rijksuniversiteit groningen / faculteit wijsbegeerte

AIM2023
Artificial Intelligence and Minority. Computational Approaches to Multilingual Historical Research

Date & time:
May 12, 2023
10:00 – 17:15

Location:
House of Connections
Grote Markt 21, Groningen, NL
and online (register here)

Program chair:
Raluca Tanasescu
r.a.tanasescu@rug.nl

Organized with the support of:
Jantina Tammes School of Digital Society, Technology and Artificial Intelligence (RUG)

Partners:
Centre for Digital Humanities (RUG)
Multilingualism and Minorities Programme (RUG)

university of groningen
jantina tammes school of digital society, technology & artificial intelligence

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No. 801653 *NaturalPhilosophy*.

erc European Research Council
Established by the European Commission

rijksuniversiteit groningen / faculteit wijsbegeerte

Conference Programme

10:15 – 11:30 – Opening Keynote Address
Christoph Purschke
"Standardization and Tool Development for NLP in Small Language Contexts"

11:35 – 12:00 – Conference Presentation #1
Senja Pollak
"Advanced Natural Language Processing Methods for Analysing LGBTIQ+ and Migration Related Texts"

12:00 – 13:30 – Lunch Break

13:45 – 14:15 – Conference Presentation #2
Chris Tanasescu
"Multilingual Multiplex Corpora and Analytical-Creative HCI Poetry in European Minority Languages"

14:15 – 14:45 – Conference Presentation #3
Martin Žnidaršič
"Natural Language Processing in ClowdFlows"

14:45 – 15:15 – Conference Presentation #4
Silvia Donker, Michiel van der Ree
"Ad Hominem: Creating a Named-Entity Recognition Model for Early Modern Latin"

15:15 – 15:30 – Coffee break

15:30 – 16:00 – Conference Presentation #5
Abigail Shanab, Diana Inkpen, Raluca Tanasescu, Andrea Sangiacomo
"Minor Figures, Major Themes. Multilingual Topic Models in an Early Modern Philosophy Corpus"

16:00 – 17:15 – Closing Keynote Address
Diana Roig-Sanz
"Cultural Diversity. An Ecological Approach to the Study of Minority"

rijksuniversiteit groningen / faculteit wijsbegeerte

Opening keynote presentation
10:15 – 11:30

STANDARDIZATION AND TOOL DEVELOPMENT FOR NLP IN SMALL-LANGUAGE CONTEXTS

Christoph PURSCHKE
University of Luxembourg (Luxembourg)
Culture & Computation Lab
christoph.purschke [at] uni.lu

Luxembourgish presents an interesting case of a small language in different regards. It has seen processes of linguistic Ausbau over the last decades that have led to its (still ongoing) standardization and establishment in the written domain (with the advent of social media). This is flanked by a dedicated language political development plan and a massive societal revaluation of the language. At the same time, Luxembourgish is still poorly implemented in school curricula leading to a low level of rule knowledge in the population.

From an NLP perspective, Luxembourgish has seen little to no tool or resource development until recently. At the same time, there are many different data sets available for text and voice processing, especially considering the small size of the speaker community. In this sense, Luxembourgish could be labeled not as an "under-resourced" but as an “under-researched” language. Given its sociolinguistic development and the recent interest in small-language contexts in NLP, Luxembourgish offers many promising starting points for research and tool development.

In this talk, I will focus on the written domain, that is, the development of written Luxembourgish over time. Using data from the news platform RTL.lu, I will demonstrate how both individual writing practices and the overall development of written Luxembourgish can be monitored and analyzed using a newly developed NLP pipeline for orthographic correction of text data. In this context, I will also discuss future avenues of research on Luxembourgish and the potential of small-language contexts as a source of innovation in NLP research.

Keywords: small languages, Luxembourgish, standardization, NLP, tool development

Biography:

CHRISTOPH PURSCHKE works as an Associate Professor in Computational Linguistics at the University of Luxembourg’s Institute for Luxembourg Studies. He is also the head of the newly founded Culture & Computation Lab (CuCo Lab), a transversal research unit for Cultural Data Science hosted at the Department of Humanities. The starting point of his research is formed by the complex relationships between human activity in the world on the one hand and the structure and dynamics of cultural symbol systems on the other. The main focus of his work is the empirical investigation of sociolinguistic issues, the implementation of computational approaches to language use and evaluation, and the development of theoretical models for the description of human cultural practice. Christoph sees the close connection of computational working methods and theoretical reflection with societal engagement and an open research practice as a special motivation for his work as well as a challenge for the future of the humanities as a whole.

rijksuniversiteit groningen / faculteit wijsbegeerte

Conference presentation #1
11:35 – 12:00

ADVANCED NATURAL LANGUAGE PROCESSING METHODS FOR ANALYSING LGBTIQ+ AND MIGRATION-RELATED TEXTS

Senja POLLAK
Jožef Stefan Institute (Slovenia)
Department of Knowledge Technologies
senja.pollak [at] ijs.si

Recent advances in the field of natural language processing are based on large language models and transfer learning. We will present selected approaches to analysing minority related topics, focusing on LGBITQ+ and migration-related discourses in Slovene texts.

First, we will present a method for diachronic analysis based on semantic change detection using contextual embeddings, and showcase it on migration-related words, comparing different time spans from the Slovene reference corpus. Next, we will present, how the method can be adapted for viewpoint detection, comparing liberal and conservative Slovene news media coverage of the LGBTIQ+ topic, where we notice also different distribution of news sentiment. Different viewpoints can be detected also by understanding the decisions behind classification models. We will present a use case on the Slovenian parliamentary debates on the topic of migrations, where we interpret the differences in speeches by left- and right-leaning members of the parliament. Last but not least, we will focus on biases in pretrained language models by using prompting and sentiment analysis.

Keywords: LGBTIQ+, migration, NLP, transfer learning, diachronic analysis

Biography:

SENJA POLLAK is a researcher and NLP group leader at Department of Knowledge Technologies at Jožef Stefan Institute (JSI) in Ljubljana, Slovenia. She was coordinator of the H2020 project EMBEDDIA (12 partners, budget 3 mil. EUR), is co-leader of RobaCOFI project (funded under the call of AI4Media), and is the leader of national project CANDAS. She has been a leader for industrial projects Kliping, TermIolar 1, 2, WP/task leader on EU projects MUSE, SAAM, PROSECCO, and institutional lead on national projects SOVRAG, KOBOS, FORMICA, and TermFrame. She has served in several conference organisation and program committees (chair of SLSP 2019) and published papers in IF journals including *Computational Linguistics*, *Natural Language Engineering*, *Terminology*, *Language Resources and Evaluation*, and the *International Journal of Lexicography*.

12:00 – 13:30
#Lunch break#

Conference presentation #2
13:45 – 14:15

MULTILINGUAL MULTIPLEX CORPORA AND ANALYTICAL-CREATIVE HCI POETRY IN EUROPEAN MINORITY LANGUAGES

Chris TANASESCU
Open University of Catalonia (Spain)
The Internet Interdisciplinary Institute
ctanasescu [at] uoc.edu

The paper presents the work carried out as part of a wider project on minority language poetries across Europe and beyond, and focuses on the portion dedicated to 10 minority, endangered, and/or vulnerable Romance languages plus Basque. The dataset has been assembled by combing online archives, websites/blogs or out-of-print anthologies (e.g., in Aromanian or Extremaduran) most of which are hard to locate and/or unmaintained, if not endangered in their turn. The experiments so far included representing the data as multiplex networks and analyzing the latter for topological features. The nodes in the multiplexes are the poems in our corpora and the layers represent features such as linguistic density or readability and computational prosody (or vector prosody) (Tanasescu 2022, MARGENTO *et al.* 2021).

While the network analysis involved foregrounds commonalities and trends across languages and corpora, it is also instrumental (alongside the above-mentioned features) in the creative-work component of the project. Such analytical-creative approach (Tanasescu 2023) and the resulting HCI poetry refers to an expanded concept of data(-based) creativity that ranges from text writing/assemblage/generation to algorithmic translation to automated corpus expansion. The novelty—by comparison to my previous #GraphPoem work outputting "computationally assembled poetry anthologies" (MARGENTO 2018 and 2024)—consists in the fractal-like methodological coherence between all levels of analytical-creativity involved (termed here "corresponsive composition), from word and line-of-verse to collage and multilingual corpus-based multiplexes.

Keywords: minority languages, multiplex network analysis, NLP, HCI, prosody

Biography:

CHRIS TANASESCU has degrees in both English and Computer Science and works at the intersection of Natural Language Processing (NLP), network analysis, poetry, and translation. He is the author, editor, or translator of over twenty books, including an internationally praised computationally assembled poetry anthology and a topic-modeling-driven collaborative poetry collection described by Servanne Monjour (Sorbonne University) as a "pioneering computational translation." As an internationally awarded intermedia poet (a.k.a. MARGENTO), he has chaired the #GraphPoem performances at the Digital Humanities Summer Institute since 2019. He currently fills the position of Senior Research Scientist with Internet Interdisciplinary Institute (GlobaLS and CoSIN3) contributing to the GlobaLS Lab's Global Translation Flows research line at UOC, and has previously served as Professor and Coordinator of Digital Humanities at University of Ottawa and Altissia Chair in Digital Cultures and Ethics at UCLouvain.

rijksuniversiteit groningen / faculteit wijsbegeerte

Conference presentation #3
14:15 – 14:45

NATURAL LANGUAGE PROCESSING IN CLOWDFLOWS

Martin ŽNIDARŠIČ
Jožef Stefan Institute (Slovenia)
Department of Knowledge Technologies
martin.znidarsic [at] ijs.si

ClowdFlows is an open-source online platform for development and sharing of data processing workflows. Its workflow development user interface allows for visual programming, as programming components can be placed on a canvas and have their inputs and outputs visually connected to form functional solutions. ClowdFlows is a research prototype that is primarily aimed at enabling simple and straightforward exposure and sharing of research work and results. A combination of visual programming interface with a simplicity of reuse of the shared workflows makes it suitable for demonstrating solutions and providing experimentation capabilities also to users with very limited programming skills.

We will introduce ClowdFlows, its user interface and provide an overview of its components, with a focus on components for natural language processing, such as tokenizers, lemmatizers, text embedding approaches and pretrained text classifiers. There will be a discussion of the benefits and drawbacks of using ClowdFlows and a demonstration of some practical examples and use cases related to potential use of ClowdFlows for research in the humanities.

Keywords: software demo, ClowdFlows, NLP data processing workflows, research workflows

Biography:

MARTIN ŽNIDARŠIČ is a senior researcher at the Department of Knowledge Technologies of the Jožef Stefan Institute in Ljubljana, Slovenia. His main research interests are in machine learning, decision modelling and text analytics. As a decision and data analyst, he was involved in several international, national and commercial projects, commonly acting as work package leader. He is teaching machine learning and artificial intelligence courses at the Jožef Stefan International Postgraduate School in Ljubljana and the Faculty of Industrial Engineering in Novo Mesto.

rijksuniversiteit groningen / faculteit wijsbegeerte

Conference presentation #4
14:45 – 15:15

AD HOMINEM: CREATING A NAMED-ENTITY RECOGNITION MODEL FOR EARLY MODERN LATIN

Silvia DONKER,[1] Michiel VAN DER REE[2]
Rijksuniversiteit Groningen (The Netherlands)
[1]Faculty of Philosophy & [2]Centre for Information Technology
[1] s.j.donker [at] rug.nl, [2] michiel.van.der.ree [at] rug.nl

How does one trace scholarly impact and influence in a large body of historical texts, written in a low-resource language? In contemporary research, citation analysis is a common method to find indicators of impact, influence or quality in scientific literature. Such research, however, relies on digital availability and standard practices and tools with regard to specific referencing systems and language usage. While we have a corpus of around 600 Latin works on natural philosophy from between 1600-1800 (acquired in a previous stage, see Sangiacomo *et al.* 2022), there are no standards to rely on, so we set out to create our own.

Extracting names from texts is commonly done with Named-Entity Recognition (NER), a form of Natural language Processing (NLP) to identify and classify named entities. Pre trained models for this exist, but are usually built for modern content and cannot naively be applied: they need to be tuned to the specific needs of the researcher. Problems arise from the idiosyncrasy of the corpus, such as outdated language and textual particularities. NER tools for historical sources are in full development, but an 'off-the-shelf' NER model that includes Latin, the *lingua franca* of Early Modern Europe, is not available. Previous work and datasets are problematic, because of a heterogeneous language over time and place, OCR errors and other difficulties (Erdmann *et al.*, 2016).

We finetune a multilingual basemodel (mBERT) for the recognition of proper names, places and groups in a large body of Early Modern Latin works dealing with natural philosophy. Because of a lack of existing tools and limited machine-readable manuscripts, a reference study of this magnitude for this period has not yet been conducted before. The corpus' inclusiveness necessarily goes beyond the canon and has the potential to reveal generational patterns or reinstate minor figures that are easy to miss through the close reading of a limited amount of literature. We will share the creation process and some preliminary findings in terms of an Early Modern reference network.

Keywords: NER, Latin, early modern science, transformer model, multilingual BERT

Biographies:

SILVIA DONKER is a PhD student working on a Digital Humanities project at the Faculty of Philosophy (RUG). As part of the ERC project The Normalisation of Natural Philosophy, she works on the reconstruction of social networks based on authors and sources in early modern natural philosophy. She is interested in knowing how networks can be used to gain an understanding of human culture and practice, whether historical or contemporary, which is why she co-founded The Patio, an interdisciplinary research group on social networks at the RUG.

MICHIEL VAN DER REE is a data scientist at the Center for Information Technology at the University of Groningen. As a member of the data science team, he supports researchers in addressing their research questions through eScience methods. His specialization lies in natural language processing and its use in the field of Digital Humanities.

15:15 – 15:30
#Coffee break#

Conference presentation #5
15:30 – 16:00

MINOR FIGURES, MAJOR THEMES. MULTILINGUAL TOPIC MODELS IN AN EARLY MODERN PHILOSOPHY CORPUS

Abigail SHANAB,[1] Diana INKPEN,[2] Raluca TANASESCU,[3] Andrea SANGIACOMO[4]
[1, 2]University Ottawa (Canada), [3]University of Galway (Ireland),
[4]University of Groningen (The Netherlands)
[1]ashan039 [at] uottawa.ca, [2]diana.inkpen [at] uottawa.ca, [3]r.a.tanasescu [at] gmail.com,
[4]a.sangiacomo [at] rug.nl

Abstract:

Major themes in early modern science, a discipline with a long textual tradition, cover not only the Aristotelian paradigm of scholastic natural philosophy, but also rivalling Renaissance and seventeenth-century conceptions of physics, with a focus on key issues related to understanding the universe, such as nature, cause, motion, or electricity. Traditionally, each of these issues has been related to one or several canonical figures, who have been studied more or less in isolation, using descriptive approaches. The work carried out by the team working on the ERC-funded grant "The Normalisation of Natural Philosophy" (Sangiacomo 2019) at the University of Groningen have been studying a 601-title corpus of natural philosophy works written in three languages (Latin, English, and French) to better understand the landscape of this discipline between 1600 and 1800 in nowadays Netherlands, Great Britain and France. By integrating NLP and network analysis methods, they represented the corpus as a multilayer network, in which each layer represented a monolingual-context (Sangiacomo *et al*. 2022).

Our presentation will explore the same corpus multilingually, offering an integrated, transnational view that straddles language and border divides. Using neural topic modeling, more specifically topic2vec with a multilingual sentence / document encoder model, we will derive topics across the three languages and will compare the results with the LDA approach previously used. Multilingual topic modeling will help us cluster canonical figures with less studied philosophers and works on the grounds of their shared interest in major themes in early modern science.

Keywords: natural philosophy, minor figures, neural networks, multilingual topic modeling, network science

Biographies:

ABIGAIL SHANAB is a fourth-year student in the Faculty of Engineering at the University of Ottawa. She is completing her Honours BSc in Computer Science with a Data Science Option. She is interested in Natural Language Processing (NLP) and Machine Learning, which is why she is working on an Honours project on multilingual topic retrieval.

DIANA INKPEN is Professor of Computer Science at the University of Ottawa (PhD uToronto). Her research is in applications of natural language processing and text mining. She is the Editor-in-Chief of the Computational Intelligence journal and an Associate Editor of the Natural Language Engineering journal. She published a book on *Natural Language Processing for Social Media* (Morgan and Claypool Publishers, Synthesis Lectures on Human Language Technologies, 3 editions), ten book chapters, more than 35 journal articles, and more than 120 conference papers. She received many research grants, from which the majority include intensive industrial collaborations.

RALUCA TANASESCU is a postdoctoral in translation and global media at the University of Galway's Moore Institute. Her work employs transdisciplinary approaches to multilingualism and minority that span complexity thinking, translation studies, literary studies and digital humanities.

ANDREA SANGIACOMO is associate professor of philosophy at the University of Groningen, where he currently teaches global hermeneutics and ancient Buddhist philosophy. He has worked extensively on early modern philosophy and science, and devoted a significant part of his research to Spinoza. He holds the Spinoza Chair at the Erasmus School of Philosophy in Rotterdam.

rijksuniversiteit groningen / faculteit wijsbegeerte

Closing keynote presentation
16:00 – 17:15

CULTURAL DIVERSITY.
AN ECOLOGICAL APPROACH TO THE STUDY OF MINORITY

Diana ROIG-SANZ
Open University of Catalonia (Spain)
The Internet Interdisciplinary Institute
dsanzr [at] uoc.edu

This lecture aims at offering some theoretical insights on how we can apply digital scholarship to address minority—here understood in terms of language, gender, ethnicity, race—in a more sustainable way. I will propose preliminary research based on my current ERC StG project and an innovative framework that will be grounded on ecological approaches and cutting-edge technology to address one of the most pressing issues we are facing today: the examination of languages and collectivities historically marginalised in the global cultural ecosystem. My goal is to raise awareness about the need to pool resources to study sustainable and unsustainable cultural practices, which will allow us to examine the specific kinds of diversity we are indeed promoting. By assuming that both our planet and our society are biodiverse, this lecture will propose a more inclusive and global vision that raises new ethics about the relevance of all the constituents of the cultural ecosystem and the dynamics among them. Thus, I address minority in terms of cultural biodiversity and as a global, relational, and multi-layered phenomenon. In this respect, this lecture will reflect on what ecological perspectives may add to literary and cultural theory (Bourdieu, Luhmann), and acknowledge a broader complexity (Latour). By incorporating an ecological dimension and notions such as that of interdependence or sustainability, I aim at challenging the role of economics or politics as the sole structuring forces. Also, I would like to discuss to what extent we should not replace terms such as those of minority, peripheral or small by other more inclusive notions such as that of diversity. Literature confirmed overlapping distribution between world's plant diversity zones and the world's languages, but we do not have ecological knowledge on, for example, cultural institutions promoting translations and multiple languages.

Keywords: minority, cultural diversity, ecological approaches, translation, digital humanities.

Biography:

DIANA ROIG-SANZ is an ICREA Full Professor at the IN3-UOC, in Barcelona. She is the coordinator of the Global Literary Studies Research Lab (GlobaLS) and the PI of the ERC Starting Grant project "Social Networks of the Past. Mapping Hispanic and Lusophone Modernity, 1898-1959." Her research interests deal with global and cultural approaches applied to literary and translation history within a digital humanities approach. She also works on sociology of translation and minor and less-translated languages and literatures. Her publications include *Bourdieu después de Bourdieu* (2014), *Literary Translation and Cultural Mediators in 'Peripheral' Cultures* (2018, with R. Meylaerts), *Literary Translation in Periodicals* (2020, with L. Fólica and S. Caristia), *Culture as Soft Power* (2022, with

E. Carbó), or *Global Literary Studies: Key Concepts* (2022, with N. Rotger). She has also published her research at the *Journal of Global History* (2019), *Culture and Social History* (2020), *Translation Spaces* (2021), or *Comparative Literature Studies* (2022). She has conducted research residencies at the Oxford Internet Institute, KU Leuven, the École Normale Supérieure, or the Amsterdam School for Cultural Analysis.

Conference presentation #6 (in absentia*)

EXPLORING THE POTENTIAL OF MACHINE TRANSLATION FOR FACILITATING MULTILINGUAL DIGITAL HUMANITIES

Lynne BOWKER
University of Ottawa (Canada)
School of Translation and Interpretation
lbowker [at] uottawa.ca

Although machine translation technology has existed for nearly 75 years, the first 50 years saw relatively little progress as developers pursued techniques that largely resembled the way in which people process language. By the turn of the millennium, the technological landscape had changed significantly: computers were faster and more powerful, and thanks in large part to the internet, more people were creating and sharing documents in electronic form, meaning that textual data was becoming plentiful. This allowed researchers to conceive of a completely different way of approaching the challenge of machine translation, leading to data-driven approaches, such as statistical and neural machine translation. In particular, neural machine translation has been a game changer because, although it is not perfect, it has increased the quality of machine translation output to a point that makes a viable draft or starting point for many purposes. In the Digital Humanities community, there is a growing increase in multilingualism, although this has been hampered by practical constraints. If researchers don't speak or understand each other's languages, how can multilingualism be achieved? While neural machine translation has the potential to help, it cannot be applied in an unchecked fashion. Rather, tool users must employ this technology in a responsible way. To do that, many users will need to improve their own machine translation literacy as a first step in integrating machine translation as part of a broader strategy for increasing multilingualism in DH. This contribution will focus on how DH scholars can improve their machine translation literacy by explaining the strengths and limitations of data-driven approaches to machine translation, considering the risks and rewards of using this technology, and exploring ways in which users can optimize the results.

Keywords: machine translation literacy, scholarly communication, machine translation evaluation, multilingualism, digital humanities

*This contribution will appear in the post-conference volume *The Politics of Translation and Multilingualism in Contemporary Techno-Humanities* (edited by Raluca Tanasescu). Due to a time conflict, Lynne Bowker's participation in the conference was not possible.

Biography:

LYNNE BOWKER is Full Professor at the University of Ottawa in Canada, where she holds a cross-appointment between the School of Translation and Interpretation and the School of Information Studies. In 2020, she was elected as a Fellow of the Royal Society of Canada for her contributions on translation technologies. During the 2022/2023 academic year, she is a NAWA Visiting Researcher in the Scholarly Communication Research Group at the Adam Mickiewicz University in

Poland, where she is working on a project about plain language and machine translation. She is the author of several books, including *Computer-Aided Translation Technology* (University of Ottawa Press, 2002), *Working with Specialized Language* (Routledge, 2002), *Machine Translation and Global Research* (Emerald, 2019) and the open access book *De-mystifying Translation* (Routledge, 2023).