Sunak's representation in the British Press

Ester Di Silvestro and Marco Venuti
University of Catania, Italy
ester.disilvestro@unict.it - marco.venuti@unict.it

Abstract

Although the number of women in relevant political roles on the international stage is growing, politics still seems a public sphere dominated by men (Liu, 2019). The gender gap in politics is visible and women represent a minority in this field. The representation of female politicians in and out of the media has always been influenced by gender stereotypes. Media can play a significant role in the reiteration of these stereotypes backgrounding other important aspects such as politicians' political agenda (Zamfirache, 2010).

This paper focuses on journalistic discourse concerning Truss' and Sunak's representation in the British press. The main aim of the analysis is to investigate if these politicians are represented similarly or differently through the employment of specific gender stereotypes. Moreover, the analysis aims to understand if their representation as gendered social actors is intertwined with particular news values. The data were collected on LexisNexis and include all the articles mentioning Truss and Sunak in headlines and lead paragraphs in British national newspapers during five specific days (their candidacy and election, and Truss' resignation). We followed Bednarek's and Caple's (2017) Discursive News Values Analysis approach in combination with qualitative (Machin & Mayr, 2023) and quantitative (Partington, Duguid & Taylor, 2013) tools aiming to identify which news values were used and paying particular attention to the gendered representation. Specifically, the quantitative approach (Partington, Duguid and Taylor, 2013) will be carried out through the software Sketch Engine (Kilgarriff *et al*., 2014) that has proven to be a very useful digital resource in the field of linguistics over the years, especially in combination with qualitative approaches allowing a more complete interpretation of data (Tognini-Bonelli, 2010: 17–18; Baker & McEnery, 2015: 2).

The preliminary results of the analysis highlight that, from a general perspective, the selected timespans influence the presence of certain news values. Another significant general trend is for tabloids to convey news values especially through images. Whereas, from a more specific perspective, some news values seem to be connected to specific gender stereotypes (e.g., the news value of personalisation is connected to Truss' role as mother and wife) by both broadsheets and tabloids.

Keywords: CADS, DNVA, gender, Truss, Sunak

REFERENCES

Baker, P. & McEnery, T. (2015). Introduction. In P. Baker & T. McEnery (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora* (pp. 1–19). Palgrave Macmillan. https://doi.org/10.1057/9781137431738_1

Bednarek, M. & Caple, H. (2017). *The Discourse of News Values: How News Organizations Create Newsworthiness*. Oxford University Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on, *Lexicography*, 1(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Liu, S.-J. S. (2019). Cracking Gender Stereotypes? Challenges Women Political Leaders Face, *Political Insight*, 10(1), 12–15. https://doi.org/10.1177/2041905819838147

Machin, D. & Mayr, A. (2023). *How to Do Critical Discourse Analysis. A Multimodal Introduction*. Sage.

Partington A., Duguid A. & Taylor C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies (CADS)*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.55

Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffe and & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14–27). Routledge. https://doi.org/10.4324/9780203856949-3

Zamfirache, I. (2010) Women and Politics – The Glass Ceiling, *Journal of Comparative Research in Anthropology and Sociology*, 1(1), 175–185.

Avoiding ‘generatese’: the optimization of NLG systems through fit-for-purpose data collections

María do Campo Bayón
Pilar Sánchez-Gijón
GELEA2LT, Universitat Autònoma de Barcelona, Spain
pilar.sanchez.gijon@uab.cat
maría.docampo@uab.cat

Paper Abstract

Most linguistic research on the use and exploitation of Natural Language Generation (NLG) systems, whether through graphical interfaces (as in the case of ChatGPT or Gemini) or without them, has primarily focused on their ability to generate text on the basis of prompts. These systems have a wide range of applications, one of which is the interlingual translation of text. They are also able to generate text from a prompt, either in response to a question or a request to perform a linguistic task. Their apparent ability to generate coherent text from another text surpasses the functionalities of any previous linguistic resource.

A translated text often retains certain traces of the source text and language, a phenomenon known as "translationese" (Baker, 1993). With the widespread adoption of machine translation, especially in certain genres, there has been an observable intensification of this phenomenon, which has been termed "post-editese" (Toral, 2019). This can be detected through measurements of specific linguistic aspects and comparisons of human and machine translations using parallel and reference corpora.

Recently, AI systems known as Large Language Models (LLMs) have begun to be used in both professional translation and translator training. The potential footprint such systems leave on translated texts could be called "generatese" (Sánchez-Gijón, forthcoming). The principle of language agnosticism that underlies NLG systems can affect not only the form of discourse (the linguistic features of a text) but also its content (the concepts and ideas it contains and how they are developed) (Sánchez-Gijón, 2022; Imran et al., 2023). This paper aims to study the impact of using small, highly fit-for-purpose data collections to optimize NLG systems by reducing the randomness of their responses and mitigating "generatese". We will explore the creation and, in particular, the description of such data collections, along with their potential for enhancing the quality of translations produced by NLG systems.

REFERENCES

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (233 – 250). John Benjamins Publishing.

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology* 15.4. https://doi.org/10.30935/cedtech/13605

Sánchez-Gijón, P. (2022). Neural machine translation and the indivisibility of culture and language. *FORUM*. Vol. 20. No. 2. John Benjamins Publishing Company, 2022. https://doi.org/10.1075/forum.00025.san

Sánchez-Gijón, P. (2024). Towards characterizing "generatese". American Translation & Interpreting Studies Association (ATISA) Conference. Rutgers U., 5-7 April 2024.

Toral, A. (2019). Post-editese: An exacerbated translationese. *Proceedings of the Machine Translation Summit*, Dublin, Ireland, 19–23 August 2019. https://doi.org/10.48550/arXiv.1907.00900

Exploring Emotions in Parliamentary Debates with a Sentiment Recognition Deep Learning Model: A Case Study of Finnish Plenary Debates on Economy and Environmental Issues 1990-2023

Kimmo Elo[1,2], Otto Tarkka[2], Juuso Laine[2], Jaakko Koljonen[2], Kristian Martiskainen[2] & Markus Korhonen[2]
[1] University of Eastern Finland, Finland
[2] University of Turku, Finland
kimmo.elo@uef.fi (corresponding author)

Paper Abstract

In the past decade, sentiment analysis has increased in popularity among scholars interested in digital communication. It has attracted multi-disciplinary attention as a tool to monitor vast data formulated in the digital sphere to examine e. g. product reviews, tweets and blogs. The digitisation of parliamentary records has introduced the new possible methods to the field of political science as "text as data" have become a more popular approach among political researchers. Recent studies have provided promising results improving our understanding of the dynamics and legislative conflicts in the parliament.

This paper analyses the use of emotions in the plenary speeches of the Finnish Parliament between 1990 and 2023, focusing on two topics, the economy and the environment. In particular, we explore and demonstrate the possibilities of machine learning based sentiment analysis in parliamentary research. We applied a fine-tuned, XLM-R-parla based emotion analysis model on a dataset consisting of 3638 speeches for economic debates and 281 speeches for environmental debates.

The results evidence a certain dominance of two contrary classes – "hopeful-optimistic-trust" on the positive, "fear-worry-distrust" on the negative side – but this finding is rather well in line with results from previous studies. Another interesting finding is, that in both analysed topics there was more variation between negative categories than between positive. Further, we found emotional speech to be more typical for both far-left and far-right parties, as well for parties in the opposition rather than in the government.

Overall, the findings have important methodological contributions for digital parliamentary research and sentiment analysis alike, as they exemplify the applicability and research potential of a deep learning model for large parliamentary speech corpora. Hence, the results offer many points of departure for future research.

Keywords: Sentiment analysis, Plenary debates, Finnish Parliament, XLMR Models, Computational Humanities, Machine Learning

Finland's navigation towards NATO: How is it portrayed in Turkish digital media?

Selcen Erten-Johansson
University of Turku, Finland
seerte@utu.fi

Paper Abstract

Putin's invasion of Ukraine triggered a chain of events prompting Finland to reassess its longstanding policy of military non-alignment. Consequently, Finland made the decision to seek NATO membership. Türkiye, whose NATO membership dates to 1951, largely affected Finland's path towards NATO integration (Kanniainen 2022; Visala & Kajander 2023). This research investigates how Finnish NATO membership is portrayed in Turkish digital media, employing a linguistic framework. The data for this study is collected from web texts of online news reports and internet forum within the Turkish media sources. These texts encompass all references to Finland with specific words such as "Finlandiya" (Finland), "Fin" and "Finli" (Finnish) published between 24 February 2022 when Ukraine was invaded, and 4 April 2023 when Finland officially joined NATO. Sourced from both pro-government and anti-government Turkish digital media, these web texts present a diverse range of perspectives that shape political and public discourse surrounding Finland's NATO membership. The analysis employs Corpus-assisted Discourse Studies (CADS) (Partington et al. 2013), combining quantitative corpus linguistics methods such as keyword analysis and qualitative methods like discourse analysis to examine language patterns, discourse structures, and lexical choices. Initial findings reveal a divergence in content focus of news reports, primarily centred on NATO, and interactive discussions, which extend to broader topics including Finnish culture and language. The divergence between the two contents is further reflected in discourse structures and lexical choices. Moreover, the study identifies potential intersections between political and public discourse, providing a comprehensive perspective on the topic. The research contributes to understanding Finland's recent history and national identity while offering insights into media literacy.

Keywords: Turkish digital media, Finland's NATO membership, web texts, political and public discourse, Corpus-assisted Discourse Studies.

REFERENCES

Kanniainen, V. (2022). Gallup Democracy in Exercising the NATO Membership Option: The Cases of Finland and Sweden. *CESifo Economic Studies 68*(3), 281-296.

Partington, A., Duguid, A. & Taylor, C. (2013). *Patterns and meanings in discourse. Theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins Publishing.

Visala, H. & Kajander, R. (2023 January 30). Turkki on hankala mutta Natolle tärkeä, koska sen avulla voidaan estää laaja sota, sanoo asiantuntija. Yle. Retrieved from https://yle.fi/a/74-20015325 .

Quasi-Parallel Corpora for Less-Resourced Languages: Parallelized Translations of Plato´s *Faidon* in Basque and Finnish.

Koldo Garai
UEF, Finnland
kgarai@uef.fi

EHU, Basqueland
koldo.garai@ehu.eus

Paper Abstract

As at the time Director-General of UNESCO Irina Bokova put it, "Language loss entails an impoverishment of humanity in countless ways. Each language – large or small – captures and organizes reality in a distinctive manner; to lose even one closes off potential discoveries about human cognition and the mind" (Bokova, Irina, 2010). The Foreword by Jordi Solé (Rehm & Way, 2023) also reflects upon languages considered not only as pure communication tools or even as vectors of culture but also as factors of identity; multilingualism is an expression of the identity of Europe.

There are 24 EU official languages, 11 additional official languages, and 54 Regional and Minority Languages (RML), protected by the European Charter for Regional or Minority Languages (ECERML) since 1992, and the Charter of Fundamental Rights of the EU. Some are Indo-European languages and some are non-Indo-European, but both share the task of defining the identity of Europe.

Out of these 90 European languages, more than half have either poor or no technological support; for instance, consider Figure 1 comparing Finnish and Spanish, and keeping in mind that Spanish has half of the technological resources of English (only European context, in both cases), and compare this against Figure 2, technological resources for Finnish, Basque, and Karelian languages (created ad hoc from the web page: *Atlas of the World's Languages in Danger - UNESCO Digital Library*, n.d.).

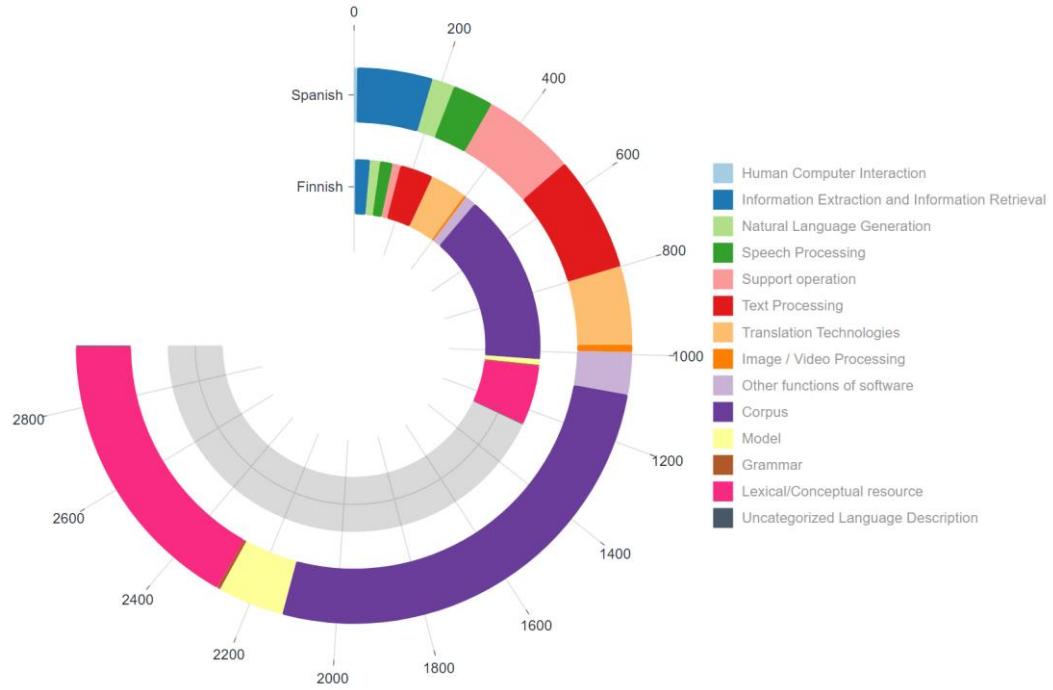

Figure: 1. Technological Factors in Spanish and Finnish

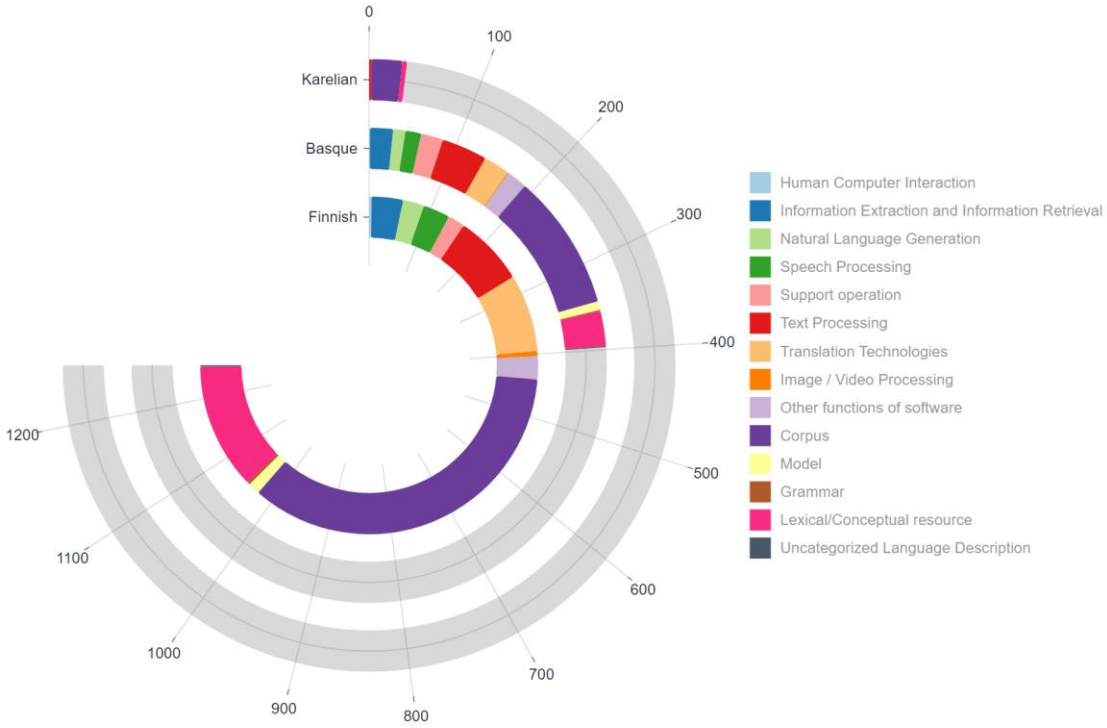

Figure: 2. Technological factors in Finnish, Basque, and Karelian languages

In the spirit of ELE, we present the first aligned Basque-Finnish corpus, both non-Indo-European languages. On the one hand, it is a finished project with the four steps for building a text-aligned corpus, and the description of the procedure can be used as a best practices manual for further prospects (Garai, 2024). On the other hand, it could be seen as a forerunner of a larger desideratum project of building a multilingual aligned corpus comprising all the European non-Indo-European languages to be used for both contrastive linguistic studies and a testbed for shared strategies and approaches to Language Technologies, given some typological convergences such as their postpositional nature or their rich morphology.

Whilst comparable corpora are made of comparable texts following some given criteria, be they from the same language or different languages, parallel refers to translations of a given text (McEnery & Xiao, 2018). Rather, here we coin the term "quasi-parallel" because one is not the direct translation of the other, but both are translations of the same omega text; in this case, Plato's Faidon (Plato, 2006 & Plato, 1978), one translated by Calamnius and the other by Zaitegi. Using already extant translations, and parallelizing them is the cheap path we are proposing for creating linguistic technologies for less-resourced languages.

As a finished study, this work travels through all four stages of building a corpus: (a) from printed text to machine-readable, (b) the standardization of the Basque text to erase graphemic idiosyncrasies to facilitate the next two steps, (c) the alignment, and (d) the automatic annotation following the Universal Dependencies (de Marneffe et al., 2021). The access to the actual outcomes will be shortly available in a repository to be announced.

Keywords: Annotation Universal Dependencies, Less-resourced languages, Parallel corpora, Plato: Finnish-Basque, Text alignment

REFERENCES

Atlas of the world's languages in danger—UNESCO Digital Library. (n.d.). Retrieved May 3, 2022, from https://unesdoc.unesco.org/ark:/48223/pf0000187026

Bokova, Irina. (2010). Preface. In C. Moseley, A. Nicolas, & Unesco (Eds.), *Atlas of the world's languages in danger* (3rd ed., entirely rev., enlarged and updated, p. 4). Unesco Paris.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Garai, K. (2024). *Quasi-Parallel Corpora for Less-Resourced Languages: Parallelized Translations of Plato´s Faidon in Basque and Finnish*. (2024009) [Master's thesis, UEF]. https://erepo.uef.fi/handle/123456789/31151

McEnery, T., & Xiao, R. (2018). *Parallel and Comparable Corpora: What is Happening?* (pp. 18–31). Multilingual Matters. https://doi.org/10.21832/9781853599873-005

Plato. (1978). *Platon. IV., Kriton eta Faidon* (I. Zaitegi Plazaola, Trans.). Euskaltzaindia.

Plato, 427? BCE-347? BCE. (2006). *Faidoni Platonin keskustelma Sokrateen viimeisistä hetkistä jasielun kuolemattomuudesta* (J. W. (Johan W. Calamnius, Trans.). https://www.gutenberg.org/ebooks/19210

Rehm, G., & Way, A. (Eds.). (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer International Publishing. https://doi.org/10.1007/978-3-031-28819-7

NLP-based Topical Analysis and Comparison of "Molokai" by Alan Brennert and "Night Calypso" by Lawrence Scott

Gordana Galić Kakkonen
University of Split
ggalic@ffst.hr

Paper Abstract

This research presents a method for literary analysis that employs automatic topical analysis. It examines the novels "Molokai" by Alan Brennert and "Night Calypso" by Lawrence Scott. By utilizing natural language processing (NLP) techniques, this study delves into and compares how each novel addresses the themes of belief systems, human relationships, health care, and the body. Our approach also allows for a more detailed analysis by subdividing the themes into sub-themes. For instance, human relationships are further divided into those dealing with relatives, friendships or authorities. We use the outputs of the NLP system to analyze the historical, societal, and cultural context of leprosy colonies. We then utilize a postcolonial theoretical perspective to support the analysis. This approach allows us to reveal the insightful ways Brennert and Scott delve into the human condition under extraordinary circumstances. Both novels offer a remarkable account of life within a leprosy colony. Furthermore, they uncover the external factors that shape the communities' response to leprosy. The comparative analysis underscores the relevance of the selected themes as lenses through which the authors explore the multifaceted experiences of their characters, shedding light on public attitudes towards the ill and the marginalized and colonized subjects. This study enhances our understanding of the thematic richness of "Molokai" and "Night Calypso" and highlights the potential of NLP tools to help uncover insights into literary texts. By combining NLP-based topical analysis with the postcolonial theoretical perspective, this study contributes to digital humanities and literary studies, offering a model for future research to explore the complex interplay of themes in literary works.

Keywords: Molokai, Alan Brennert, Night Calypso, Lawrence Scott, natural language processing, automatic topical analysis, postcolonial theory, leprosy

Comparing French and Swedish web registers using multilingual word vectors

Saara Hellström
University of Turku, Finland
sherik@utu.fi

Paper Abstract

The web features a wide variety of registers (Biber, 1988), i.e., situationally defined language use with different purposes (e.g., blogs, news, recipes), in numerous languages. Yet online language use in other languages than English (Biber & Egbert, 2018) remains largely unexplored. Moreover, comparisons across languages are manually conducted which is time-consuming and prone to subjective interpretations. Our study expands web register research to French and Swedish and examines the register characteristics using multilingual word vectors allowing the analysis of registers in one multilingual space without manual comparison. Our research aims at answering the following questions: 1) What kind of keyword groupings does the clustering of the word embeddings reveal? and 2) What (dis)similarities does the clustering of the word embeddings reveal about the languages and registers? Our data consists of the newly established FreCORE and SweCORE corpora including similarly register-annotated web documents. In our analysis, we first extract the keywords, i.e., the statistically overrepresented words indicating what the texts are about (Scott & Tribble, 2006, pp. 55-59), from the corpora using text dispersion keyness (Egbert & Biber, 2019) to get the language specific characteristics for the registers. Then, using the fastText tools, we transform the keywords into word vectors, i.e., linguistically motivated, numerical representations of words derived from a language model. The word vectors present words in one multilingual space where semantically similar words are represented by similar vectors even across languages. Finally, to examine the cross-lingual similarities of the keywords and what they tell about the registers, we cluster the word vectors with KMeans. Nineteen clusters offer the best fit to the data. Our analysis shows that the clusters group keywords based on their topical or grammatical features: e.g., the cluster POLITICS/POWER (topic) includes pouvoir – makt (power; authority), people – folket (people) while the cluster STANCE (grammar) features pense – tänker (thinks), vrai – sant (true). Moreover, the keywords in each cluster tend to belong to certain dominant registers, and these prominent registers and clusters are often the same in both French and Swedish. The keywords within a register group coherently which suggests that clustering could be a viable method to group keywords computationally instead of the laborious manual grouping. These findings suggest that there are more cross-linguistic similarities than dissimilarities between the French and Swedish web registers.

Keywords: web register, keywords, multilingual word vectors, clustering

REFERENCES

Biber, D. (1988). Variation Across Speech and Writing. Cambridge University Press.

Biber, D. & Egbert, J. (2018). Register Variation Online. Cambridge University Press.

Egbert, J. & Biber, D. (2019). Incorporating text dispersion into keyword analyses. Corpora, 14(1), 77–104.

Scott, M. & Tribble, C. (2006). Textual Patterns: Key Words and Corpus Analysis in Language Education. John Benjamins Publishing Company.

Using deep learning to examine cross-linguistic similarities of registers

Erik Henriksson, Amanda Myntti, Veronika Laippala
University of Turku, Finland
erik.henriksson@utu.fi

Paper Abstract

This study examines the use of multilingual deep learning to analyze cross-linguistic similarities of registers – situationally defined text varieties such as news or reviews (Biber 1988). Register studies have repeatedly shown that differences in the situational context of a text are reflected in its linguistic characteristics. However, little is known about register variation across languages (see, however, Biber 2014; Li et al. 2023). One of the reasons for this is the lack of methods enabling the analysis of registers without the manual interpretation of register characteristics in each language. In this study, we apply multilingual deep learning to fill this gap. We examine cross-linguistic similarities of registers using the deep learning model XLM-R (Conneau et al. 2020). Specifically, we target eight registers and eight languages: English, French, Swedish, Finnish, Turkish, Urdu, Chinese, and Farsi. First, using the multilingual CORE corpora (Laippala et al. 2022) and XLM-R, we train a multilingual register identification model. The model learns to classify documents to register classes and creates document vectors that represent the documents in one multilingual vector space. This allows us to examine registers and their similarities across languages by calculating document similarities in the vector space. Second, we extract keywords for the registers using the trained model and the model explanation method SACX (Rönnqvist et al. 2022). This enables the analysis of the linguistic motivation behind the learnt model. We group the keywords using semantic and grammatical criteria and analyze the registers and their similarities across languages based on these groupings. Furthermore, we compare our findings to previous studies based on more frequently applied statistical methods, such as multi-dimensional analysis. Preliminary results show that the model learns to identify the registers at a nearly human-level performance. In the vector space, the documents are structured to language-independent and register-specific groupings. This shows that the model has learnt language-independent representations of the registers. Furthermore, the analysis of the keywords shows that the learning is based on linguistically motivated features. For instance, the keywords feature semantic properties such as stance and functional features such as reporting verbs that characterize registers across languages – and have been identified as register characteristics in previous studies focusing on individual languages. Thus, our findings support the existence of register universals (Biber 2014) and encourage the use of multilingual deep learning for cross-linguistic corpus analyses.

Keywords: multilingual machine learning, web-as-corpus, web registers, register universals, keyword extraction

REFERENCES

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D. (2013). In Gray, Bethany. (2013). Interview with Douglas Biber. Journal of English Linguistics 41(4), 359–379. https://doi.org/10.1177/0075424213502237.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. Languages in Contrast, 14(1), 7-34. https://doi.org/10.1075/lic.14.1.02bib

Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440-8451. Association for Computational Linguistics.

Laippala, V., Salmela, A., Rönnqvist, S., Aji, A. F., Chang, L.-H., Dhifallah, A., Goulart, L., Kortelainen, H., Pàmies, M., Prina Dutra, D., Skantsi, V., Sutawika, L., & Pyysalo, S. (2022). Towards better structured and less noisy Web data: Oscar with Register annotations. In Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), 215-221. Association for Computational Linguistics.

Li, H., Dunn, J. & Nini, A. (2022). Register variation remains stable across 60 languages. Corpus Linguistics and Linguistic Theory, 19(3), 397-426. https://doi.org/10.1515/cllt-2021-0090

Rönnqvist, S., Kyröläinen, A.-J., Myntti, A., & Laippala, V. (2022). Explaining Classes through Stable Word Attributions. In Findings of the Association for Computational Linguistics, 1063-1074. Association for Computational Linguistics.

Corpus-assisted critical discourse analysis on LGBTQ+ segregation and internal migration in Finland

Jarmo Harri Jantunen
University of Jyväskylä, Finland
jarmo.h.jantunen@jyu.fi

Abstract

Although an inclusive city is generally the normative framework for urban development, opportunities for inclusion are not the same for everyone. This presentation focuses on urban places and meaning-making of these places for Finns whose gender and sexual identity are considered non-normative. Furthermore, the discourses of inclusion and exclusion and how these have influenced people's willingness to stay in certain places or move away are discussed (see e.g., Gerhards 2010; Poston et al. 2017). These experiences are linked to LGBTQ+ segregation, LGBTQ+ ghettos (see e.g., Aldrich 2004), and the stigmatisation of certain places; identity issues and fear of discrimination can lead to self-segregation and the choice of environments that offer social protection, which in turn can lead to the segregation of those areas (Ghaziani 2014).

The presentation takes a critical approach to the experiences and meaning-making associated with urban places, and the notions of inclusion, equality, and rights that they evoke. The research is based on two datasets; the first from a survey conducted via the Webropol application (521 responses), and the second from the Suomi24 Corpus (City Digital Group, 2021), which is a collection of posts from the 'Finland24' discussion forum. The Webropol survey asked participants about their experiences of the places where they have lived as a child and an adolescent, reasons for settling in their current location, possible reasons for internal migration, and experiences that have affected the quality of life. The Suomi24 Corpus was chosen to represent "general" opinion, as the forum ranks among one of the most popular for discussions among the general public in Finland.

The data were analysed using corpus-assisted critical discourse analysis, i.e. a combination of quantitative corpus analysis (collocation analysis and grouping of collocations into semantic categories) and qualitative critical discourse analysis (close reading of concordance lines) methods. The data shows there is a clear trend for internal migration towards large urban areas and growth centres; one of the motivating factors for this being the sense of inclusion in a community that supports identity. The creation of such places is also carried out by members of minorities themselves through self-segregation and strategies to avoid stigmatised places. In turn, the Suomi24 data shows that the internal migration of LGBTQ+ is seen as problematic in the "general" discussion (mostly among non-LGBTQ+ people), as certain urban areas, such as Helsinki, are seen to be adversely overcrowded with people from minority groups.

Keywords: CACDA, inclusion, internal migration, LGBTQ+ segregation, urban areas

REFERENCES

Aldrich, R. (2004). Homosexuality and the City: An Historical Overview. *Urban Studies*, *41*(9), 1719–1737. http://www.jstor.org/stable/43201476

City Digital Group (2021). The Suomi24 Sentences Corpus 2001–2020. *The Language Bank of Finland.* http://urn.fi/urn:nbn:fi:lb-2021101526

Gerhards, J. (2010). Non-Discrimination towards Homosexuality: The European Union's Policy and Citizens' Attitudes towards Homosexuality in 27 European Countries. *International Sociology 25*(1). https://doi.org/10.1177/0268580909346704

Ghaziani, A. (2014). *There Goes the Gayborhood?* Princeton University Press. https://doi.org/10.1515/9781400850174

Poston, D. L., Compton, D. R., Xiong, Q., & Knox, E. A. (2017). The Residential Segregation of Same-Sex Households from Different-Sex Households in Metropolitan USA, circa-2010. *Population Review 56*(2). https://dx.doi.org/10.1353/prv.2017.0005.

Sentiment analysis for detecting suicidal youths' positive and negative encounters with public service providers

Pyry Kantanen
University of Turku, Finland
pyry.kantanen@utu.fi

Kati Kataja
Tampere University, Finland
kati.kataja@tuni.fi

Leo Lahti
University of Turku, Finland
leo.lahti@utu.fi

Paper Abstract

Despite a long-term downwards trend in suicide rates, Finland still experiences a higher prevalence of suicide compared to many other EU countries (Eurostat, 2024). Investigating the life courses of young people who have either attempted to or have committed suicide can yield new knowledge that can be used to influence policies and mitigate the effect of negative life trajectories or completely prevent them. Employing a broader framework of social autopsy, we examine the social and political conditions surrounding suicide incidents.

Our methodology involves 20 semi-structured interviews with young adults who have attempted suicide and family members of young persons who have committed suicide. As the data involves a wide spectrum of emotional and valuated expressions, we apply the novel method of automated sentiment analysis to detect youths' encounters with service providers and classify them as positive, negative, and neutral. Sentiment analysis has widely been used in classifying product reviews and customer feedback, but more recently also for more varied tasks, such as detecting possible markers mental states such as depression or anxiety in social media posts (Tana, Shcherbakov & Espinosa-Leal, 2022). In our work, this method complements more traditional qualitative methods of close reading, thematic analysis and narrative analysis. Given the sensitive nature of our datasets, it is critical that machine learning models can be run locally in a secure environment. As the interviews were conducted in Finnish, we utilise the FinBERT model fine-tuned with FinnSentiment dataset. FinnSentiment consists of 27,000 polarity-annotated sentences drawn from the Suomi24 social media corpus (Lindén, Jauhiainen & Hardwick, 2023).

In this presentation, we will discuss the pilot results that demonstrate how sentiment analysis can complement the methodological toolkit of social autopsy. We will highlight the advantages and constraints of utilizing a machine learning model trained on social media corpus for the purpose of analysis of interview data. Our work is a part of the Young Despair research project that, in addition to qualitative interviews, utilize register data and official records including police reports, coroner's autopsies, and forensic toxicology findings in studying young peoples' suicidality. All these concerted efforts aim to enhance our understanding of the multifaceted determinants contributing to youths' suicidal

behaviors, thereby informing targeted intervention strategies and policy initiatives aimed at suicide prevention.

Keywords: suicide, young people, sentiment analysis, machine learning, research interviews, sensitive data.

REFERENCES

Eurostat (2024). *Death due to suicide, by sex* [Data set]. Retrieved August 14, 2024, from https://doi.org/10.2908/TPS00122

Linden, K., Jauhiainen, T., & Hardwick, S. (2023). FinnSentiment: A Finnish Social Media Corpus for Sentiment Polarity Annotation. *Language Resources and Evaluation, 57*(2), 581-609. https://doi.org/10.1007/s10579-023-09644-5

Tana, J., Shcherbakov, A., & Espinosa-Leal, L. (2022). Sentiment analysis of depression related discussions in the Suomi24 discussion forum. *Informaatiotutkimus*, 41(2–3), 157-162. https://doi.org/10.23978/inf.122663

Acknowledgements/Disclaimers

This research was funded by the Strategic Research Council (SRC) established within the Research Council of Finland, grant numbers 352600, 352604 and 352602.

Manual data collection & qualitative analysis for social media data – "luddite" meme researcher insecurities in the age of AI

Reeta Karjalainen
University of Jyväskylä, Finland
reeta.a.karjalainen@jyu.fi

Paper Abstract

Digital data and digital research methods have undoubtedly opened a whole new world for researchers. However, they have also raised fears and insecurities about research ethics and researcher's role. For some, AI and other innovative digital tools seem too scary to learn or take control of. This paper focuses these personal and professional insecurities about digital, automated research methods as researcher on memes who utilizes manual data collection and qualitative analysis in the age of AI, alongside with presenting the overall PhD project about representations of mental health and mental illness in Internet memes. The data of this paper consists of approximately 900 manually collected mental health themed memes collected under two social media and Internet platforms: Instagram and Imgur.com. On Instagram, the data is collected manually by taking screenshots of memes under hashtags #mentalhealthmeme and #mentalillnessmemes, as well as three accounts focusing on mental health related themes. On Imgur.com, the data is detected and collected on the Most Viral page, especially under larger image cluster posts called Meme Dumps. The data is collected gradually throughout the research process, with more intensive data collection periods implemented. During the time of writing, the latest more intensive data collection phase was conducted in January 2024. Moreover, data collection method called reverse snowball method (Särmä, 2014: 99) is utilized in during the whole research process. The aim of this study is to examine different semiotic ways these memes represent mental health issues by analysing the data with multimodal critical discourse analysis. In this paper I present the insecurities about utilizing digital data collection and analysis tools in meme research, as well as the arguments for not utilizing them in relation to the overall PhD project. Fears of not learning AI or other data scraping tools include being left behind of digital development, cherry-picking research data, and being labelled as a luddite not willing or not being able to learn new methods. However, arguments for more manual, researcher-oriented methods in collecting social media data, will also be discussed. Overall, the aim of this paper is to seek encouragement from peers, as well as new ideas and support in utilizing digital research methods, and most importantly, how to connect those methods to qualitative and critical research.

REFERENCES

Särmä, S. (2014). Junk feminism and nuclear wannabes: Collaging parodies of Iran and North Korea. Tampere: Tampere University Press.Cambridge, MA: MIT Press.

Practical solutions for digitally administering and scoring of a children's speechreading test

Jaakko Kauramäki[1,2], Satu Saalasti[1,3], Kerttu Huttunen[1]
[1]*University of Oulu, Finland*
[2]*University of Helsinki, Finland*
[3]*University of Eastern Finland, Finland*
(jaakko.kauramaki@helsinki.fi)

Paper and Poster Abstract

Use of visual information about speech is pronounced in situations in which auditory information is degraded because of, for example, background noise or reverberation. In speechreading (often also called lip reading), information about lip, jaw and tongue movements but also the visual cues of facial expressions are used to perceive the message of a speaker. People with hearing loss try to make use of speechreading for complementing the insufficient auditory information caused by their hearing problem, but interindividual differences are large and reliable assessment methods are needed.

Aims of the research project *Gaze on lips?* (https://www.oulu.fi/en/projects/gaze-lips) are twofold: to construct, standardize and validate the pre-recorded Speechreading Test for Finnish Children (SPETFIC; Huttunen & Saalasti, 2023) with automatized scoring and to find out the developmental trajectory of primary school-age children's speechreading skills. Data are currently being compiled from 8- to 11-year-old hearing children to have the age norms for the SPETFIC.

In the current presentation, we report the practical solutions of administering SPETFIC both in-person and remotely. Remotely collected samples are managed by REDCap tools hosted at the University of Oulu, Finland. REDCap (Research Electronic Data Capture; Harris et al. 2009; 2019) is a secure, web-based software platform designed to support data collection. We implemented remote testing by utilizing screen sharing of a Zoom meeting (Zoom, 2024) so that SPETFIC is run on a test administrator's computer. For testing the stability and speed of the Internet connection and the capabilities of screen sharing of Zoom video call, a specific frame drop estimation test was constructed. If there were issues causing excessive frame dropping, a one-time direct access link to the REDCap running SPETFIC was conveyed to the child via chat channel of the Zoom.

SPETFIC includes the automatic scoring of the results, shown both as total and section specific score on screen after finishing the test. Additionally, the item-by-item and summary results of the test can be downloaded as comma-separated (CSV) files. After the validation phase, speech and language therapists testing the children at clinics can choose to separately administer either the section A (easier words), the section B (more difficult words) or the sentence section. Having these sections enables tapping of a fairly wide skill spectrum and following up of skill improvement along the child's maturation and

intervention. Automatic scoring rules out scoring errors in both research and clinical use of the test.

Keywords: Finnish language, lipreading, online testing, speechreading, visual speech processing.

REFERENCES

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

Huttunen, K., & Saalasti, S. (2023). *Lasten huuliolukutesti (Speechreading Test for Finnish Children)* [Unpublished; test in validation phase].

Zoom. (2024). Zoom (v6.1.0) [Software]. Retrieved from https://zoom.us/

Gendered recruiting in social media: a case study in network marketing

Daria Kosinova
Institution, country
email@email.com

Paper Abstract

My doctoral dissertation explores the novel phenomenon of recruiting new workers on social media platforms such as Instagram. In particular, I focus on analyzing those personal narratives that recruiters share on social media in order to attract potential workers and build communities. Being interested in 'life writing' practices, I study self-presentation and authoring of one's own story from a new angle - as a social recruiting practice in the direct selling and network marketing industry. Social media offer recruiters the chance to tap into the passive candidates market to approach new talents and those who are not actively seeking new job opportunities. To recruit new people, direct selling professionals build long-term relationships with the audiences that function as affectively engaging communities. My theoretical starting point for approaching social media recruiting is the affect theory. The theory offers an alternative understanding of the 'life writing' practices that focus on a moment of recognition of the need for change in one's work-life relationship. I explore how affects and emotions are evoked, triggered, and build up in social media practices and what insights the theory of affect brings on hiring in the digital age and recruiting passive candidates. I examine a contemporary form of the direct selling industry as a form of the gig economy and female direct selling professionals as independent contractors. The gig economy is a part of a larger transformation of the economy, where all kinds of online platforms take center stage. My research widens the understanding of the gig economy and brings a more gendered perspective to the debates by focusing on a women-centered industry and its practices. In my empirical study, I develop the concept of gendered social recruiting. The methodological approach of this research combines data from digital ethnography and qualitative semi-structured interviews with direct sellers from three EMEA region countries (Finland, the United Kingdom, and South Africa). I focus on countries that belong to the same region inside the company and according to my observations share similar social recruiting practices while also open up tensions when get into comparison. It allows studying the international societal phenomenon that is important both in a global and national context.

Keywords: Digital labour, gender, affect, social media, digital ethnography, gendered social recruiting

AI Literacy for Study and Working Life – University Students' Experiences from the Pilot Course

Päivi Kousa
University of Jyväskylä, Finland
paivi.m.kousa@jyu.fi

Jenny Tarvainen
University of Jyväskylä, Finland
jenny.h.tarvainen@jyu.fi

Presentation Abstract

This presentation introduces a pilot course "AI Literacy for study and working life" at the University of Jyväskylä (JYU), Centre for Multilingual Academic Communication (Movi). The voluntary course was held in Spring, 2024 and it was for all students in JYU. The course covered topics such as AI literacy and ethics, AI in research and writing process, and machine translation and AI in language learning process. In this study, we focus on AI ethics and the students' perceptions of how they think AI literacy skills could benefit them in their studies and in their future working life. The topic is studied through the following research questions:

1. What are students' opinions about AI and AI Ethics a) at the beginning of the course b) at the end of the course?

2. How do the students see AI literacy skills in their a) studies b) in their future working life?

The data was collected through the course's online platform, Howspace. There was also an ethnographical touch since the researchers have been planning and teaching the course alongside other teachers. The current data (N = 38) can be considered small, but it is the first part of a longitudinal study. The data consists of students' online discussions and mind maps, reports, reflective written task, and teachers' notes. During the content analysis, the data is first coded into smaller units, which are then combined into categories and finally, to larger thematic areas. The aim is to find not only similarities but also differences between students and their perceptions. The content analysis is carried out by two researchers in several phases using atlas.ti software.

The results are expected to give new insights on university students' perceptions of AI ethics and AI literacy since there are not so many empirical studies in the field (Laupichler, *et al*., 2022). The aim is to increase general understanding of how university teaching and teacher education should be developed to promote students' AI-related future skills for working life (e.g., Dignum, 2021). On a larger scale, the results bring added value to the research of this topic and stimulate discussion about the development of higher education AI-pedagogy and how to keep up to date in a rapidly changing working life.

Keywords: AI Ethics, AI Literacy, AI Pedagogy, Future Skills, Higher Education

REFERENCES

Dignum, V. (2021). The role and challenges of education for responsible AI. *London review of education. 19*(1), 1-11. DOI:10.14324/LRE.19.1.01

Laupichler, M., Aster, A., Schirch, J., & Raubach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence 3*(1). DOI:10.1016/j.caeai.2022.100101

Ambiguous grammatical forms and power relations. A statistical analysis of Latvian corpora

Sergei Kruk
Riga Strandis University
sergei.kruk@gmail.com

Paper Abstract

Qualitative and quantitative analysis of Latvian laws and policy documents has identified grammatical forms that make texts ambiguous (Kruk 2021; 2024). The reason is that some grammatical forms combine the characteristics of different parts of speech. It is up to the receiver to decide whether a particular word is to be interpreted as a noun, verb, adjective, or adverb. Meanwhile, the sender reserves the possibility to reject the interpretation as wrong. Ambiguity therefore can maintain power relations since it permits the sender to bear no responsibility for the content. The paper seeks a statistical foundation for such a claim in corpora of the Latvian language. If ambiguous grammatical forms (AGF) are helpful in power relations, then they must be less frequent in those specialized corpora and subcorpora whose primary topic is not the maintenance of power relations. The frequencies of four AGFs (nominalizations, indeclinable participles, cascades of genitive and subjunctive mood) were counted in 19 specialised sub/corpora using the Sketch Engine programme. The extracted data were normalised per million words and processed in SPSS. The two-step cluster analysis identified four clusters. The lowest incidence of AGF was found in fiction, poetry, literary and children's magazines; the highest was found in legal documents and Ph.D. dissertations. Linear regression has found a strong dependence of indeclinable participles on nominalizations. Qualitative analysis of randomly selected fragments reveals that the overuse of AGF by legislators and doctoral students weakens the illocutive force of propositions. Creating an impression of 'serious' discourse, the authors, in fact, conceal the sense of propositions. They can blame critical readers for their inability to find the appropriate information in the text and for the lack of adequate expertise to grasp the complex language structures of the law or academic discipline. The paper discusses the applicability of corpus-driven approach in critical discourse analysis.

Keywords: Corpus linguistics, corpus statistics, corpus driven research, Latvian language, ambiguity in language

REFERENCES

Kruk, S. (2020). Uzticības, sadarbības un vienotības konceptu izpratne Nacionālajā attīstības plānā 2021.–2027. gadam. Akadēmiskā Dzīve 56: 131–147.

Kruk, S. (2024). Ambiguous grammars of legal discourse. Lettonica 53.

Predicting Child Language Outcomes Across Diverse Longitudinal Cohorts: A Machine Learning Approach

Marja Laasonen*, Rosa González Hautamäki, Federico Màlato, Jade Plym, Sini Smolander, Eva Arkkila, Pekka Lahti-Nuuttila, Sari Kunnari, Penny Levickis, Cristina McKean, & Patricia Eadie
* *University of Eastern Finland, Finland*
marja.laasonen@uef.fi

Paper Abstract

Introduction: Developmental language disorder (DLD) is a highly prevalent condition that significantly impacts children's language development, academic success, and overall well-being. Despite extensive research efforts, we do not fully understand the factors influencing typical language development and the emergence of DLD. This study employs machine learning techniques to predict and examine the generalisability of child language outcomes across two diverse longitudinal cohorts: the Helsinki Longitudinal Specific Language Impairment Study (HelSLI) in Finland and the Early Language in Victoria Study (ELVS) in Australia.

Specifically, we asked: (1) Can group membership (typical development, TD vs. low language outcome, LLO) be predicted using a comprehensive set of cognitive/neuropsychological variables, including linguistic measures, in the ELVS dataset? (2) Which cognitive/neuropsychological and linguistic variables are the most informative predictors of group membership in the ELVS dataset? (3) How do the predictive models and critical variables identified in the HelSLI study compare to those from the ELVS dataset, and what does this reveal about the generalisability of findings across populations?

Methods: We employed an ensemble machine learning approach for two-class classification problems, specifically Random Forest (RF). RF provides an accurate model that explores insights into which variables contribute most toward predictive accuracy.

Results: The previous cross-sectional HelSLI studies showed that using a machine learning approach, neuropsychological and linguistic variables could accurately (85–90 %) classify TD and DLD in 3–7-year-old monolingual and sequentially bilingual children. Among the neuropsychological set, language and verbal memory were most important in classifying monolingual and bilingual children. In the linguistic set, the best classifiers differed between the language groups, with variables corresponding to language production being superior in classifying monolingual groups and language comprehension in bilingual groups. The results with the ELVS dataset indicate how similar variables contribute to the prediction of group membership.

Discussion: The findings of this study have important implications for our understanding of the complex, multifactorial nature of DLD and the development of more effective, culturally sensitive approaches to early identification and intervention. By comparing the predictive models and key variables across the HelSLI and ELVS datasets, we aim to

identify universal and language-specific factors that shape language development and contribute to the emergence of language disorders.

Keywords: cognitive/neuropsychological factors, cross-linguistic, developmental language disorder, low language outcome, machine learning

Towards Open Source Ecosystem for European Music Data

Leo Lahti, Pyry Kantanen, Akewak Jeba

University of Turku
leo.lahti@utu.fi

pitkant@utu.fi

Paper Abstract

The Open Music Europe project aims to reshape the European music data landscape by identifying data sources, developing data collection methods, and crafting policy-relevant indicators to underscore significance of data. The core scientific focus of the project is on enhancing data interoperability and accessibility through the integration of best practices in data science into an open source software ecosystem. The project pioneers best practices in data science and integrates them into an accessible open source software ecosystem that enables non-specialist stakeholders to gather and utilize data from multiple sources effectively. This software ecosystem, which includes a set of open source components and interactive cloud-based applications, has been implemented and is actively maintained. We demonstrate the use of these tools, and in particular the use of the eurostat R package in data retrieval and analysis. We show how users can add metadata by utilizing special data containers where additional metadata contents can be obtained from the Eurostat SDMX API. The framework supports conversions to various linked data standards and formats, greatly facilitating interoperability between data standards and openly available methodology and advancing data provenance, data citations, and reproducible research. Analysis of the European music industry complements the ongoing research efforts focusing of other forms of cultural production in the field of computational humanities. Our work demonstrates in particular how interoperability across data standards can significantly contribute to the advancement of FAIR and open data initiatives, helping to ensure more open sharing and utilization of music data in academic research as well as more broadly in society.

Keywords: Computational Humanities, Open Music, Open Source Ecosystem

Do we use bad language more with friends or with acquaintances? Evidence of "fuck" from societal big data

Mikko Laitinen, Paula Rautionaho, Masoud Fatemi & Mikko Halonen
University of Eastern Finland
mikko.laitinen@uef.fi

Paper Abstract

This presentations focuses on the uses of "fuck" in digital social networks from social media. Social media outlets have so far been predominantly treated as massive text collections, but they can be effectively used to investigate the role of social networks in shaping human communication. We use user-generated texts from 5,660 social networks (with 435,345 users and 7.8 billion words) from three settings (UK, US, and Australia). With embedded network information, this massive dataset enables us to investigate how network properties, that of the size and the strength of the network, influence the use of offensive words in these three settings. Our findings show that the use of "fuck" is nearly 20 times more frequent in social media material than in texts in traditional text corpora. The observations also show that Americans use "fuck" most frequently, while Australians least frequently but they are highly creative with spelling variants of the word. We find that people on social media swear more with acquaintances than with friends, but only in smaller networks - in larger networks of >100 people, the differences level out. Overall, this study highlights the benefits of using social media data that can be enriched to allow access to the social networks that people interact in.

The Trust Divide: Chatbots' Superior Performance and Skeptical Students

Martin Laun[1], **Katharina Hirt**[2], **Eva L. Wyss**[2], **Fabian Wolff**[1]

[1]*Bielefeld University, Germany*
[2]*University of Koblenz, Germany*
martin.laun@uni-bielefeld.de

Paper Abstract

Chatbots have emerged as fundamental components of innovative teaching approaches to address the significant challenge of catering to students' diverse learning backgrounds with limited resources (Lazarides & Chevalère, 2021; Zhang & Aslan, 2021). However, their usefulness depends on their ability to provide high-quality, reliable outcomes and create trustworthiness (Kaplan et al., 2021). Against this background, the present study investigated the trustworthiness of chatbots in supporting educational tasks and the extent to which this trustworthiness is justified. The study involved 189 students from vocational nursing schools who created medical care plans with and without ChatGPT's (Model GPT-4) support. Additionally, ChatGPT was asked to solve the tasks without students being involved. Experts then evaluated all three sets of plans. This allowed us to compare the quality of the answers solved by the chatbot to the solutions by the students. To examine the trustworthiness of the chatbot, another independent care plan was evaluated by the students following experimentally manipulated feedback that it was created either by a human or a chatbot. Statistical analyses revealed that students' beliefs about the source of the care plan -explicitly informed as either chatbot-generated or human-created- significantly affected their perceptions of trustworthiness. Students who believed a chatbot solved the task perceived it as less trustworthy than those who thought a human created the plan. However, analyses of the expert ratings revealed that the solutions created by the chatbot were actually of higher quality than those created by humans. Therefore, the mistrust in the chatbots was not justified. This finding suggests that students underestimate chatbot performance in educational settings.

Keywords: chatbots, education, trustworthiness, quality, human computer interaction

REFERENCES

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors*, *65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Lazarides, R., & Chevalère, J. (2021). Artificial intelligence and education: Addressing the variability in learners' emotion and motivation with adaptive teaching assistants. *Bildung Und Erziehung*, *74*(3), 264–279. https://doi.org/10.13109/buer.2021.74.3.264

Zhang, K., & Aslan, A. (2021). AI technologies for education: Recent research & future directions. *Computers and Education. Artificial Intelligence*, *2*, 100025. https://doi.org/10.1016/j.caeai.2021.100025

Acknowledgements/Disclaimers

Our research was funded by Grant 16DHBKI039 from the German Federal Ministry of Education and Research (BMBF).

Data-rich History for 19th Century Literature in Finland

Kati Launis
University of Eastern Finland, Finland
kati.launis@uef.fi
Aino Mäkikalli
University of Turku, Finland
ainmak@utu.fi
Viola Parente-Čapková
University of Turku, Finland
viocap@utu.fi
Veli-Matti Pynttäri
University of Eastern Finland, Finland
veli-matti.pynttari@uef.fi
Osma Suominen
National Library of Finland, Finland
osma.suominen@helsinki.fi

Abstract

Consortium *Digital History for Literature in Finland* (Research Council of Finland, 2022–26)[1] focuses on renewing current understanding of the literary history of the long 19th century – which, in Finland, covers the period of autonomy, 1809–1917 – by combining detailed historical analysis with a dedicated bibliographic data science framework, presented by Julia Matveeva, Osma Suominen and Leo Lahti in their abstract. Collaboration between researchers from the universities of Turku and Eastern Finland and the National Library of Finland makes it possible, for the first time, to build a data rich view of the history of Finland's publishing landscape. We will systematically mine the bibliographic metadata for all available published fictional works that are included in the National Bibliography Fennica and complement this by close reading of the Finnish and Swedish language literary texts in either digital or traditional paper format.

In this presentation, we are focusing on the research done in the WP1 (Literary history in the 19th century Finland, UEF) and WP 3 (Digital resources, The National Library of Finland). Based on the curated corpus – which is currently under development – of about 2800 first editions of the published fictional works that are included in the National Bibliography Fennica, we ask, what kind of fiction was published in Finland during the 19th century. What kind of changes and patterns rise above others? Which genres can be considered as prolific literary forms in the literary scene? The titles are of particular interest as *paratexts* (Genette 1987/1997), as they contain a wealth of genre terms: what do the titles tell us about the way contemporaries structured the genre?

This multidisciplinary project is framed in the field of digital humanities. It stems from the first wave of such initiatives in Finland; elsewhere, such digital or data rich literary

[1] https://sites.utu.fi/digital-history-literature-finland/en/frontpage/

history has been developed e. g. by Franco Moretti (2000/2013) and Katherine Bode (2018). The method used in the project is a combination of *bibliographic data science* (Lahti & al 2019), *data-rich literary history* (Bode 2018) as well as *meso-analysis* (Saint-Amour 2019, Parente-Capková, Launis & Westerlund 2023) or *resourceful reading* (Bode ja Dixon 2010), which combine the distant and close reading of literature. We use both new-empirical and eResearch techniques and put emphasis on empirical bibliographical study (Bode & Dixon 2010); our approach is defined by the use of digital methods and expertise in data science, bibliographic and library knowledge, and a solid knowledge of the literary history and the 19th century context.

Keywords: bibliographic data science, data-rich literary history, Finnish literature, meso-analysis, National Bibliography Fennica.

REFERENCES

Bode, K. (2018). *World of Fiction. Digital Collections and the Future of Literary History.* University of Michigan Press.

Bode, K. & Dixon, B. (2010). Resourceful Reading: A New Empiricism in the Digital Age. In K. Bode & R. Dixon (Eds.), *Resourceful Reading: The New Empiricism, EResearch and Australian Literary Culture*, 1–27. Sydney University Press.

Genette, G. (1987/1997). *Paratexts. Tresholds of Interpretation*. Transl. Jane E. Lewin. Cambridge University Press.

Lahti L., Marjanen J., Roivainen H. & Tolonen, M. (2019). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly* 57(1), 5–23. http://dx.doi.org/10.1080/01639374.2018.1543747

Moretti, F. (2000/2013). *Distant Reading*. Verso.

Parente-Čapková, V., Launis, K. & Westerlund, J. (2023). Digitaaliset metadata-arkistot ja mesoanalyysi kulttuurisen vaihdon kartoittamisessa. *Avain - Kirjallisuudentutkimuksen aikakauslehti*, vol 20, nro 1, 100-111. https://journal.fi/avain/article/view/127486/77833.

Saint-Amour, P. (2019). The Medial Humanities: Toward a Manifesto for Meso-Analysis. *M/m,* vol. 3, cycle 4. https://doi.org/10.26597/mod.0092.

Imaginaries of Ownership and Sustainability: A Corpus-Assisted study

Ella Lillqvist
University of Eastern Finland & University of Vaasa, Finland
ella.lillqvist@uef.fi

Paper Abstract

In recent years, public discussion on economic change or, rather, a deliberate transformation of the economy, has become increasingly common. This study examines the idea on ownership in the context of these discourses of economic change. In particular, the study focuses on how the role of ownership of material property is shaped in relation to "futures of sustainability" (Adloff & Neckel, 2019; Degens, 2021) and other imaginaries of future good life.

The corpus of this study was formed by retrieving both news texts and social media texts from a large collection of Finnish online textual materials aggregated by the company Legentic. The search term used was *omistaminen* ('ownership') and it was also required that the text mentioned a word referring to economic change, namely *kiertotalous* ('circular economy'), *jakamistalous* ('sharing economy'), or *alustatalous* ('platform economy'). The search terms were treated as lemmas. The final corpus totals approximately 426,000 running words and spans the years 2015-2023.

This study takes a corpus-assisted discourse studies approach (e.g. Ancarno, 2020; Mautner, 2016), analysing keywords, as well as n-grams and collocates related to ownership and the new economy concepts.

In this way, this study sheds light on, among other things, who we talk about as owners in Finland, what are essential things to own, and how the social meaning of ownership is described. The focus is on what kind of change is being talked about in the context of these dimensions of meaning and how change is linked to sustainability. There is a broad consensus in the data that ownership is and will become less desirable and will be replaced by services, sharing, and other forms of co-use. In the corpus, the importance of ownership for sustainability is seen narrowly, as mainly related to the ownership of goods, whereby real estate, land and financial property, as well as new forms of property such as carbon quota are excluded.

Keywords: corpus-assisted discourse studies; discourse; future; imaginary; ownership

REFERENCES

Adloff, F., & Neckel, S. (2019). Futures of sustainability as modernization, transformation, and control: a conceptual framework. *Sustainability Science, 14*(4), 1015–1025. https://doi.org/10.1007/s11625-019-00671-2

Ancarno, C. (2020). Corpus-Assisted Discourse Studies. In A. Georgakopoulou & A. De Fina (Eds.), *The Cambridge Handbook of Discourse Studies* (pp. 165–185). Cambridge University Press. https://doi.org/DOI: 10.1017/9781108348195.009

Degens, P. (2021). Towards sustainable property? Exploring the entanglement of ownership and sustainability. *Social Science Information, 60*(2), 209–229. https://doi.org/10.1177/05390184211011437

Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to CDA. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse studies* (3rd ed., pp. 122–143). SAGE Publications.

Defining the core characters and events of a fictional narrative by two-mode social network analysis

Lauri Luoto
University of Turku, Finland
laeflu@utu.fi

Leo Lahti
University of Turku, Finland
leo.lahti@utu.fi

Kati Launis
University of Eastern Finland, Finland
kati.launis@uef.fi

Abstract

Inspired by the increased availability of digitized texts and the development of relevant technologies, as well as rise of the digital literary studies and the method of distant reading (Moretti, 2013), social network analysis has increasingly been used to analyze fictional narratives. The main lines of research focus on revealing patterns within a wider corpus and modelling interaction between characters within a single literary work (Chao et al., 2019). The present study adds to past research by analyzing two essential story elements in parallel, the plot and the characters. Plot is a structure connecting a series of events based on their logical relationships. Patterns of characters' co-participation in the events allows readers to see how characters and plot define each other thus building a co-dependent relationship between these two elements of the story. Data for the article is drawn from the renowned Finnish war novel *The Unknown Soldiers* by Väinö Linna (2015 [1954]). A two-mode dataset was constructed to show how the 58 characters of the novel co-appear in the 88 sections of the book, each of which illustrates an episode encountered by the troops. As a depiction of war, the novel describes various forms of interaction, including multiple characters sharing spaces and experiences together without any bilateral dialogue. The article contributes in part to the use of core-periphery analysis as an indicator of characters' positions. Among the methods created for identifying the core-periphery structure, and whether such structure can be found, there are varying and often inconsistent assumptions about how the core and the periphery are connected to each other. A binary typology distinguishes between two-block models and k-cores (Gallagher R. J. et al., 2021). The former partitions a network into a binary hub-and-spoke layout while the k-core decomposition divides the network into a layered hierarchy. In the present article, the discussion is extended to two-mode networks and a qualitative verification of the methods. Both the two-block model (Borgatti et.al., 2018) and the k-core (Cerinšek & Batagelj, 2015) are implemented for two-mode networks. The article assesses how the selection of the core-periphery algorithm affects the ability to reveal relevant observations about the protagonists and key events of the novel. Finally, perspectives provided by the network measures are compared with those of past scholarship.

Keywords: two-mode networks, duality, core-periphery structure, Väinö Linna, Finnish war novel

REFERENCES

Borgatti, S., Everett, M., & and Johnson, J. (2018). *Analyzing Social Networks*. Sage.

Cerinšek, M., & Batagelj, V. (2015). Generalized two-mode cores, *Social Networks 42*. https://doi.org/10.1016/j.socnet.2015.04.001

Chao, D. et al. (2019). Representing stories as interdependent dynamics of character activities and plots: A two-mode network relational event model. *Digital Scholarship in the Humanities 34*, 471-481. https://doi.org/10.1093/llc/fqy062

Gallagher R. J. et al. (2021). A clarified typology of core-periphery structure in networks. *Science Advances 7*. https://doi.org/10.1126/sciadv.abc9800

Linna, V. (2015 [1954]). *The Unknown Soldiers*. Penguin Books.

Moretti, F. (2013). *Conjectures on World Literature. Distant Reading.* Verso.

Automating data curation for the Finnish national bibliography Fennica

Julia Matveeva
University of Turku, Finland
yulia.matveeva@utu.fi

Osma Suominen
National Library of Finland, Finland
osma.suominen@helsinki.fi

Leo Lahti
University of Turku, Finland
leo.lahti@utu.fi

Abstract

The consortium Digital History for Literature in Finland (Research Council of Finland, 2022–26) differs from earlier research on Finnish literary history by making use of digital collections and new methods in data science, which enable the use of the collection as a whole.1 Here we present a dedicated bibliographic data science framework tailored for the specific context of the consortium research purposes. We will examine how data science methods can improve our understanding of literary history and how it's told, and how reliable the information can be. [1,2,3] The Finnish National Bibliography, Fennica, consists of over one million records from 1488 to the present and includes diverse data types such as books, newspapers, maps, and other documents from 1488 to the present day. The source data contains ambiguous information, missing or erroneous entries, however. Any refinement efforts will include context-specific choices that depend on the research use case. We have previously shown how selected subsets of the collection can be refined automatically to support large-scale statistical analyses of book printing during the years 1500-1800[1, 3]. In our present version of the workflow2, we've scaled up the previous analyses of 70 thousand records to cover all Fennica records and improved the data curation workflows. To cater to the research objectives of the project, we've integrated signum data and created a focused subset covering the years 1809-1917 for each metadata category. Furthermore, we've added a genre subfield derived from a broader leader field to enable genre identification at a bibliographic level, specifically focusing on books. Our aim has been to replicate a curated list, adhering to predefined criteria such as UDC classification, language, genre, and signum data. This automated process mirrors and supports the manual list creation method utilized by the Literary History subproject within the consortium. Future efforts could encompass the integration of further complementary sources, such as the rich information on authors, publishers, and geographic places available in the public domain. These efforts contribute to achieving the consortium's goals, which involve leveraging digital collections and methodologies to broaden the conventional understanding of Finnish literary history. Specifically, we aim to map Finnish and Swedish language fiction from the 19th century into an enriched format conducive to large-scale statistical analyses and the development of reproducible data science workflows.

Keywords: digital humanities, workflow, bibliography, fennica, data

REFERENCES

1. Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly,* 57(1), Special Issue. https://doi.org/10.1080/01639374.2018.1543747

2. Lahti, L., Mäkelä, E., & Tolonen, M. (2020). Quantifying bias and uncertainty in historical data collections with probabilistic programming. *CEUR Workshop Proceedings on Computational Humanities Research, 2723, Short* Paper 46. https://ceur-ws.org/Vol-2723/short46.pdf

3. Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2019). A quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 52(1), 57-78. https://doi.org/10.1080/01615440.2018.1513177

Tweets in Karelian: from data collection to the content analysis

Ilia Moshnikov
Karelian Institute, University of Eastern Finland, Finland
ilia.moshnikov@uef.fi

Eugenia Rykova
University of Eastern Finland, Finland
Catholic University of Eichstätt-Ingolstadt, Germany
eugenryk@uef.fi

Paper Abstract

The internet in general and social media in particular offer a new domain for the use of minority languages, which is important from the perspective of language vitality and language revitalisation. In our presentation we focus on the visibility of the Karelian language on X (formerly known as Twitter). Karelian is an endangered minority language spoken in Russia and Finland. According to the latest census and research, the total number of Karelian speakers is roughly about 20,000 people (Sarhimaa 2017; Federal State Statistics Service 2021).

We present our data collection strategy based on the use of language-related keywords and hashtags. The data was scraped from X (Twitter) using the Postman API software (Postman, 2023). The multilingual dataset combines many different languages, with Finnish dominating. Our final data consists of 2625 entries written entirely or partially in Livvi, South and Viena Karelian. The language and Karelian dialects were labelled manually by the first author of the study, who is a native Livvi-Karelian speaker. The visibility of Karelian on X has increased significantly in recent years, with Livvi-Karelian being the most prominent dialect (Moshnikov and Rykova 2023). Automatic language detection (Jauhiainen et al. 2022) identified Livvi-Karelian (or a mix of dialects including Livvi-Karelian) as such with 99.7% sensitivity, and South Karelian and Viena Karelian as Livvi-Karelian with 90% and 73.8% sensitivity, respectively.

We also analysed the topics of Twitter (X) entries written in Karelian. Ten main topics were identified manually by close reading each entry. Since the data was collected using keywords and hashtags related to the Karelian language itself, most of the entries are related to the language and vocabulary in sense of translation or language learning. However, personal tweets are the most numerous among the original entries. Tweets about the status of the Karelian language and the process of language revitalisation are particularly interesting from a research perspective as well as individual use of the language. In our data it is also possible to analyse tweets about the Karelian language written in Finnish and Russian.

Keywords: automatic language recognition, data scraping, Karelian, minority languages, X (Twitter).

REFERENCES

Federal State Statistics Service. (2021). *Vserossijskaja perepis' naselenija 2020 [Russian Census 2020]*. https://rosstat.gov.ru/vpn/2020.

Jauhiainen, T., Jauhiainen, H., & Lindén, K. (2022). HeLI-OTS, Off-the-shelf language identifier for text. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3912–3922. Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.416/.

Moshnikov, I. & Rykova, E. (2023). Little Big Data: Karelian Twitter Corpus. *Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023), 14–15 September 2023, University of Mannheim, Germany*, pp. 142–147. https://doi.org/10.14618/1z5k-pb25.

Postman. (2023). *Postman API Tool*. https://www.postman.com/.

Sarhimaa, A. (2017). *Vaietut ja vaiennetut. Karjalankieliset karjalaiset Suomessa [Silent and being forced to be silent: Karelian-speaking Karelians in Finland]*. Tietolipas 256. Helsinki: Finnish Literature Society.

Analyzing Finnish Inflectional Classes through Discriminative Lexicon Models

Alexandre Nikolaev
University of Eastern Finland
Alexandre.nikolaev@uef.fi

Harald R. Baayen
University of Tuebingen
harald.baayen@uni-tuebingen.de

Yu-Ying Chuang

Paper Abstract

Descriptions of complex nominal systems make use of inflectional classes. Inflectional classes bring together nouns which have similar stem changes and use similar exponents in their paradigms. Establishing what the inflectional classes of a language are is far from trivial, and across grammatical descriptions, the number of classes distinguished can vary considerably. Although inflectional classes can be very useful for language teaching as well as for setting up finite state morphological systems, it is unclear whether inflectional classes are cognitively real, in the sense that native speakers would need to discover these classes in order to learn how to properly inflect the nouns of their language.

This study investigates whether the Discriminative Lexicon Model can understand Finnish inflected words without setting up inflectional classes, using a dataset with 55271 inflected nouns of 2000 high-frequency Finnish nouns of 49 inflectional classes. Two DLM models were constructed, one using Endstate of Learning (EOL), and the other using Frequency-Informed Learning (FIL). Both models were given the task of predicting Finnish FastText embeddings from words' forms, represented by trigram-based vectors. The models were trained on 40694 word tokens, and evaluated on 14577 held-out low-frequency tokens. Overall accuracy on test data was higher for EOL (78%) than for FIL (50%), which is to be expected since EOL represents optimal learning with infinite exposure. Importantly, for both models, accuracies increased for inflectional classes with more types, more lower-frequency words, and more hapax legomena. Importantly, the accuracy of the DLM models mirrors the productivity of the inflectional classes. The model struggles more with novel forms of unproductive and less productive classes, and performs far better for unseen forms belonging to productive classes.

These results demonstrate that the inflectional system of Finnish nouns can be learned without hand-crafting of inflectional classes. Crucially, the extent to which generalization is possible matches the productivity of the inflectional classes distinguished by linguistic

analysis. We are currently investigating whether deep discriminative learning can provide even more accurate mappings, but we anticipate that with these even more powerful mappings, models will outperform what can be expected for human native speakers.

Keywords: Discriminative Lexicon Model, Finnish, FastText, Word embeddings, Inflectional morphology

Atlas of Finnish Literature 1870–1940

Asko Nivala & Jasmine Westerlund
University of Turku
njwest@utu.fi

PaperAbstract

The project Atlas of Finnish Literature 1870–1940 (the Alfred Kordelin Foundation – Major Cultural Projects 2022–2024) participates in the research tradition of literary cartography. Literary cartography, which emerged in the late 1990s as part of the spatial turn in the humanities, has, in the 2000s, adopted the methods of digital humanities. Our project applies these new methods for the first time to the study of Finnish fiction. In the project Atlas of Finnish Literature 1870–1940, we have extracted geographical information from two corpora of literary texts (Project Lönnrot and Project Gutenberg) and developed an interactive web application where you can plot the spatial references from texts on a map. The texts were transformed from plain texts to TEI/XML and then processed with named entity recognition and linking tools. In this presentation, we introduce the process we have used: NER, geocoding and linking to external data sources. We will also introduce the Atlas of Finnish Literature 1870–1940 web application, which is open to the public, and give research examples of the kind of information that can be found with the application. Our webpage contains 846 works from 1870 to 1944 that are geocoded. The application can be used for historical research, literary studies and geography. The database can be filtered and viewed by work, by author or by location. One can see, for example, where Santeri Ivalo's Anna Fleming (1898) is situated, what is the spatial scope in Minna Canth's works or which books mention Helsinki. Finally, we have added full-text search to the application, which brings interesting new possibilities to the study of Finnish fiction.

Keywords: Literary geography, digital humanities, maps, Finnish literature, cultural history, named entity recognition

Artificial Melodies: Investigating the Limits of AI in Replicating Human Songwriting

Maria Claudia Nunes Delfino, Tony Berber Sardinha
Institution, country
email@email.com

Paper Abstract

The emergence of Artificial Intelligence (AI) chatbots has ignited debates regarding the potential replacement of human-generated texts, particularly structured content such as weather forecasts or financial reports. However, its application in creative domains like poetry or songwriting is increasingly acknowledged. This study aims to investigate the capacity of AI to replicate creative human writing, specifically focusing on song lyric composition in English. To achieve this goal, we conducted two Lexical Multi-Dimensional analyses (LMDA; Berber Sardinha and Fitzsimmons-Doolan, 2024), employing a curated corpus of song lyrics spanning diverse musical genres (including country, pop, rap, rock, and soul), which encompassed both chart-topping and random lyrics. Additionally, we generated a comparative corpus of artificially produced lyrics using ChatGPT, Google's Gemini, and Meta's Llama. The corpus consisted of 4000 lyrics, evenly split between human-authored and AI-generated texts, with each subcorpus comprising 400 lyrics per style. The first analysis involved conducting an additive LMDA based on the dimensions of variation identified by Author (2023). These dimensions were derived from a large corpus of over 100,000 song lyrics, each tagged for semantic class using the USAS semantic tagger. The dimensions are the following: 1) Materialism and Superficiality, 2) Alterity and Interpersonal Dynamics, 3) Mysticism and Transcendence, and 4) Romanticism and Personal Quest. After scoring each of our lyrics on these dimensions, we ran a Discriminant Function Analysis (DFA) to classify the lyrics as either human-authored or AI-generated. The results showed a 64% accuracy rate in identifying AI-generated songs and an 84% accuracy rate for human-authored songs. The second LMDA used the actual vocabulary of the songs, rather than semantic classes. Key lemmas were extracted for each authorship condition, which were then subjected to factorial analysis, resulting in a five-dimensional model: 1) Social Justice versus Romance, 2) Reality versus Transcendence, 3) Rural versus Urban, 4) Individualism versus Collectivism, and 5) Extroversion and Physicality versus Introversion and Emotions. This model demonstrated efficacy in classifying lyrics, achieving a 69% success rate for AI-generated lyrics and 90% for human-authored songs. Overall, the findings indicate a significant discrepancy between AI and human songwriting capabilities. Only 30.90% of AI-generated songs closely resembled those written by humans, suggesting that while AI can replicate human songwriting vocabulary, it falls short in generating discourses that align with human musical expressions. Conversely, human-authored songs were accurately identified with a high degree of precision (90.05%), underscoring the distinct and irreplicable aspects of human creativity in songwriting.

Keywords: Corpus Linguistics, Popular music, Multi-Dimensional Analysis, Artificial Intelligence

A cancer of Finnish or a great happiness to us all? A corpus-assisted discourse study on the English language in the Suomi24 discussion forum

Heli Paulasto, Lea Meriläinen
University of Eastern Finland
heli.paulasto@uef.fi

Paper Abstract

The Increased use of English has brought about a massive change in the sociolinguistic landscape of the Nordic countries during the past few decades (see Peterson & Beers Fägersten 2024). This change is not welcomed by all: in Finland, the prominent role of English in the public sphere is also a cause for concern and produces heated discussion in the media as well as among the general public, whose views both reflect and contribute to the broader discourses pertaining to English (Saarinen & Ennser-Kananen 2020). In this study, we examine a popular online discussion forum to uncover layman views related to the English language in Finland and the emerging attitudes and ideologies. The study utilizes digital methods and data in the form of corpus-assisted discourse analysis. Our focus is on the word englanti 'English (language)' and its common collocates in Suomi24 Corpus (Lagus et al. 2016), based on Finnish language discussions on the Suomi24 online forum in the years 2018-2020. Through the study of collocates, i.e. commonly co-occurring words (see Scott & Tribble 2006), it is possible to identify recurring themes, which are then further analyzed qualitatively through the lens of language ideologies and discourses. Our research questions are: 1) What are the most common collocates of the word englanti in the Suomi24 Corpus and, based on the collocates, what are the central topics of discussion on English in Finland? 2) What kinds of discourses can be identified around these collocates, and what do the findings reveal about linguistic ideologies and the current sociolinguistic climate in Finland? The emerging discourses reflect language speakers' reactions to and attitudes towards the changing linguistic sphere in Finland. While Finland as well as other Nordic countries have always been multilingual, English has fairly rapidly begun to occupy the space of national languages, which causes tensions. One of the most prevalent themes in the discussions is English as a threat vs. opportunity, which is also observed in public opinions in other Nordic countries (Mortensen 2024). By comparing the findings to studies conducted in other Nordic countries, it is possible to gain a broader perspective into the public discussion on the role of English in Finland.

Keywords: Corpus-assisted discourse analysis, language ideologies, English language

REFERENCES

Author, A. A. (Year of publication). *Title of work: Capital letter also for subtitle*. Publisher Name. DOI (if available)

Lagus, K., Pantzar, M., Ruckenstein, M. & Ylisiurua, M. 2016. Suomi24. Muodonantoa aineistolle. Valtiotieteellisen tiedekunnan julkaisuja, Nro 10. Helsinki: Helsingin Yliopisto.

Mortensen, J. 2024. Beyond threat or opportunity: English and language ideological tensions in the Nordic countries. In Peterson, E. & Beers Fägersten, K. (eds.), English in the Nordic Countries. Connections, Tensions, and Everyday Realities. New York & London: Routledge, 104-124.

Peterson, E. & Beers Fägersten, K. (eds.) 2024. English in the Nordic Countries. Connections, Tensions, and Everyday Realities. New York & London: Routledge.

Saarinen, T. & Ennser-Kananen, J. 2020. Ambivalent English: what we talk about when we think we talk about language. Nordic Journal of English Studies 19(3): 115–129.

Scott, M. & Tribble, C. 2006. Textual patterns. Key words and corpus analysis in language education. Amsterdam: John Benjamins.

#WOMENINSTEM: A Corpus-Based Multimodal Critical Discourse Analysis of STEM Identity Construction and Advocacy Performance on Instagram

Laura Sofia Pensabene
University of Catania, Italy
laura.pensabene@phd.unict.it

Paper Abstract

Digital spaces have paradigmatically altered the way we communicate, giving people the opportunity to speak freely and reach potentially infinite audiences (Sergeant & Tagg, 2014).
Such affordances, typical of social media, have been progressively and intrinsically exploited by minorities, thus allowing them to resist social, cultural and institutional power (Buktus, 2023) and renegotiate identities.
Historically, women have been quite marginalised in STEM (Science, Technology, Engineering, Mathematics) fields: still in 2023, they made up on average only 28% of the STEM workforce globally (Piloto, 2023); prejudices and biases are actually at the core of the persistence of such *status quo,* resulting in the perception of STEM fields as male-dominated (Lee, 2008).
Nevertheless, in recent years an increasing number of STEM women has undoubtedly started to take advantage of social networks affordances (Montgomery, 2018) to construct public counter-discourses against patriarchal institutions and culture, as it is the case with hashtag feminism (Linabary et al., 2020; Semenzin 2022).
Indeed, by recounting their day-to-day experiences, women aim to 'own' the narrative of what being a woman in STEM actually is and implies, redefining therefore their STEM identity (Kim et al., 2018), providing at the same time genuine representation and inspiration for the "next STEMM gen".
This works aims at analysing how STEM identity and advocacy have been both discursively and visually constructed and performed during the two weeks surrounding both the International Day of Women and Girls in Science (11th February) and the International Women's Day (8th March).
Our dataset stems from a search of the #womeninstem: hanging out on Instagram, we identified several accounts of women working both in Academia and Industry in different STEM fields; we therefore selected 15 accounts belonging to the former category. Firstly, the linguistic and visual content of the posts selected for that day were analysed employing Multimodal Critical Discourse Analysis tools (Machin & Mayr, 2012); afterwards, an *ad hoc* corpus of comments of the posts was created to observe wordlists, collocates, and frequencies (Hunston, 2022): both methods combined helped us investigate how both textual and visual content were used to perform STEM identity and advocacy and helped us reach the preliminary conclusion that such identity is performed discursively (with women constantly appealing to the women in the STEM community), multimodally (with

visual elements fostering diverse representation) and lastly by means of co-construction through the comment section.

Keywords: Corpus Linguistics, Identity, Instagram, Multimodal Critical Discourse Analysis, STEMinism

REFERENCES

Buktus, C. M. (2023). *Social Media, Marginalised Identity and Liminal Publics* [Doctoral Thesis, University of Technology Sydney]. Opus Lib Uts Edu Repository. https://opus.lib.uts.edu.au/handle/10453/173479

Hunston, S. (2022). *Corpora in Applied Linguistics* (2nd ed.). Cambridge University Press.

Kim, A., Sinatra, G. M. & Seyranian, V. (2018). Developing a STEM Identity Among Young Women: A Social Identity Perspective. *Review of Educational Research*, 20(10), 1-37. https://doi.org/10.3102/0034654318779957

Lee, J. A. (2008). *Gender equity issues in technology education: A qualitative approach to uncovering the barriers*. [Doctoral Thesis, Carolina State University]. NC Repository. https://repository.lib.ncsu.edu/items/0c73fd52-3730-49a3-acde-9269c97dd8de

Linabary, J.R., Corple, D.J. & Cooky, C. (2020). Feminist activism in digital space: Postfeminist contradictions in# WhyIStayed. *New Media & Society*, 22(10), 1827–1848. https://doi.org/10.1177/1461444819884635

Machin, D. & Mayr, A. (2012). *How to do Critical Discourse Analysis. A Multimodal Introduction*. Sage Publications Ltd.

Montgomery, B. L. (2018). Building and Sustaining Diverse Functioning Networks Using Social Media and Digital Platforms to Improve Diversity and Inclusivity. *Frontiers in Digital Humanities*, 5(22), 1-11. https://doi.org/10.3389/fdigh.2018.00022

Piloto, C. (2023, March 13). *The Gender Gap in STEM: Still Gaping in 2023*. MIT Professional Education. https://professionalprograms.mit.edu/blog/leadership/the-gender-gap-in-stem

Seargeant P. & Tagg C. (Eds). (2014). *The Language of Social Media Identity and Community on the Internet*. Palgrave.

Semenzin, S. (2022). “Swipe up to smash the patriarchy”: Instagram feminist activism and the necessity of branding the self. *AG AboutGender,* 11(21), 113-141. https://doi.org/10.15167/2279-5057/AG2022.11.21.1990

Event-based Experience Sampling of Music Listening with the MuPsych app

William Mathew Randall
University of Jyväskylä
will.m.randall@gmail.com

Paper Abstract

The In the age of the internet and smartphones, music listening has become a more portable, accessible, and personalised experience. As this personal style has unique potential for influencing the emotions and well-being of listeners, it is important to understand the complete range of variables involved. An innovative solution to this challenge comes from the mobile app MuPsych, which utilises the experience sampling method (ESM) to capture music listening experiences as they occur in everyday life. The app presents participants with a series of questions at the moment they start listening to music on their phone, allowing for real-time and ecologically valid data measurement of listening experiences. Data from these music reports are combined with individual variables, through a battery of psychological surveys presented within the app. To complement these sources of self-report data, the app can also collect track and artist data, physiological data from wearable devices, and weather data. The main purpose of research using the MuPsych app has been to develop a comprehensive model of how music influences emotional states, through a complex interaction of music, listener, and context variables. The app is also available to all music researchers, as a tool to investigate various phenomena related to the listening experience, through custom studies. In the future, the data collected by MuPsych will be used to develop a music recommender, which will create playlists based on listener mood, activity, and reason for listening, while supporting emotional health and well-being.

Keywords: Experience sampling, Music and Emotions, Data collection

Digital Methodologies in Forensic Linguistic Authorship Analysis: Social Media Data and Computational Approaches in Geolinguistic Profiling

Dana Roemling
University of Birmingham, UK | University of Helsinki, Finland
danaroemling@gmail.com

Paper Abstract

Research and case work in forensic authorship profiling focuses on inferring social characteristics of unknown authors from their texts, such as age, gender or first language influence, while drawing on foundational work laid in sociolinguistics (Nini, 2018). However, inferring the regional background of an author has received limited attention, although one of the most prominent cases in forensic authorship profiling was resolved recognising the regionalism "devil strip" in a ransom note (see Shuy, 2001).

With computational methods and large corpora of natural language data being available, this study moves away from the traditional approach to geolinguistic profiling by spotting regionalisms and using dictionaries or dialect atlases in the hopes of placing the word in question. For this the study employs a corpus of 21 million German social media posts from the platform Jodel (Hovy & Purschke, 2018) and provides an evaluation of the regionally distributed data in the corpus. Given that geolocated social media data is often sparse and centred on cities, the study uses ordinary kriging (see Wackernagel, 2003), i.e. geospatial statistics, to interpolate the data for unobserved locations, thus enhancing the resolution for location prediction while visualising the results to make them more accessible. Further, the study presents an algorithm to predict the dialect region of an author in question and discusses both the explainability of the prediction in the forensic context and the accuracies reached. The study finds that apart from being a reference tool for qualitative analyses in forensic investigations, this corpus also allows us to extract linguistic features relevant for forensic analyses that are not based on previous knowledge.

Not only does this research advance the field of forensic authorship profiling by reducing the reliance on an analyst's expertise to spot regionalisms, but it also illustrates how interdisciplinary research in linguistics, NLP, digital technologies and forensic science can improve the delivery of justice.

Keywords: authorship analysis, corpus linguistics, dialect classification, forensic linguistics, geospatial statistics

REFERENCES

Hovy, D., & Purschke, C. (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. *Proceedings of the 2018 Conference on EMNLP*, 4383–4394. https://doi.org/10.18653/v1/D18-1469

Nini, A. (2018). Developing forensic authorship profiling. *Language and Law / Linguagem e Direito*, *5*(2), 38–58.

Shuy, R. W. (2001). DARE's role in linguistic profiling. *DARE Newsletter, 4*(3), 1–5.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer.

Acknowledgements

Dana Roemling was supported by the UKRI ESRC Midlands Graduate School Doctoral Training Partnership ES/P000711/1.

They thank Dirk Hovy and Christoph Purschke for sharing their data with them.

They wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

AI-based Personalized Feedback in Speech Therapy for People with Aphasia

Eugenia Rykova
University of Eastern Finland, Finland,
Catholic University of Eichstätt-Ingolstadt, Germany
eugenryk@uef.fi

Paper Abstract

In the field of speech and language therapy, artificial intelligence has been used in diagnostics, therapy, and assistive systems for people with aphasia (PWA) (Adikari et al., 2023; Azevedo et al., 2024; Pottinger & Kearns, 2024). AphaDIGITAL project (TDG, 2021) focuses on developing such a mobile application for German-speaking PWA that will provide personalized multilevel feedback with the help of Automatic Speech Recognition (ASR) and further text analysis. To build the corresponding pipeline (Rykova & Walther, 2024), the following questions are addressed.

Which existing ASR solutions are suitable for the task-specific speech of German-speaking PWA?

More than 50 open-source ASR solutions were evaluated with the help of several speech recordings from different corpora (Rykova, Walther, & Zeuner, 2022). Thirteen models were selected and tested with atypical speech, including two small datasets of PWA's speech (Rykova & Walther, in press - a). Based on Character Error Rate (CER), HITS (the number of precisely recognized words) and the number of empty outputs, four open-source ASR models were selected for the pipeline (Fleck, 2022; Grosman, 2022; Guhr, 2022; NVIDIA, 2022). These models are to a greater or lesser extent robust to speaker gender and age. The experiments suggest that for better single-word recognition the audio samples should be not too short and pronounced neither too slowly nor too fast (i.e. intentionally speeded up) (Rykova & Walther, in press - b).

How can selected ASR solutions be improved and/or adapted for the purposes of speech and language therapy?

In the absence of adequate data for ASR models' (re)training, applying the knowledge about non-standard (aphasic and dialect) phonetic features post-hoc to ASR output was attempted. Aphasic features included recognition of syllables as separate words and vowel prolongation. Dialect features were selected from the Thuringia-Upper Saxon dialect group (Wallraff, 2007; Rocholl, 2015; B. Siebenhaar, personal communication, January, 2024). The method combined generating alternative pronunciations based on non-standard patterns (Masmoudi et al., 2014) and using alternatives for evaluation (Ali et al., 2017), and proved to work on the recordings of German aphasia test naming and repetition tasks (Huber, 1993).

How can a combination of selected ASR solutions and existing tools for semantic and grammatical analysis serve for speech production errors analysis?

If the answer of the speaker is not recognized as fully correct or containing phonetic/phonemic errors, it is subject to semantic analysis. It must be compared to the target in terms of their semantic relationship and distance. The current semantic analysis pipeline is built upon GermaNet – a semantic network for the German language (Hamp & Feldweg, 1997). It includes recognition of hyponymy/hypernymy, belonging to the same semantic (sub)category, and different lexical and conceptual relationships, derived from GermaNet. If the answer is not recognized as an existing word, a search for close orthographic matches is performed, and the match that is semantically the closest to the target is subject to the relationship analysis described above. This approach has been tested and described in detail in Rykova & Walther (2023).

Keywords: Aphasia; automatic speech recognition; speech and language therapy; digital health.

REFERENCES

Adikari, A., Hernandez, N., Alahakoon, D., Rose, M. L., & Pierce, J. E. (2024). From concept to practice: A scoping review of the application of AI to aphasia diagnosis and management. *Disability and Rehabilitation*, *46*(7), 1288-1297.

Ali, A., Nakov, P., Bell, P., & Renals, S. (2017, December). WERd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 141-148). IEEE.

Azevedo, N., Kehayia, E., Jarema, G., Le Dorze, G., Beaujard, C., & Yvon, M. (2024). How artificial intelligence (AI) is used in aphasia rehabilitation: A scoping review. *Aphasiology*, *38*(2), 305-336.

Fleck, M. (2022). *Wav2vec2-large-xls-r-300m-german-with-lm.* Retrieved from https://huggingface.co/mfleck/wav2vec2-large-xls-r-300m-german-with-lm. Accessed September 12, 2022.

Grosman, J. (2022). *Fine-tuned XLSR-53 large model for speech recognition in German*. Retrieved from https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german. Accessed September 12, 2022.

Guhr, O. (2022). *wav2vec2-large-xlsr-53-german-cv9.* Retrieved from https://huggingface.co/oliverguhr/wav2vec2-large-xlsr-53-german-cv9. Accessed September 12, 2022.

Hamp, B., & Feldweg, H. (1997). GermaNet – a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Huber, W. (1983). *Aachener Aphasie Test (AAT) [Aachen Aphasia Test]*. Verlag für Psychologie Hogrefe, Göttingen, Zürich.

Masmoudi, A., Khmekhem, M. E., Esteve, Y., Belguith, L. H., & Habash, N. (2014, May). A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition. In *LREC* (pp. 306-310).

NVIDIA. (2022). *NVIDIA Conformer-Transducer Large (de)*. Retrieved from https://huggingface.co/nvidia/stt_de_conformer_transducer_large. Accessed September 12, 2022.

Pottinger, G., & Kearns, Á. (2024). Big data and artificial intelligence in post-stroke aphasia: A mapping review. *Advances in Communication and Swallowing*, (Preprint), 1-15.

Rocholl M.J. (2015). *Ostmitteldeutsch – eine moderne Regionalsprache? Eine Untersuchung zu Konstanz und Wandel im thüringisch-obersächsischen Sprachraum [East-Central German – a modern regional language? An investigation into constancy and change in the Thuringian-Upper Saxon language area]*. Hildesheim, Zürich, New York: OLMS.

Rykova, E., & Walther, M. (2023, March). Concept for semantic error analysis in a mobile application for speech and language therapy support. In *Konferenz Elektronische Sprachsignalverarbeitung* (pp. 127-133). TUDpress, Dresden.

Rykova, E., & Walther, M. (2024). AphaDIGITAL – digital speech therapy solution for aphasia patients with automatic feedback provided by a virtual assistant. *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS).*

Rykova, E., & Walther, M. (in press - a). Evaluation of German ASR solutions for speech and language therapy support of people with aphasia. *LOQUENS.*

Rykova, E., & Walther, M. (in press - b). Linguistic and extralinguistic factors in automatic speech recognition of German atypical speech. *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*.

Rykova, E., Walther, M., & Zeuner, E. (2022, August). *aphaDIGITAL — Avatar-based digital speech therapy solution for aphasia patients: first evaluation*. Poster session presentation at the 35th Fonetiikan Päivät, Joensuu, Finland. Available at https://www.researchgate.net/publication/364676432_aphaDIGITAL_-Avatar-based_digital_speech_therapy_solution_for_aphasia_patients_first_evaluation.

TDG — Translationsregion für digitale Gesundheits-versorgung [Translational Region for Digital Healthcare]. (2021). *AphaDIGITAL: Entwicklung einer digitalen, dezentralen sprachtherapeutischen Versorgung [Development of digital, decentralized speech therapy solutions]*. Retrieved from https://inno-tdg.de/projekte/aphadigital/. Accessed January 25, 2023.

Wallraff, U. (2007). *Ausgewählte phonetische Analysen zur Umgangssprache der Stadt Halle an der Saale [Selected phonetic analyses of the colloquial language of the city of Halle an der Saale]*. [Doctoral Dissertation, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany].

Acknowledgements/Disclaimers

AphaDIGITAL project was sponsored by German Federal Ministry of Education and Research under funding code 03WIR3108A via the TDG innovation ecosystem (Translationsregion für digitale Gesundheitsversorgung [Translational region for digital

healthcare]) and „WIR! – Wandel durch Innovation in der Region" [Change through innovation in the region] program.

Word proximity and dependencies in parliamentary discourse in Finnish parliament

Kirsi Sandberg, Juho Karvinen, Aarne Ranta, Jyrki Nummenmaa
University of Tampere
kirsi.sandberg@tuni.fi

Paper Abstract

The In digital humanities, so-called distant reading often relies on methods based on statistics on proximity of words used in the given textual context. Prior to the data analysis, textual data is typically lemmatized and stop words are removed, which from linguistic perspective erases a layer of textual meaning as well as the actual informational content from the sentences (see e.g. Lambrecht 1996). Even if the content words can provide an overall comprehension of discourse entities mentioned in a given text, the majority of research questions in the fields of humanities and social sciences are concerned with how the entities are spoken about as well as the explicated relations between the entities, both typically expressed grammatically: with function words, inflection and word order. This study sets out to evaluate the extent to which a computational analysis based on grammatical relations (as in syntactic dependencies) - instead of word proximity - can capture central features of temporal relations expressed in parliamentary debate discourse and elaborate the methodological approaches to parliamentary data and political temporality. Parliamentary discourse is nothing but straightforward; discourse entities and processes MPs refer to during parliamentary debates are typically abstract and expressed with complex noun phrases or infinitive constructions. Due to this, the rhetorical wordings under scrutiny can comprise the core of the expression as well as stand as a more distant frame or a modifier for what actually is stated. We report three cases where a noun referring to time is used in different syntactic positions: head of a noun phrase, modifier of a noun phrase and as one of the main arguments in the clause (subject, object). We focus on plenary sessions of Finnish parliament, a showcase for highly inflectional language with flexible word order. The data set consists of the official records of Finnish parliamentary debates from 1980 to 2022 (see Andrushchenko et al. 2021) and is dependency-parsed with the Finnish neural parser (Turku NLP, Kanerva et al. 2018). Universal Dependencies provide a language independent framework thus it also enables systematic comparisons between languages used in different parliaments. The analyses are obtained from a machine-learned parser in a standardized syntax tree format. They are then sent to a rule-based pattern matching tool, which finds subtrees and sequences of trees satisfying relevant conditions such as "sentence where the word 'future' is used as subject or object", or "sequence of sentences in the past tense".

Keywords: universal dependencies, parliamentary records, temporality

REFERENCES

Andrushchenko, M., Sandberg, K., Turunen, R., Marjanen, J., Hatavara, M., Kurunmäki, J., Nummenmaa, T., Hyvärinen, M., Teräs, K., Peltonen, J. & Nummenmaa, J. (2021) Using parsed and annotated corpora to analyze parliamentarians' talk in Finland. Journal of the Association for Information Science and Technology (JASIST) https://doi.org/10.1002/asi.24500

Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. Proceedings of the conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/K18-2013

Lambrecht, K. (1996) Information Structure and Sentence Form : Topic, Focus, and the Mental Representations of Discourse Referents. Cambridge: Cambridge University Press

Navigating the ethical and legal dimensions of Human-AI co-creativity in Interaction Design

Anca Serbanescu
anca.serbanescu@polimi.it

Paper Abstract

In our rapidly evolving contemporary landscape, characterized by the pervasive influence of technological progress, artificial intelligence (AI) emerges as a significant catalyst driving transformative shifts within society(Serbanescu, 2024; Serbanescu & Nack, 2024). However, amidst its potential lies a veil of uncertainty regarding the ethical and legal responsibilities incumbent upon its creators and users. While numerous ethical guidelines have been promulgated by diverse entities, spanning from corporate entities (IBM, 2019; FLI, 2021; Microsoft, 2022) to governmental bodies (EU, 2023; United Nations, 1948; Wiewiórowski & Wojciech, 2022), a noticeable dearth of commensurate legal frameworks governing the actions of designers within the realm of human-AI co-creativity persists. This contribution seeks to shed light on the ethical aspects surrounding human-AI co-creativity, delving into the theoretical underpinnings through a critical examination of existing literature. By analyzing two case studies—DesignPal (Rezwana & Maher, 2023) and AniThings (Marenko & Van Allen, 2016)— the aim is to elucidate the ethical considerations arising from the collaborative nexus between humans and AI within the creative process. As the dynamics of human-AI co-creativity are scrutinized, pivotal inquiries surface: What ethical implications emerge from the symbiotic relationship between humans and AI in creative endeavors? How should the mantle of responsibility be shared between human designers and AI systems throughout the creative process? However, the delineation of a designer's accountability in the development of AI systems remains opaque, encompassing not only ethical but also legal dimensions. Thus, this study endeavors to elucidate the extent of a designer's responsibility within the existing scholarly discourse, aiming to clarify the ethical and legal obligations inherent in co-creating with AI support systems. By traversing the blurred boundary between ethical considerations and legal obligations in human-AI co-creativity, this study contributes to a more comprehensive understanding of the complex interplay between technology and ethics, paving the way for informed decisionmaking and responsible innovation in interaction design.

REFERENCES

EU, (2023). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/en/txt/?uri=celex%3a52021pc0206

IBM, (2019). Fundamentals. IBM Design for AI. https://www.ibm.com/design/ai/fundamentals

FLI, (2021). AI Principles. 2022(August 16,). https://futureoflife.org/2017/08/11/ai-principles/

Marenko, B., & Van Allen, P. (2016). Animistic design: How to reimagine digital interaction between the human and the nonhuman. Digital Creativity, 27(1), 52–70.

Microsoft, (2022). Microsoft responsible AI standard, V2.

Rezwana, J., & Maher, M. L. (2023). User Perspectives on Ethical Challenges in Human-AI Co-Creativity: A Design Fiction Study. 62–74.

Serbanescu, A. (2024a). Human-AI Co-Creativity. Understanding the Relationship between Designer and AI systems in the field of Interactive Digital Narrative. Politecnico di Milano.

Serbanescu, A., & Nack, F. (2024). Towards an analytical framework for AI-powered creative support systems in interactive digital narratives. Journal of Entrepreneurial Researchers.

United Nations. (1948). Universal declaration of human rights. https://www.un.org/en/about-us/universaldeclaration-of-human-rights

Wiewiórowski & Wojciech. (2022). EDPS Homepage (European Data Protection Supervisor). https://edps.europa.eu/_en

Exploring the Potential of AI-Generated Texts to Replace Human-Written Content in Language Education

Marilisa Shimazumi, Tony Berber Sardinha
Institution, country
email@email.com

Paper Abstract

The incorporation of AI-generated texts into educational materials is an emerging topic of interest, particularly concerning the potential application of AI in crafting tasks for language teaching. The goal of our research is to examine the capability of AI-generated texts to replicate the characteristics of English coursebook text samples used in language instruction. This analysis enables assessing the viability of replacing traditional human-written instructional content with AI-generated texts in educational settings. Our investigation is set against the background that textbook texts, conventionally employed as exemplars in language education, are specifically tailored and revised to match a certain level of difficulty, and thus do not fully represent authentic language usage in everyday scenarios. Nevertheless, these coursebook texts exhibit a distinct form of human authorship, shaped by the instructional requirements of students learning a second language. The ability of AI to produce simplified texts that are on par with those created by humans remains an open question. To fill this gap, we conducted a Multi-Dimensional Analysis (Biber, 1988, 1995; Berber Sardinha & Veirano Pinto, 2014, 2019) of our English Language Teaching textbook corpus (ELTT corpus), encompassing 106,840 words from 500 texts across 19 different registers. These texts, sourced from 43 books by major publishers over 25 years (1996 to 2021), spans B2 and C1 levels, with an equal number of texts from each level. Five dimensions were identified, namely (1) Persuasion, speaker engagement, and personal opinion vs Expression of analysis and technical information; (2) Expressive, interactive, speculative discourse with stance marking; (3) Formal, informative, detailed composition; (4) Narrative and descriptive accounts; (5) Summarized abstracted overviews. Each dimension comprises a set of correlated grammatical features performing the major functions corresponding to the dimensions. As a comparison sample, we created an AI-generated corpus (AI-ELTT corpus) using ChatGPT to simulate textbook texts, resulting in 500 comparable texts. In general, the results showed that AI EFL coursebook text models are different from human counterparts. First, AI struggles with producing texts that emphasize persuasion, speaker engagement, and personal opinion. Instead, AI-generated texts are characterized by the expression of analysis and technical information. Secondly, AI faces difficulties in producing language that is expressive, interactive, and speculative with stance marking, reducing the incidence of these features. Given these differences, it was possible to successfully differentiate AI from human texts in more than 80% of cases.

Keywords: Multi-Dimensional Analysis, Language Teaching, Artificial Intelligence

REFERENCES

Biber, D. (1988). Varia%on Across Speech and Wri%ng. Cambridge: Cambridge University Press.

Biber, D. (1995). Dimensions of Register Varia%on - A Cross-Linguis%c Comparison. Cambridge: Cambridge University Press.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2014). Mul%-Dimensional Analysis, 25 years on: A Tribute to Douglas Biber. Amsterdam: John Benjamins.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). Mul%-Dimensional Analysis: Research Methods and Current Issues. Abingdon: Bloomsbury.

A Dialectometric Study of Low Saxon Syntactic Variation through Time

Janine Siewert
University of Helsinki, Finland
janine.siewert@helsinki.fi

Paper Abstract

We present a corpus-based dialectometric study of synchronic and diachronic syntactic variation in literary Low Saxon, where we focus on aggregate similarity on the one hand and the occurrence of particular structures on the other. These results are then compared to our previous studies targeting other levels of representation as well as to findings from traditional dialectology. Our two major research questions for this study are: (1) Does the overall syntactic similarity of the dialect groups change over time and, if yes, how? (2) Do certain structures considered characteristic for Low Saxon decrease in frequency in the written language as well, as found in studies on spoken language?

The major part of our Modern Low Saxon data comes from the LSDC dataset Siewert et al. (2020) and is divided into two time periods (1800–1939 and 1980–2022) and six major dialect groups (Figure 1). Our findings will be compared to the Reference Corpus Middle Low German / Low Rhenish ReN-Team (2019).

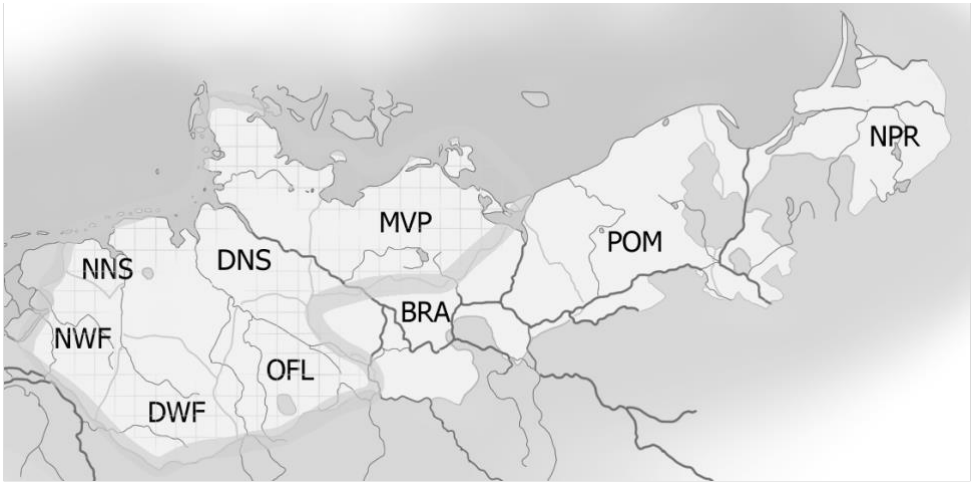

Figure 1. Low Saxon dialects included in our study: NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish – West Pomeranian, OFL: Eastphalian. Other dialects: BRA: Brandenburgish, POM: East Pomeranian, NPR: Low Prussian

In addition to our own research, recent dialectometric studies of Low Saxon have appeared by, for instance, Buurke et al. (2022) and Bartelds and Wieling (2022). A slightly older study is by Lameli (2016) who re-analysed the Wenker atlas data and found a north-south split in German Low Saxon.

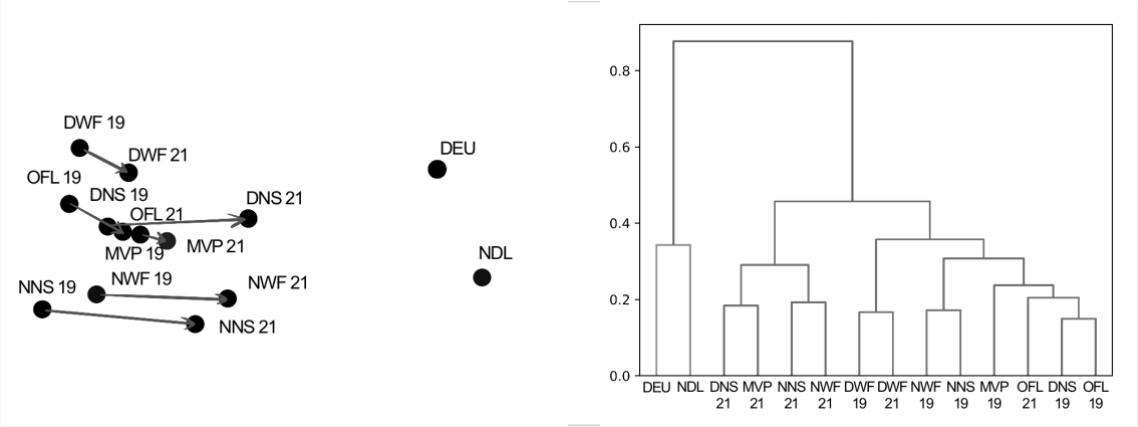

Figure 2. PoS level PCA and hierarchical clustering. NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish – West Pomeranian, OFL: Eastphalian. 19: 1800–1939, 21: 1980–2022

In previous experiments, we have compared aggregate distances in Modern Low Saxon, Standard Dutch and Standard German at the levels of characters, PoS (Part-of-Speech) tags and morphological features (explained here: https://universaldependencies.org/guidelines.html) from whole corpora. Here, we have found different trends at the different levels of representation. Whereas Dutch Low Saxon seems to approach Standard Dutch at all levels, the picture for German Low Saxon is more diverse: While we find a comparable trend of German Low Saxon approaching Standard German at the PoS level (Figure 2), when adding morphological information, the northern dialects appear to approach Standard Dutch (Figure 3). Furthermore, similar to Lameli, we find a north-south division in German Low Saxon to be more prominent than the traditionally assumed east-west division (compare, e.g., Schröder, 2004).

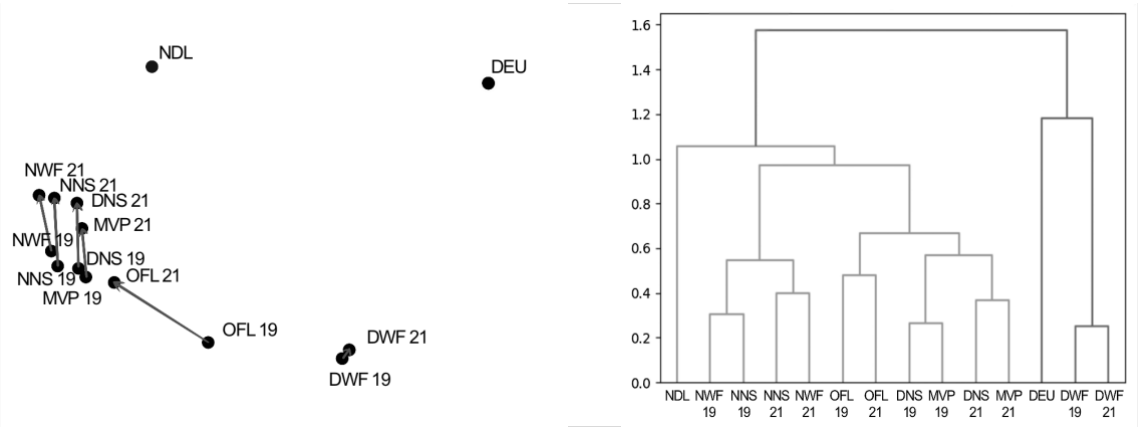

Figure 3. PoS and morphological features, PCA and hierarchical clustering. NNS: Dutch North Saxon, NWF: Dutch Westphalian, DNS: German North Saxon, DWF: German Westphalian, MVP: Mecklenburgish – West Pomeranian, OFL: Eastphalian. 19: 1800–1939, 21: 1980–2022

To complement our previous studies, we make use of syntactic relations that and lemmata to look at structures that PoS tags do not sufficiently differentiate. In addition to the aggregate similarity based on syntactic relations, we particularly want to investigate the occurrence of structures that according to Elmentaler and Borchert (2012) are often presented as characteristic for Low Saxon in textbooks and grammar books but which they have not found to be particularly frequent in the spoken language.

Keywords: computational dialectology, diachronic variation, dialectometry, Low German, Low Saxon

REFERENCES

Bartelds, Martijn & Wieling, Martijn (2022). Quantifying language variation acoustically with few resources. In Marine Carpuat, Marie-Catherine de Marneffe & Ivan Vladimir Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the As-sociation for Computational Linguistics: Human Language Technologies* (3735-3741). Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.273.

Buurke, Raoul Sergio Samuel Jan, Sekeres, Hedwig G, Heeringa, Wilbert, Knooihuizen Remco & Wieling, Martijn (2022). Estimating the level and direction of aggregated sound change of dialects in the northern Netherlands. *Taal & Tongval*, *74*(2), 183-214.

Elmentaler, Michael & Borchert, Felix (2012). Niederdeutsche Syntax im Spannungsfeld von Kodex und Sprachpraxis. In Robert Langhanke, Kristian Berg, Michael Elmentaler & Jörg Peters (Eds.), *Germanistische Linguistik 220 – Niederdeutsche Syntax* (101-135). Georg Olms Verlag.

Lameli, Alfred (2016). Raumstrukturen im Niederdeutschen. Eine Re-Analyse der Wenkerdaten. *Niederdeutsches Jahrbuch .Jahrbuch des Vereins für niederdeutsche Sprachforschung, 139*, 131-152.

ReN-Team (2019). *Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200-1650), 2019*. URL http://hdl.handle.net/11022/0000-0007-D829-8. Archived in Hamburger Zentrum für Sprachkorpora. Version 1.0. Publication date 2019-08-14.

Schröder, Ingrid (2004). Niederdeutsch in der Gegenwart: Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher (Ed.), *Niederdeutsche Sprache und Literatur der Gegenwart* (35-97). Georg Olms Verlag.

Siewert, Janine, Scherrer, Yves, Wieling, Martijn & Tiedemann, Jörg (2020). LSDC - a comprehensive dataset for Low Saxon Dialect Classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects* (25-35). International Committee on Computational Linguistics (ICCL).

Acknowledgements/Disclaimers

This work has been supported by the Academy of Finland through project No.~342859 “CorCoDial – Corpus-based computational dialectology”.

Linguistic practices and identity construction on social media: Italian Americans on Instagram

Federica Silvestri
University of Catania, Italy
federica.silvestri@phd.unict.it

Paper Abstract

Between the 19th and the 20th centuries, Italian migrants who arrived in the United States built their identity through the cultural encounter with their foster society (Michaud, 2011). Nowadays, a significant number of younger Italian Americans have shown greater interest in their heritage (De Fina, 2014), and many have turned to social media to engage with it. Since social media facilitate exchanges within ethnic communities (Dekker et al. 2015), this work aims to test the hypothesis that Italian Americans deploy linguistic practices on social media to engage with their heritage and shape their ethnic identity (Longo, 2023). In this initial phase of this qualitative research, to deal with a manageable yet representative pilot dataset, I gathered and examined 250 comments posted by Italian American Instagram users under posts shared by accounts discussing Italian American culture. To construct my dataset, starting from March 31st, 2024, I clicked on the first five accounts listed after having typed "Italian American" into the Instagram search bar. I, then, opened the ten most recent posts for each account; finally, I selected the first five comments for each post and ended up with a total of 250 comments, which were examined by means of a linguistic-pragmatic approach. At this stage of my research project, being a pilot study, I preferred focusing on methodology prioritising each step of the analysis over quantity reliance, therefore I selected and worked on 250 comments only. The analysis of the data has shown three main aspects, which will be corroborated (or not) by the analysis of a larger dataset. Firstly, all the Italian Americans whose comments were examined are native speakers of English as L1, thus showing the effects of the language shift that characterised the Italian American experience between the first and the second generation and that caused members of the latter to lose proficiency in their heritage languages (Haller, 2011). Secondly, in spite of their limited skills in either Italian or their parents' and grandparents' regional dialects, many Italian American Instagram users proudly defend their heritage, and some even share their families' experiences after their arrival in the U.S., thus shaping their own identity via their interactional linguistic practices (Bucholtz & Hall, 2005) and through their ancestors' experiences. Thirdly, some features typical of online language (Yus, 2011) are used by Italian Americans on Instagram not only to communicate with one another but also to signal allegiance to their roots.

Keywords: digital ethnography, ethnic identity, Italian Americans, social media language, sociolinguistics

REFERENCES

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, *7*(4-5), 585–614. https://doi.org/10.1177/1461445605054407

De Fina, A. (2014). Language and identities in US communities of Italian origin. *Forum Italicum*, *48*(2), 253–267. http://dx.doi.org/10.1177/0014585814529227

Dekker, R., Belabas, W., & Scholten, P. (2015). Interethnic Contact Online: Contextualising the implications of social media use by second-generation migrant youth. *Journal of Intercultural Studies*, *36*(4), 450–467. https://doi.org/10.1080/07256868.2015.1049981

Haller, H. W. (2011). Varieties, use and attitudes of Italians in the U.S.: The dynamics of an immigrant language through time, in T. Stehl (Ed.), *Sprachen in mobilisierten Kulturen: Aspekte der Migrationsinguistik* (pp. 57–70). Universitätsverlag Potsdam. https://d-nb.info/1218860448/34

Longo, S. (2023). eEthnicity: Social media, Italian Americans, and Cultural Identity. *Proceedings of The World Conference on Social Sciences*, *2*(1), 24–44. https://doi.org/10.33422/worldcss.v2i1.98

Michaud, M. C. (2011). The Italians in America, from transculturation to identity renegotiation. *Diasporas*, *19*, 41–51. https://doi.org/10.4000/diasporas.1788

Yus, B. (2011). *Cyberpragmatics. Internet-mediated communication in context*. John Benjamins. https://doi.org/10.1075/pbns.213

The image of Chinese international students on Chinese social media *Zhihu*

Yanni SUN
King's College London, China
K2368699@kcl.ac.uk

Abstract

COVID-19 and the surging nationalism and populism sentiments made Chinese international students (CIS) targets of online vigilantism on Chinese social media and they face alienation in the homeland (Feng, 2020; Gao, 2022) apart from discrimination overseas (Russell, 2020). To obtain comprehensive understanding about this issue, this research examines the image of CIS in a comparatively large corpus of Chinese social media discourses through a corpus-based critical discourse approach facilitated by statistical analysis. The Discourse-Historical Approach (Wodak, 2015) commonly used for media presentation studies is adopted to theorise and categorise the findings.

328 posts of 280995 Chinese characters on a major Chinese social media *Zhihu* were collected. Major referential expressions of CIS were identified and classified by browsing the general word and keyword lists and examining their concordances (Table 1). Predication analysis based on concordances of the most frequent referential expressions of CIS in the corpus *留学生(们)international student(s)*. I compared comments from CIS with comments from other *Zhihu* users to reveal in- and out-group differences. Chi-square tests were conducted to identify significant differences in their use of referential and predication expressions.

Table 1 Exemplary referential expressions of CIS

Category		**Examples**
Victim		受害者 victim;难民 refugee; 苦命人 luckless people
Trouble or degenerate		人渣 scum; 巨婴 giant baby
Meritocracy		人才 talent;无名之辈 nobody; 废物 trash
Nationalism	Real Chinese	中国人 Chinese;同胞 compatriot;爱国主义者 patriot
	Fake Chinese	恨国党 the gang who hate the country;大英帝国的子民 subjects of the British empire
Populism	The privileged	富二代 the rich second generation; 精英 elite;
	Ordinary people	普通人 ordinary people; (人)民 the people

It is found CIS were alienated and stigmatised as the problematic "other" through frames of trouble or degenerate, meritocracy, nationalism, populism, collectivism, and misogyny in the corpus though some comments try to challenge those frames and depict CIS as well-behaved people, victims, the socio-culturally marginalized, patriots, ordinary people without privileges or high socioeconomic status, talents, individuals with rights, and cosmopolitans. Comparative analysis of comments from CIS and other *Zhihu* users reveals both groups use stigmatising discourses against CIS. Apparently, tensions not only exist between CIS and non-CIS but also wihtin CIS. The major difference is that CIS are more likely to object to the "trouble or degenerate" and "meritocracy" frames, present CIS as "socio-culturally marginalized or isolated", recount reverse culture shocks for CIS, and depict CIS as cosmopolitans while non-CIS group is more likely to oppose the "victim" frame, stigmatize CIS as trouble or degenerates, position them in a meritocratic hierarchy, and perceive them from a collectivism (especially pro-collectivism) stance.

Keywords: Corpus linguistics; DHA; media image; Chinese international students; Covid-19

REFERENCES

Feng, Z. (2020). Being a Chinese student in the US: 'Neither the US nor China wants us. BBC News. https://www.bbc.com/news/world-us-canada-53573289

Gao, Z. (2022). Political identities of Chinese international students: Patterns and change in transnational space. *International Journal of Psychology*, *57*(4), 475-482.

Russell, A. (2020, March 17). The rise of coronavirus hate crimes. *The New Yorker*.https://www.newyorker.com/news/letter-from-the-uk/the-rise-of-coronavirus-hate-crimes

Automatic Language Proficiency Assessment of Written Texts: Training a CEFR classifier in L2-Finnish

Jenny Tarvainen
University of Jyväskylä, Finland
jenny.h.tarvainen@jyu.fi

Ida Toivanen
University of Jyväskylä, Finland
ida.m.toivanen@jyu.fi

Ari Huhta
University of Jyväskylä, Finland
ari.huhta@jyu.fi

Presentation Abstract

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), is a framework commonly used to assess the proficiency level of language learners (e.g. Martiyk & Noijons, 2007). It is also utilized for the assessment of language proficiency for citizenship purposes in Finland (Rocca et al.., 2020). To study the suitability of using a deep learning model for the CEFR classification task, we develop and present a language proficiency classifier for Finnish as a second language (F2) written texts. The classifier has been trained to recognize the six Common European Framework of Reference (CEFR) proficiency levels from A1 (basic user) to C2 (proficient user). During the development process, we seek answers to the following questions:

1. Is there enough Finnish learner language data for training deep learning models?

2. Training with existing datasets, how well can a deep learning model detect different CEFR levels?

3. How does the model compare to other CEFR-models?

The FinBERT (Virtanen et al., 2019) language model has been further trained with the datasets of (1) the International Corpus of Learner Finnish (ICLFI), (2) The Advanced Finnish Learner's Corpus (LAS2), (3) subcorpus of young Finnish learners in the Cefling project, and (4) The Finnish Subcorpus of Topling - Paths in Second Language Acquisition. These datasets provide a volume of c. 8000 texts, combining to c. 1.5 million tokens.

After training of models, the best accuracy we obtain is a score of 76.8 %. Accuracy is calculated by dividing the number of correctly classified samples by the total number of samples (see, e.g., Tharwat, 2020), so for instance, 100 % accuracy would mean that the model classified all the data samples correctly. The results indicate that there is room for improvement in model performance and a need for more CEFR-annotated Finnish learner language training data. However, the wrong classifications were mostly only off by one proficiency level (e.g., if the annotation should have been A2, the classifier should label it as A1 or B1). One should also keep in mind that even human assessors do not always agree, and a similar one-off phenomenon could also present itself (see, e.g. Yancey et al., 2023).

Keyword: Automatic Writing Assessment, CEFR, Classifier, L2 Proficiency, LLM

REFERENCES

Martinyuk, W. & Noijons, J. (2007). The use of the CEFR at national level in the Council of Europe Member States. *The Common European Framework of Reference for Languges (CEFR) and the development of language policies: challenges and responsibilities, Strasbourg, 6-8 February 2007.* Council of Europe.

Rocca, L., Carlsen, C. H, & Deygers, B. (2020). *Linguistic Integration of adult migrants: requirements and learning opportunities. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants.* Council of Europe.

Tharwat, A. (2020). Classification assessment methods. *Applied computing and informatics 17*(1),168-192. DOI:10.1016/j.aci.2018.08.003

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint* arXiv:1912.07076.

Yancey, K., Laflair, P., Verardi, G. A., & Burstein, J. (2023). Rating Short L2 Essays on the CEFR Scale With GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576-584). Association for Computational Linguistics.

Can the archives become as cool as a museum? Data Circulation in the Budapest Time Machine

Creative stakeholders & citizen scientist around the Budapest City Archives

Ágnes Telek
Budapest City Archives
telek@bparchiv.hu

Paper Abstract

The Data is power and not only for present information. Users of all kinds turn to archives as a foundation nowadays, entitled to access the data and archival documents they need. The research room of Budapest City Archives (BCA) manages 3000 researchers per year. 56-58% of them are interested in architectural sources of various types, therefore this phenomenon needs to be addressed urgently. Our online databases and digitized materials give smooth access to the citizen scientists and average users to discover stories & histories of the city and its buildings. Many of them don't even have to visit us in person. A huge joint database offers several Hungarian cultural institutions' materials on a very user-friendly surface. Hungaricana is not only a home for databases and digitized cultural heritage but gives us an opportunity to synchronize and unify different layers of data on the same page at the same time with the help of georeferenced historical map layers. The idea of the Time Machine is to travel back to the past just as we do nowadays with the help of Google Street View. This requires a huge amount of work, but we have taken a few steps on the road already: switching between the different time layers offered by the historical maps presents a great opportunity to discover the changes in the structure of the city. We also added four types of georeferenced archival documents, so anyone can reveal the former inhabitants or shops of the building, or to check the original drawings & some vintage postcards, photographs of the neighborhood. Many of the latter have been annotated by AI. 3D reconstructions of some historical buildings are already available on the site, so we can compare different states of the city and the building density. Offering easy access to our materials, some of our stakeholders were able to start building their own databases and create new documents based on our collection. The recurring cooperation with the Budapest100 (celebration of 100-year-old houses with volunteer researchers) and the ÓE YBL Architecture Faculty lead us to an agreement on offering the digital copies from our materials in exchange for the newborn documents they create. We will archive the house data sheets and the 3D building models and publish them as well, which creates "data circulation". In the framework of our joint project, "City Memories" we are working together with two city archives, Stockholm and Copenhagen, to bring the architectural archives closer to the wider audience, sharing our experiences int Best Parctice Guides.

Keywords: 3D modelling, archives, database, architecture, urban history, data circulation, historical maps, university students, volunteer researchers, new born digital documents

Finding Patterns across Multiple Time Series Datasets: Democracy in the Twentieth-century Political Discourses in the United Kingdom, Sweden, and Finland

Risto Turunen
University of Jyväskylä, Finland
risto.j.turunen@jyu.fi

Pasi Ihalainen
University of Jyväskylä, Finland
risto.j.turunen@jyu.fi

Abstract

This paper analyses the contextual variation of nouns and adjectives related to democracy in the United Kingdom, Sweden, and Finland in the twentieth century. We compare parliamentary data (*Hansard*, *Riksdag*, and *Eduskunta*) against press data (UK: *Guardian* and *Times*, Sweden: *Dagens Nyheter* and *Svenska Dagbladet*, Finland: *Helsingin Sanomat* and *Suomen Kuvalehti*). While our parliamentary datasets (Ihalainen et al. 2022) encompass several political ideologies simultaneously, the selected newspapers can broadly be categorized into conservative and liberal strands. By including both newspapers with diverse political leanings as well as parliamentary speeches, our study offers a fresh perspective on the relation between democratic discourses produced by politicians and journalists.

The approach includes visualizing the main similarities and differences in the use of democratic vocabulary between multiple historical time series datasets, as well as applying cross-correlation analysis to automatically find identical patterns between parliament and media or across different nations. The similarity of various word frequency time series charts is evaluated using the Pearson correlation coefficient (PCC), which can vary from -1 to 1. A value of 1 indicates a perfect positive correlation, where every increase in word frequency in dataset A is matched by a simultaneous increase in dataset B. Conversely, a value of -1 indicates a perfect negative correlation, where every increase in word frequency in dataset A corresponds to a simultaneous decrease in dataset B. The closer the PCC values are to 0, the weaker the relationship between the two variables (Derrick & Thomas 2004). The strengths of the PCC are its mathematical simplicity, easy interpretability, and tolerance for noise, while its main limitation is sensitivity to extreme outliers which can be mitigated by identifying and addressing outliers before conducting analysis.

Our findings indicate that the cross-correlation is strongest between similar political terms in the same dataset, e.g., the relative frequency of “democracy” and “democratic” over time in a national parliament (in *Hansard* 0.91, *Riksdag* 0.76, and *Eduskunta* 0.65). Another strong set of cross-correlations can be observed when the same political term appears in different datasets from the same country, e.g., the frequency of “democracy” in liberal and conservative press (in the UK 0.87, in Sweden 0.82, and 0.61 in Finland). The

most important finding from a historical viewpoint is the statistically strong cross-correlation between media and parliamentary discourses, with values ranging from 0.55 to 0.76 for the term “democracy”. Transnational correlations of political terms were not as strong as intra-national correlations, but they were clearly evident in the PCC values, e.g., for the frequency of “democracy” they varied from 0.58 to 0.68 between three parliaments under investigation. The shared patterns between three parliamentary democracies include general increase in the use of “democracy” over time, with notable peaks in the 1930s as a reaction to totalitarianism, around the year 1968 related to the rise of social movements, and in the early 1990s with the fall of the Eastern bloc.

Keywords: newspapers, parliamentary speeches, text mining, time series

REFERENCES

Derrick, T., & Thomas, J. (2004). Time series analysis: The cross-correlation function. In N. Stergiou (Ed.), *Innovative Analyses of Human Movement* (pp. 189–205). Human Kinetics Publishers.

Ihalainen, P., Janssen, B., Marjanen, J., & Vaara, V. (2022). Building and testing a comparative interface on Northwest European historical parliamentary debates: Relative term frequency analysis of British representative democracy. In *Digital Parliamentary Data in Action* (pp. 52–68). CEUR Workshop Proceedings, Vol. 3133. http://ceur-ws.org/Vol-3133/paper04.pdf.

Wevers, M., Gao, J., & Nielbo, K. (2020). Tracking the consumption junction: Temporal dependencies between articles and advertisements in Dutch newspapers. *Digital Humanities Quarterly*, 14(2). http://www.digitalhumanities.org/dhq/vol/14/2/000445/000445.html

Hiking with Machine's Eyes: A Computer Vision Exploration of Nature Photography in Instagram

Juhana Venäläinen
University of Eastern Finland
juhana.venalainen@uef.fi

Paper Abstract

This paper explores the methodological and epistemological implications of using computer vision to analyse visual representations of Finnish recreational nature sites in social media. The study focuses on the Instagram imagery of two nature sites in Finland, while also being informed by ethnographic walking interviews that focus on how the uses of digital media transform the representations and experiences of nature and its fragility. By combining and contrasting machine learning techniques with qualitative inquiry (cf. Maltezos et al., 2024), the study aims at making sense of how the complex interplay between algorithmic visual cultures and the quotidian uses of technology shapes our environmental relations.

Through scraping the Instagram API with a hashtag-based approach, a large dataset of images was collected about the two fieldwork sites: Patvinsuo National Park in Lieksa, Finland, and Viiankiaapa Mire Reserve in Sodankylä, Finland. The images were analysed using Google's Inception v3 API for image embeddings and further through unsupervised machine learning methods (hierarchical clustering, principal component analysis) in Orange data mining platform. These methods facilitated constructing a visual taxonomy of nature representations as well as a set of dichotomous factors that supposedly describe a part of the dataset's variance as captured by the embedding algorithm. This visual taxonomy highlights AI's proficiency in object detection, while the categorisations of landscapes images were harder to interpret in a cultural context. Notably, the process of hierarchical clustering creates pairings of which some are predictable but others very unexpected ("selfies and canoes"), challenging us to consider the embedded values, assumptions, and often invisible data labour that shape AI's understanding (cf. Denton et al., 2021; Carah et al., 2022).

The study asserts that the proliferation of digital photography on social media in combination with ethnographic approaches provides a rich basis for exploring how boundaries between virtual and on-site nature are currently being blurred, and how this entanglement transforms our relations with nature. Algorithms not only categorise but co-create our visual digital cultures, and thus there is a need to critically assess their underlying tendencies and biases. The research also underscores the AI's methodological limitations in visual content analysis. While AI offers an efficient method to manage and categorise large image datasets, the interpretative nuance of human analysis remains essential, particularly

for contextually rich images. A mixed-method approach can thus yield a more holistic understanding of nature's digital representations.

Keywords: nature imagery, computer vision, machine learning, ethnography, social media studies, algorithmic cultures, cultural representation, visual culture

REFERENCES

Carah, N., Angus, D., & Burgess, J. (2022). *Tuning machines: an approach to exploring how Instagram's machine vision operates on and through digital media's participatory visual cultures*. Cultural Studies, 36(3), 456-478. https://doi.org/10.1080/09502386.2022.2042578

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). *On the genealogy of machine learning datasets: A critical history of ImageNet*. Big Data & Society, 8(1), 1-15. https://doi.org/10.1177/20539517211035955

Maltezos, V., Luhtakallio, E., & Meriluoto, T. (2024). *Bridging ethnography and AI: a reciprocal methodology for studying visual political action*. International Journal of Social Research Methodology, 27(2), 234-249. https://doi.org/10.1080/13645579.2024.2330057

IN SEARCH OF THE INVISIBLE GPT in An Investigation of Hidden Semantic Information

Katarzyna Wiśniewska, PhD
University of Eastern Finand
University of Rijeka
katarzyna.wisniewska@uef.fi

Bedikt Prak,Dr.
University of Rijeka

Paper Abstract

The surge of interest in artificial intelligence (AI) stems from the remarkable fluency demonstrated by Large Language Models (LLMs) in language comprehension. Our study focuses on uncovering hidden semantic information, particularly how OpenAI's Generative Pre-trained Transformer (GPT) 4.0 model excels in this area. Informed by documented obstacles in neural machine translation (MT) (Koehn and Knowles, 2017; Wan et al., 2022), challenges in text-generation models (Wang et al., 2023), recent advancements in leveraging LLMs for natural language processing (NLP) tasks (Riemenschneider and Frank, 2023) and human-like translation strategies (He et al., 2024), our aim to evaluate GPT's proficiency in discerning and articulating semantic nuances, comparing its performance with human judgement. Employing a comparative analysis framework, our study scrutinises a selection of translated sentences from literary and audiovisual texts that have been extensively studied within the framework of Talmy's (2000) force dynamics by Wiśniewska (2022 and 2023). This conceptual framework illuminates how language expresses causality and dynamic relationships between entities, often using the metaphor of physical forces acting on objects. For instance, in the sentence She persuaded him to go, the force exerted by she leads to the action of him going. Methodologically, our approach is multifaceted, integrating quantitative measures with qualitative analysis to assess the fidelity of translations in capturing hidden semantic information. Custom prompts are employed to elicit translation and self-evaluation of GPT output as metadata. We explore relationships between English, Finnish, and Polish, with a focus on verb phrases and sentences embedded with force dynamics meanings. Evaluation encompasses idiomatic correctness and language conventions, providing technical details alongside a nuanced discussion of semantics in JSON format. Anticipated findings indicate that while GPT demonstrates remarkable proficiency in rendering surface-level meanings, its ability to identify and articulate hidden semantic information varies based on linguistic context, complexity, and the specificity of prompts. Human translation evaluators display a more nuanced understanding, leveraging linguistic intuition to analyse translations rich in hidden meanings. This study contributes to the growing body of research on AI-assisted translation by shedding light on the capabilities and limitations of LLMs in uncovering implied semantic information. By elucidating the interplay between human translators,

AI, and MT systems, this research advances our understanding of translation studies in the digital age of NLP. Drawing on insights from studies such as Alzahrani et al. (2024), it emphasises the importance of rigorous methodology and careful consideration of evaluation metrics in assessing the performance of translation models.

REFERENCES

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Aloitaibi, N., Altwairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. (2024). "When benchmarks are targets: Revealing the sensitivity of Large Language Model leaderboards." ArXiv. Preprint. arXiv.2402.01781v1. February 1, 2024 [access: 11/03/2024].

He, Z., Lian, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang X. (2024). "Exploring human-like translation strategy with Large Language Models." In: Transactions of the Association for Computational Linguistics. pp. 229–246.

Koehn, P. and Knowles, R. (2017). "Six challenges for neural machine translation." In: Proceedings of the First Workshop on Neural Machine Translation. Vancouver, Canada. August 4, 2017. Association for Computational Linguistics, pp. 28–39.

Riemenschneider, F. and Frank, A. (2023). "Exploring Large Language Models for classical philology." ArXiv. Preprint. arXiv:2305.13698. May 23, 2023 [access: 10/04/2024].

Talmy, L. (2000). Toward a Cognitive Semantics: Vol. 1. Concept Structuring Systems. Cambridge: The MIT Press.

Wan, Y., Yang, B., Wong, D. F., Chao, L. S., Yao, L., Zhang, H., and Chen, B. (2022). "Challenges of neural machine translation for short texts". In: Computational Linguistics, 48 (2). pp. 321–342.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., and Wang, G. (2023). "GPT-NER: Named entity recognition via Large Language Models." ArXiv. Preprint. arXiv:2304.10428. October 7, 2023 [access: 10/04/2024].

Wiśniewska, K. (2022). Description of Force Dynamics and Cognitive Retention in Literary and Audiovisual Translation. Unpublished manuscript. PhD thesis. University of Eastern Finland, Joensuu, Finland.

Wiśniewska, K. (2023). Understanding Cognitive Retention in Translation: An Exploration of a Descriptive Tool Focusing on Cognitive Semantics of Force Dynamics. Manuscript submitted for publication.

Abstracts:

Posters

The three universities' cooperated management studies in the specialist training in medicine and dentistry

Mirkka Forssell, Marjo Tourula, Anna-Liisa Suominen, Hanna Tenhunen, Elias Vaattovaara
Tampere University
mirkka.forssell@tuni.fi

Poster Abstract

Specialist training in medicine and dentistry in Finland includes a compulsory ten-credit course in management. National teaching modules are human resource management, communication, structure, organization, legislation, and financing of social and health care services. In 2024, three universities (UEF, Oulu, and Tampere) entitled to provide specialist training organized a pilot during which each university provided a one-day webinar common to all participants. Each university is responsible for hosting one of the three webinars. The aim of the research The research aims to get information about the experiences of doctors and dentists in specialist training from the joint webinars at three universities. The purpose is to develop management training to meet the changing needs of the healthcare service system. Methodology in-brief In this research, information is collected using a structured questionnaire from specializing doctors and dentists in connection with three pilot sessions, which will be carried out from February 2024 to December 2024. Participation in the study is voluntary. There were 172/296 participants aged from under 30 to 65 who answered the first questionnaire. The structured questionnaire contained multiple-choice questions and open questions. The open questions have been analyzed using the Affinity diagram method. Preliminary findings The preliminary research results showed issues related to accessibility occurred during the webinar day. Despite the accessibility issues, 86% of the participants considered the lecture implemented as a webinar to be the best arrangement, and only 10% on-site teaching and 3% independent study online. Of the respondents, 52% reported problems with sound quality and audibility, and 19% with network connections. The open-ended answers to the cause of why students report webinars as the best lecture option, analyzed with an affinity diagram, revealed five clusters: no need to commute, flexibility, interaction between participants, work and family life balance, and ability to concentrate. The research revealed that 77% of participants considered the engaging and most educative part of the webinar to be the discussion with other participants who were geographically located in different parts of Finland.

Keywords: management, medicine, dentistry, sustainability

REFERENCES

1. https://www.laaketieteelliset.fi/ammatillinen-jatkokoulutus/johtamisopinnot

2. Lucero, A. (2015). Using Affinity Diagrams to Evaluate Interactive Prototypes. In: Abascal, J., Barbosa, S., Fetter, M., Gross, T., Palanque, P., Winckler, M. (eds) Human-Computer Interaction – INTERACT 2015. INTERACT 2015. Lecture Notes in Computer Science, vol 9297. Springer, Cham. https://doi.org/10.1007/978-3-319-22668-2_19

Data Sources for Automatic Classification and Analysis of Texts from Egyptian Antiquity

Tommi Jauhiainen, Heidi Jauhiainen, Marja Vierros
University of Helsinki, Finland
firstname.surname@helsinki.fi

Erik Henriksson
University of Turku, Finland
firstname.surname@utu.fi

Poster Abstract

In this poster, we present the aims and the current state of the research project "Automatic Classification and Analysis of Texts from Egyptian Antiquity", funded by the Kone Foundation.

In short, the project aims to develop new state-of-the-art language technological methods for automatically processing textual documents from Egypt dating from the 8th century BCE to the Arab conquest in the 7th century CE. The project investigates the extensive textual evidence from the region as a whole, including the texts in both the Greek and the Egyptian languages. We aim to develop new and improved methods for automatically identifying languages and dialects (Jauhiainen et al. 2019) and detecting the text's date and place of origin (Jauhiainen et al. 2023). Furthermore, one of our goals is to build automated methods for finding loan words between languages by creating converters between all the target languages' transliteration systems (Jauhiainen & Jauhiainen 2023).

A large part of the project is dedicated to collaboration between the project and various entities that own the copyright to the existing machine-readable texts within the focus of the research (Jauhiainen 2022). We will identify sources for machine-readable texts pertinent to our study and, if they are not openly available, negotiate with the rightsholders for suitable access to the texts to use in the project.

We will create a database of all sources where relevant machine-readable text collections are available. The listing will be openly available on the project's website and updated throughout the project's lifespan. We will contact the entities and persons behind the text collections and aim to get the data as exports from their system instead of reverting to methods like web scraping. We have already identified several sources for texts that are usable by the project.

For the texts primarily written in Greek, we use all the transcribed texts available through the Papyri.info project as our data. Currently, the papyri.info collection contains metadata for over 100,000 texts, of which more than 50,000 are transcribed. In addition to the document data from papyri.info, we already have access to several thousand inscriptional texts from the Packard Humanities Institute's collection.

Thesaurus Linguae Aegyptiae (TLA), a digital publication platform, includes machine-readable texts written in Egyptian using either Hieroglyphic, Hieratic, or Demotic scripts. The TLA is the largest ongoing project collecting and publishing machine-readable ancient Egyptian texts, and their collection is continuously increasing. We expect the latest

form of the logographic Egyptian writing, Demotic, to be most interesting regarding language contact, as it was used while the Greeks ruled Egypt.

Keywords: corpora, egyptology, language identification, multilinguality, papyrology

REFERENCES

Jauhiainen, H. (2022). Encoding Hieroglyphic Texts. In K. Berglund, M. La Mela, & I. Zwart (Eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)* (pp. 244-250). Article 22 (CEUR Workshop Proceedings ; Vol. 3232). CEUR-WS.org.

Jauhiainen, H., & Jauhiainen, T. (2023). Transliteration Model for Egyptian Words. In A. Rockenberger, J. Tiemann, & S. Gilbert (Eds.), *DHNB2023 : Conference Proceedings* (pp. 149-164). (Digital Humanities in the Nordic and Baltic Countries Publications; Vol. 5, No. 1). University of Oslo Library. https://doi.org/10.5617/dhnbpub.10659

Jauhiainen, T., Henriksson, E., Vierros, M., & Jauhiainen, H. (2023). *Automatic detection of place and time for Greek texts in Egypt*. Poster session presented at International Congress of Egyptologists, Leiden, Netherlands.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). *Automatic Language Identification in Texts: A Survey*. Journal of Artificial Intelligence Research, 65, 675-782. https://doi.org/10.1613/jair.1.11675

Automated tool for sharing experiential knowledge: the case of Human Science section in the Digital Citizen Science Center of University of Jyväskylä

Anna Kajander, Eerika Koskinen-Koivisto
University of Jyväskylä
anna.k.kajander@jyu.fi

Poster Abstract

The Digital Citizen Science Centre is a multidisciplinary project that aims to develop mobile applications for the needs of different citizen science projects. It consists of four research groups from four faculties which work in collaboration with the digital services and the Centre of Open Knowledge of the University of Jyväskylä. The technology is based on a mobile application Research for JYU Mobile (RFJM).

This poster introduces the Human science section of the project which is focused on research of everyday experiences and meaningful places of nature. We will create a pilot model of an AI based and automated mobile tool for interviewing and collecting qualitative research data. We will analyze the everyday experiences and interactions with nature of different kinds, which could mean various elements and locations, such as forests, mires, plants, animals as well as urban environments such as parks etc. The application developed in the project will provide citizens with an opportunity to share their in-situ experiences of encountering nature and explore other users' experiences of different natural sites.

Our aim in the human science project is to explore how an automated tool is accepted and used by citizens and how large quantities of qualitative data could be analyzed and used in research of affective everyday experiences. We will also consider the ethical challenges of creating such experiential data and motivating citizens in using the application to share their experiences.

Towns on the Eastern Border of the Swedish Great Power Power Structures and Commercial Networks in the Second Half of the 17th Century

Kimmo Katajala, Jenni Merovuo, Antti Härkönen
University of Eastern Finland
kimmo.katajala@uef.fi

Poster Abstract

The Towns on the Eastern Border of the Swedish Great Power project explores the power and trade networks of five cities in the late 17th century. Nyen, Kexholm, Sortavala, Brahea and Kajaani located closest to the eastern border. Despite their small size, they were important commercial centers of their regions. The project examines interaction within and between cities, but also connections with other regions of the Swedish Empire and Russia. The aim is to bring a comparative perspective to microhistorical research, enabled by new digital methods. In cooperation with the National Archives of Finland, HTR technology is used to produce a digitized corpus of sources that is analysed with digital network analysis tools. Network analysis, comparison of cities and a close reading of sources characteristic of microhistory are expected to reveal new features of the patterns of behaviour of urban communities in the early modern era.

The workflow of the project is as follows. The project has about 700 pages of machine-readable material ready for teaching the HTR-program's algorithm. The approximately 10,000 document pages in the audited cities are interpreted into an electronically readable format with the help of the Transkribus program in the National Archives of Finland. Coordinates can be added to the XML file format, for example, for place names where the detected point of text is located in the machinescanned image. When creating the observation matrices, the PDF file format enables search functions based on OCR technology.

Observation matrices are created with Excel from the corpus for network analysis. Visone, which operates in a Java environment, has been chosen as the network analysis tool. The internal and external interactions of cities are primarily seen as explicit social networks. Through their descriptions, the systemic meanings of interaction and dependence are sought. Visual network descriptions produced with the Visone application may be accompanied by attributes that describe time and place (geographic coordinates).

The data can be combined with a digital map model, whereby the networks are also presented as distances to nodes that are connected to each other (links). The directions of interaction (knot, ingrade, or out-grade) and strength (intensity) may also be included in the review. From the city, place or group under consideration (hub), it is possible to determine its characteristics in the network, i.e. whether it is a "star" in the middle of the

Practical solutions for digitally administering and scoring of a children's speechreading test

Jaakko Kauramäki[1,2], Satu Saalasti[1,3], Kerttu Huttunen[1]
[1]*University of Oulu, Finland*
[2]*University of Helsinki, Finland*
[3]*University of Eastern Finland, Finland*
(jaakko.kauramaki@helsinki.fi)

Paper and Poster Abstract

Use of visual information about speech is pronounced in situations in which auditory information is degraded because of, for example, background noise or reverberation. In speechreading (often also called lip reading), information about lip, jaw and tongue movements but also the visual cues of facial expressions are used to perceive the message of a speaker. People with hearing loss try to make use of speechreading for complementing the insufficient auditory information caused by their hearing problem, but interindividual differences are large and reliable assessment methods are needed.

Aims of the research project *Gaze on lips?* (https://www.oulu.fi/en/projects/gaze-lips) are twofold: to construct, standardize and validate the pre-recorded Speechreading Test for Finnish Children (SPETFIC; Huttunen & Saalasti, 2023) with automatized scoring and to find out the developmental trajectory of primary school-age children's speechreading skills. Data are currently being compiled from 8- to 11-year-old hearing children to have the age norms for the SPETFIC.

In the current presentation, we report the practical solutions of administering SPETFIC both in-person and remotely. Remotely collected samples are managed by REDCap tools hosted at the University of Oulu, Finland. REDCap (Research Electronic Data Capture; Harris et al. 2009; 2019) is a secure, web-based software platform designed to support data collection. We implemented remote testing by utilizing screen sharing of a Zoom meeting (Zoom, 2024) so that SPETFIC is run on a test administrator's computer. For testing the stability and speed of the Internet connection and the capabilities of screen sharing of Zoom video call, a specific frame drop estimation test was constructed. If there were issues causing excessive frame dropping, a one-time direct access link to the REDCap running SPETFIC was conveyed to the child via chat channel of the Zoom.

SPETFIC includes the automatic scoring of the results, shown both as total and section specific score on screen after finishing the test. Additionally, the item-by-item and summary results of the test can be downloaded as comma-separated (CSV) files. After the validation phase, speech and language therapists testing the children at clinics can choose to separately administer either the section A (easier words), the section B (more difficult words) or the sentence section. Having these sections enables tapping of a fairly wide skill spectrum and following up of skill improvement along the child's maturation and

intervention. Automatic scoring rules out scoring errors in both research and clinical use of the test.

Keywords: Finnish language, lipreading, online testing, speechreading, visual speech processing.

REFERENCES

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377-381. https://doi.org/10.1016/j.jbi.2008.08.010

Huttunen, K., & Saalasti, S. (2023). *Lasten huuliolukutesti (Speechreading Test for Finnish Children)* [Unpublished; test in validation phase].

Zoom. (2024). Zoom (v6.1.0) [Software]. Retrieved from https://zoom.us/

Letter_2_Santa.py – Tapping Big Data from the Arctic Circle

Gabriele Liebert, Sandra Reimann, Michael Rieβler, Amelie Tahedl
University of Eastern Finland
michael.riessler@uef.fi

Poster Abstract

Our poster presents the results of a pilot project which aims at building the Santa Claus Letter Corpus. These letters – sent to Finland from around the world – feature text and art, mostly handwriting enriched with drawings. The senders are primarily children. The physical collection at the National Archives contains 25 shelf meters of letters. So far they have been catalogued only in bunches, according to country and year of origin. We have started to examining the collection in 2023, digitised parts of it, enriched the cataloguing metadata, run tests for quantitiative analyses, and carried out first qualitative analyses. Our original focus has been on letters which we expected to be written in either German, Finnish, Swedish, or Russian. But we found out immediately that the language diversity is higher than the sender's poststamp suggests, e.g. letters sent in Finnish from Sweden or in English from Germany. The main results of our pilot were: 1) The documentation of workflows and data standards for digitisation, 2) Preliminary (manual) indexing according to language, artwork, and texttype, 3) Experimenting with computational methods for indexing the letters (format-, language-, and text recognition), 4) Pragmatic analysis of a subset of German-language letters (name anonymised, in press).

Keywords: Finland, Linguistic Data Science, Cultural Studies, Art Education, Pragmatics

REFERENCES

name anonymised (in press) „Briefe an den Weihnachtsmann in Finnland – eine unerforschte Textsorte. Kategorisierung und textpragmatische Auswertung“

network, a “liaison”, a “bridge” or a so-called "gatekeeper". In this way, abstracted network descriptions of the interaction between places and people make it possible to investigate commercial systems between cities and bourgeois on the eastern border and compare networks within and between cities.

The project studies also the different aspects of segregation inside the cities. In segregation research, QGIS geoinformatics software is used for visual representations of the social dimensions of urban space. This is done by creating a digital map from town plan drawings to which various collected real estate data can be added, such as occupation derived from the titles of the plot owners, wealth based on tax information, ethnic background derived from the name (Swedish, Finnish, German, Karelian, Russian) and thus also an assessment of the person's religion. In this way, a picture can be obtained of the spatially ordered social characteristics of urban areas.

Keywords: Digital humanism, handwritten text recognition, network analysis

Handwritten Text Recognition (HTR) model for historical documents from 17th to 20th centuries – Using TrOCR

Marttila Riikka, Joska Sanna, Lipsanen Mikko, Föhr Atte, Jokipii Ilkka
National Archives of Finland
firstname.lastname@kansallisarkisto.fi

Poster - Abstract

Handwritten historical documents remain an important part of research materials in different fields even today. When historical documents are digitized, modern text analysis methods can be applied to make them easier to use and analyse. However, texts must first be recognized from images containing text and converted into machine-readable form, and for a long time this has not been possible for handwritten texts. Without the advancements in (open source) machine learning methods and computational power in recent years, handwritten text recognition on a larger scale would not have been possible (Muehlberger et al. 2019). In a research project conducted at the National Archives of Finland, we utilize the pretrained Transformer-based Optical Character Recognition (TrOCR) model developed by Microsoft (Li et al. 2022). It combines an image Transformer encoder and a text Transformer decoder for optical character recognition, replacing traditional CNN- and RNN-based approaches and eliminating the need for additional language models for post-processing accuracy. TrOCR is pre-trained on synthetic data and fine-tuned on human-labeled datasets, demonstrating superior performance on both printed and handwritten text recognition tasks.

We aim to compare the performance of various HTR models developed specifically for the handwriting styles of individual centuries against a super model trained on a comprehensive dataset from the 1600s to 1900s. The goal is to train HTR models to perform with sufficient accuracy on documents in both Finnish and Swedish languages. As an important part of the strategy of the National Archives of Finland high-performing HTR model development can make handwritten historical documents more accessible and easier to use as source materials in many research fields (Lahtinen & Katajisto 2020, Paju et al. 2020).

This research has access to 26800 pages of annotated data. Annotation here refers to the transcription of texts and the marking lines around text lines. On average, one page consists of 30 lines of text. The data is randomly divided into training, validation, and test datasets and weighted across different centuries and languages (Swedish and Finnish) to ensure a sufficiently representative samples from each century and both languages. The training dataset is used for training the HTR model, while the validation set is automatically used by TrOCR for model validation to identify the best model configuration. The test dataset is used to compare different models against each other. Different models are compared, and model accuracy is evaluated using the Character Error Rate (CER) value.

Keywords: digital humanities, handwritten text recognition, historical research, historical documents, machine learning

REFERENCES

Lahtinen, A. & Katajisto, K. (2020). *Handwritten text recognition opens up a treasure trove of information on Finnish society. In J. Nuorteva & P. Happonen (Eds.),* The National Archives of Finland Strategy 2025: Perspectives for the Future (pp.18-19).

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F. (2022). *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models.* arXiv: 2109.10282v5

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., . . . Zagoris, K. (2019). *Transforming scholarship in the archives through hand-written text recognition: Transkribus as a case study.* Journal of Documentation, Vol. 75 No. 5, pp. 954-976. https://doi.org/10.1108/JD-07-2018-0114

Paju, P., Oiva, M., & Fridlund, M. (2020*)*. Digital and distant histories: Emergent approaches within the new digital history. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 3–18). Helsinki: Helsinki University Press. https://doi.org/10.33134/HUP-5-1

Acknowledgements

We would like to thank everyone who participated the project with special thanks to all the annotators as well as the volunteers and researchers who wrote the transcriptions.

Collecting digital research data using smart devices from deaf and hard of hearing children training speechreading

Minna Mundoli, Reetta Viljanen, Jaakko Kauramäki, Katja Dindar, Satu Saalasti, Kerttu Huttunen
University of Oulu, Finland
University of Eastern Finland
University of Helsinki, Finland
(minna.mundoli@oulu.fi)

Poster Abstract

Deaf and hard of hearing (DHH) people may face challenges in speech recognition during everyday social interaction. Speech processing is multimodal, which means that people do not rely on the auditory channel alone to understand speech, but also benefit significantly from visual cues (Massaro & Jesse, 2012). *Speechreading*, or *lipreading*, is the use of visual speech to aid in speech recognition, especially in DHH people. Speechreading can be strengthened with training, and, lately, there has been growing interest in developing effective computerized training programs (Buchanan-Worster et al., 2021; Pimperton et al., 2019; Tye-Murray et al., 2022). Speechreading training in childhood could increase the benefits for speech recognition, and the systematicity of training could be improved with digital solutions.

We developed a mobile speechreading training application Optic Track (openly available after the research period) and aimed to find out how the amount and quality of using the app is related to the change of the speechreading skills of Finnish DHH children aged 8–11 years. Children participating in the *Gaze on lips?* (https://www.oulu.fi/en/projects/gaze-lips) study use either Android or Apple devices (cellular phones, tablet computers) for training for eight weeks. Users of the app look at videos of people silently speaking single words, sentences and short narratives and accomplish discrimination and matching tasks and compile three-word-sentences. The app provides a versatile set of exercises, such as memory games, bingo games and tasks in which the user selects their response out of two to five alternatives. During the training period, the research version of the app collects data such as the time used for practicing, and the success in accomplishing different tasks.

After the training period, the data collected by Optic Track are transferred to the researcher's device without saving any sensitive data. Data transfer is done either via USB cable or wirelessly. In the wireless option, the data are transferred using a QR-code linked with FileSender, a secure filesharing service. The binary file is then converted into a comma separated (CSV) file to have the data in a more easily readable text format and to allow further processing with statistical programmes.

Preliminary results show that Optic Track provides reliable data that can be used in exploring the relationships between, for example, the training time with the Optic Track app, the skill improvement across time in the tasks the app contains and the possible changes in speechreading skills tested with a speechreading test.

Keywords: intervention, lipreading, mobile application, speechreading, visual speech processing

REFERENCES

Buchanan-Worster, E., Hulme, C., Dennan, R., & MacSweeney, M. (2021). Speechreading in hearing children can be improved by training. *Developmental Science*, *24*(6), e13124. doi: 10.1111/desc.13124.

Massaro, D. W., & Jesse, A. (2012). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.) *The Oxford Handbook of Psycholinguistics* (pp. 19-36). Oxford University press.

Pimperton, H., Kyle, F., Hulme, C., Harris, M., Beedie, I., Ralph-Lewis, A., Worster, E., Rees, R., Donlan, C., & MacSweeney, M. (2019). Computerized speech-reading training for deaf children: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research, 62*(8), 2882-2894. DOI: 10.1044/2019_JSLHR-H-19-0073

Tye-Murray, N., Spehar, B., Sommers, M., Mauzé, E., Barcroft, J., & Grantham, H. (2022). Teaching children with hearing loss to recognize speech: Gains made with computer-based auditory and/or speechreading training. *Ear and Hearing*, *43*(1), 81-91. https://doi.org/10.1097/AUD.0000000000001091

Quantitative and qualitative approach to Finnish Twitter during the Covid-pandemic: Topics, attitudes, and emotions

Jenna Saarni
University of Turku, Finland
jensaay@utu.fi

Otto Tarkka
University of Turku, Finland
ohitar@utu.fi

Poster Abstract

The ways we discuss crises affect our understanding of major events and the world in general (Seeger & Sellnow, 2016). These kinds of discussions can even change our behaviour (Mustafa-Awad & Kirner-Ludwig, 2017); thus the study of crisis communication from a linguistic perspective is essential. During the Covid-19 pandemic, social media became an effective arena for crisis communication, and it brought together different actors from decision-makers and healthcare professionals to ordinary citizens through communication and interaction (Spencer, 2023). However, crises are often events in which people react strongly while they try to understand the situation (Bednarek et al., 2022). On social media platforms, discussions get easily heated when different emotions, experiences, and opinions collide. In this poster presentation, we describe how the global health crisis was represented on a popular microblogging site by addressing the following research questions: (1) What kind of topics are discussed on Finnish Twitter during the Covid-19 pandemic? and (2) What kind of attitudes and emotions are attached to these topics? To answer these questions, we utilise a large corpus of 375,322 tweets in Finnish from January 2020 to August 2021. We adopt a multidisciplinary approach to the data as we use complementary quantitative and qualitative methods that allow us both to examine the data as a vast entity and to explore the linguistic meanings in more detail. First, we use the unsupervised machine learning method of topic modelling to automatically identify topics and keywords attached to them (Blei et al., 2003). Next, we study the attitudes and emotions attached to these topics with the framework of evaluative parameters (Bednarek, 2010). Based on the results, the topic model identified 35 pandemic-related topics that cover, for example, emotions and protective measures in healthcare, briefings and news broadcasts, associations offering support services, masks, and quarantine and infection rates. The analysis of the evaluative parameters shows that expressions of emotivity, mental state, importance and necessity were attached to these topics.

Keywords: Covid-19, crisis communication, discourse analysis, evaluation, topic modelling

REFERENCES

Bednarek, M. (2010). Evaluation in the news. A methodological framework for analysing evaluative language in journalism. *Australian Journal of Communication, 37*(2), 15– 50.

Bednarek, M., Ross, S.A., Boichak, O., Doran, Y. J., Carr, G., Altmann, E. G., & Alexander, T. J. (2022). Winning the discursive struggle? The impact of a significant environmental crisis event on dominant climate discourses on Twitter. *Discourse, Context & Media, 45*, 1–13.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Mustafa-Awad, Z., & Kirner-Ludwig, M. (2017). Arab women in news headlines during the Arab Spring: Image and perception in Germany. *Discourse & Communication, 11*(5), 515– 538.

Seeger, M. W. & Sellnow, T. L. (2016). *Narratives of crisis: Telling stories of ruin and renewal*. Stanford University Press.

Spencer, A. (2023). International communication. In S. M. Croucher & A. Diers-Lawson (Eds.), *Pandemic communication* (pp. 215–230). Routledge.