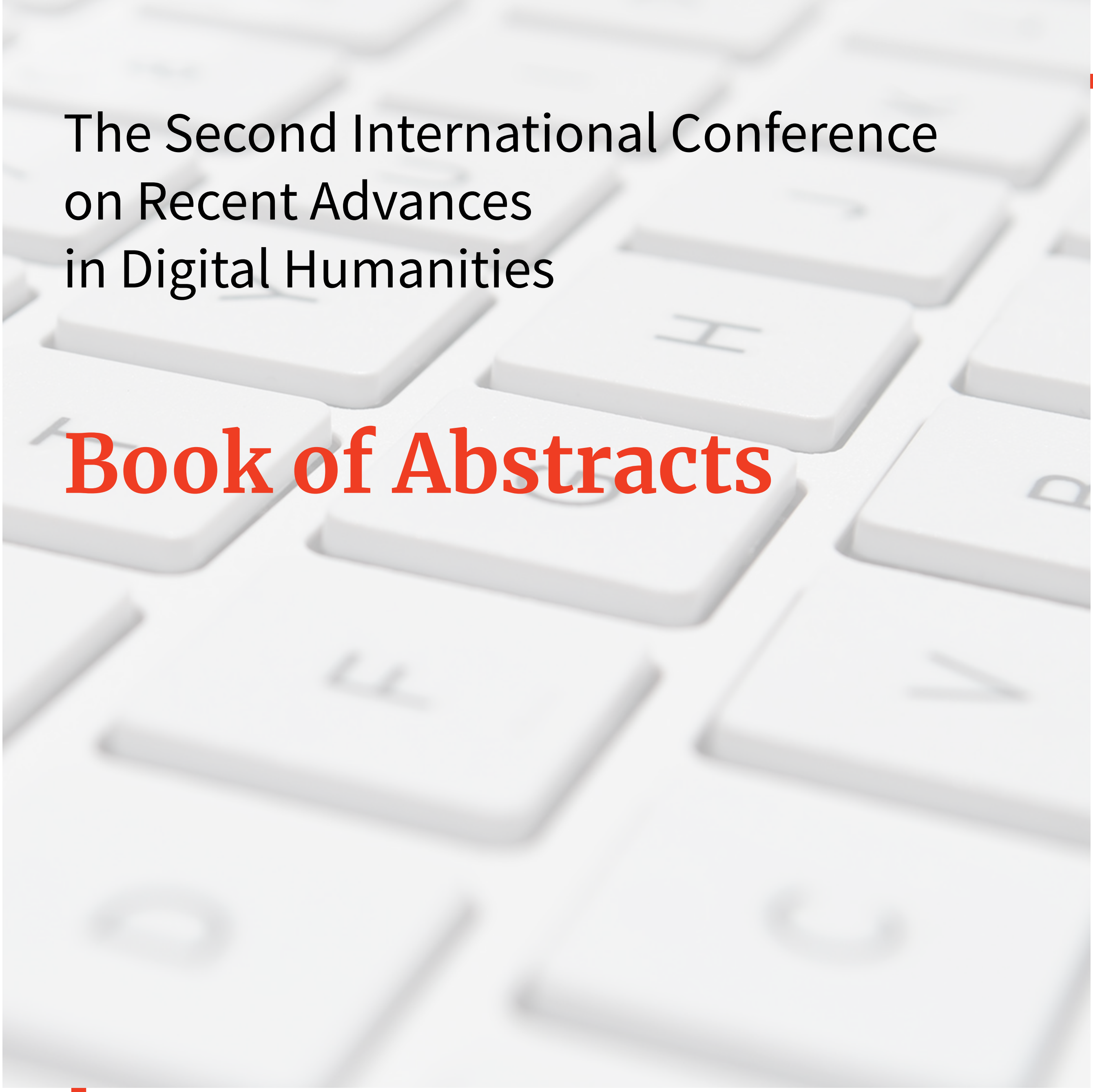
Universitatea de Vest din Timişoara

codhus
CENTRE FOR CORPUS RELATED DIGITAL APPROACHES TO HUMANITIES

RADH 2023
The Second International Conference on Recent Advances in Digital Humanities



The Second International Conference on Recent Advances in Digital Humanities

Book of Abstracts

Online:
https://stream.meet.google.com/stream/aca2a123-ef89-4a9f-9249-570426aeff25
(EET Time zone)

17-18 Noiembrie 2023

Timişoara, Romania

Programme Committee Chair

Anca DINU	University of Bucharest

Programme Committee

Anca Dinu	University of Bucharest, Romania
Mădălina CHITEZ	West University of Timișoara, CODHUS
Liviu P. DINU	University of Bucharest, HLT
Mihnea Dobre	University of Bucharest
Paolo Rosso	Technical University of Valencia, Spain
Chris Tănăsescu	Open University of Catalonia, Spain
Ioana Galleron	Paris 3 Sorbonne Nouvelle, France
Alexandru Nicolae	University of Bucharest, Romania
Simona Georgescu	University of Bucharest, Romania
Andrea Sgarro	University of Trieste, Italy
Marcos Zampieri	George Mason University, USA
Alina Țigău	University of Bucharest, Romania
Roxana Rogobete	West University of Timisoara, Romania
Alexandra Lițu	University of Bucharest, Romania
Loredana Bercuci	West University of Timisoara, Romania
Alina Resceanu	University of Craiova, Romania
Florentina Nițu	University of Bucharest, Romania
Alexandru Oravițan	West University of Timisoara, Romania
Ion Resceanu	University of Craiova, Romania
Roxana Patraș	Alexandru Ioana Cuza University of Iasi
Adrian Cîntar	West University of Timisoara, Romania

General Editors

Anca DINU & Mădălina CHITEZ

Editing Assistants

Roxana ROGOBETE	West University of Timișoara
Cristina ONEȚ	West University of Timișoara

Cover photo credit: Sergi Kabrera (source: Unsplash)

RADH 2023: CONTENTS

Keynote Lectures

Papers

I. LANGUAGE RELATED DIGITAL HUMANITIES

II. NLP OUTREACH

III. Culture, History and Language

KEYNOTE SPEAKERS

Recent Applications of the German Wordnet in the Digital Humanities

Erhard Hinrichs

In my presentation, I will demonstrate the added value of integrating semantic information from wordnets in a variety of digital humanities applications. After an introduction to wordnets, I will turn to a range of use cases from recent studies in digital humanities for German that range from computational literary studies to computational lexicography, semantic information retrieval, and language learning and that have all benefitted from the integration of wordnet information.

More specifically, I will focus on the wordnet for German that is called GermaNet and that we have developed in my research group at the University of Tübingen since 2003 and that is still being extended on a yearly basis. GermaNet covers the base vocabulary of contemporary German and has been extended on the basis of word frequency lists derived from large digital corpora.

Erhard Hinrichs is Professor of General and Computational Linguistics at the University of Tübingen, Germany and a senior researcher at the Leibniz Institute for German Language in Mannheim, Germany. He is an honorary member of the Foundation of Logic, Language and Information and an honorary member of the Linguistic Society of America. He is currently the Scientific Speaker of the German National Infrastructure Consortium Text+ (https://text-plus.org/) and has served as a member of the Executive Board of the ESFRI-Initiative CLARIN (www.clarin.eu) and as CLARIN National Coordinator for Germany (https://www.clarin-d.net/en/) for many years. His current research focuses on computational semantics, language resources for digital humanities applications, and machine learning approaches to linguistic modelling.

In search of lost smells. Tracing the olfactory history of Europe with NLP

Sara Tonelli

In recent years, European cultural heritage institutions have invested heavily in large-scale digitisation: we hold a wealth of object, text and image data which can now be analysed using artificial intelligence. However, a methodology for the extraction of scent-related information from large amounts of texts is still missing, as well as a broader awareness of the wealth of historical olfactory descriptions, experiences and memories contained within heritage datasets. In this talk I will describe ongoing activities towards this goal, focused on text mining and semantic processing of olfactory information, which have been carried out within the ODEUROPA project. The aim of the project is to develop a "computer nose": design AI strategies to capture references to smells in the past from digital heritage collections. In this talk, I will provide a step-by-step explanation on how ODEUROPA employs advanced NLP to extract smell-related information from texts, and how these extracted bits of information are then encoded in a graph-like database built on Semantic Web technologies (the European Olfactory Knowledge Graph, or EOKG). From here, we can extract storylines and follow smell sources, noses, and olfactory practices over time and space.

Sara Tonelli is the head of the Digital Humanities research group at FBK. She is currently involved in the H2020 ODEUROPA project, where she leads the work-package related to olfactory information extraction. She is also part of several other interesting European projects, for example PROTECTOR, STAND BY ME and SCAN2, more details here. She is also involved in the HYBRIDS MSCA network, whose goal is to fight disinformation using AI and human-in- the-loop approaches. Between January 2021 and December 2022 she was also the scientific coordinator of the KID ACTIONS European project, aimed at addressing cyberbullying among children and adolescents through interactive education and gamification.

On the definition and the automatic detection of hate speech. A tale of pulling translators into NLP

Alberto Barrón-Cedeño

The democratisation of the web has opened the door for people to freely expressonline. Unfortunately, malicious users take advantage of it to spread hate. In this talk, I will present our efforts on the creation of automatic approaches for the automatic identification of different kinds of hate speech over the last three years. I will pay special attention to discuss open questions, such as where is the boundary between hate speech and &quot;freedom of expression&quot;, which actions should be taken when in the presence of a hateful post or community, and what exactly means hate speech (but from a societal and from a legal perspective). Aspects that make deep learning models struggle to properly identify hate speech (such as lexical creativity, lack of annotated data or level of expliciteness).

Alberto Barrón-Cedeño is an associate professor at Università di Bologna (Italy). Before that, he was a Scientist at the ALT group of Qatar Computing Research Institute (Qatar; winner of the King Salman International Academy for the Arabic Language Award 2023), and an Alain Bensoussan fellow at Universitat Politècnica de Catalunya (Spain). He obtained his PhD on AI from Universitat Politècnica de València (Spain). Alberto is interested in the automatic analysis of diverse qualities of text, such as originality, relevance, and intent; also across languages. He has co-organised various editions of both the PAN and the CheckThat! Labs at CLEF, a SemEval shared task and served as general chairman of the 2022 edition of CLEF, in Bologna; he has published in the top forums of NLP and IR (e.g., ACL, EMNLP, SIGIR, ECIR, IP&amp;M, LRE, KNOSYS).

PAPERS

I. Language Related Digital Humanities

FANFICTION ANALYSIS: HUMAN VS AI-GENERATED TEXTS

Anca Dinu & Andra Maria Florescu
University of Bucharest

Fanfiction represents a written text made by fans regarding any subject they are a fan of such as books, movies, cartoons, celebrities, games, etc. Since its debut, in the 70s, its popularity was constantly growing, due to the fact that fanfiction is a space of freedom, inclusiveness and creativity outside of canonical works. It inspired new books, movies, tv series and shows and it recently gained attention from the academic community.

This study aims to computationally analyze and compare fanfiction texts (fanfics for short) from three different sources: two popular websites, namely Wattpad and Archive of Our Own (AO3), and fanfics generated by ChatGPT. To this end, we have chosen fanfics based on the novel "Fire &amp; Blood" by the famous writer George R.R. Martin, the creator of the series of books which were also the base of the TV Show series "Game of Thrones" and "House of the Dragon". The data is balanced, comprising 50 carefully selected fanfics (of a minimum length of 10000 words) from each of the three sources, 150 fanfics in total. In addition, we included the original text consisting of more than 200000 words. All data have been preprocessed by cleaning the metadata, removing stopwords, tokenizing the text, lowercasing the text, lemmatizing the words, and removing non-character elements.

The main objectives of this research are to:

- discover common themes (topic modelling);
- assess and compare the readability of the fanfics;
- perform sentiment analysis, assigning positivity and negativity scores to the fanfics from all the three sources;
- investigate the semantic distance between the fanfics from all the three sources and the novel;
- evaluate ChatGPT&#39;s (version 3) ability to generate fanfics similar to those authored by human authors;
- Automatically classifying machine versus human – written fanfics.

For these purposes we used NLP methods and DistilBERT, Voyant, and Sketch Engine.

The code was written in Python, using Visual Studio Code and Google Collab. We also used ChatGPT and BingAI (version 4) to assist in writing and improving the code.

Some of the interesting differences we have found between the fanfics from the three sources are:

- ChatGPT-generated fanfics appear to have slightly higher similar topics to the novel thanWattpad and Ao3 fanfics;

- while human-written fanfics are balanced w.r.t. the positive/negative proportion (24 of the 50 Wattpad fanfictions listed have been labeled as having positive sentiment, while 26 have been labeled as negative emotion), ChatGPT-generated fanfics have all been labeled as positive, with an average sentiment score of 0.997.
- fanfics from AO3 have the highest textual similarity to the book, according to the Jaccard similarity metric, with the most comparable fanfic having a similarity score of 0.5171; Wattpad fanfics are likewise relatively similar to the book, with the most similar fanfic getting a similarity score of 0.4731; ChatGPT fanfics are much less similar to the book, with the most similar fanfic having a similarity score of 0.1376. In terms of the Cosine similarity measure, Wattpad and AO3 fanfics are very similar to the book, having the top similar fanfic earning a similarity score of 0.66, while ChatGPT fanfics are less similar to the book, with the most comparable fanfic having a similarity score of 0.5094;
- the automatic classification of human written fanfics was much more difficult than the one for ChatGPT-generated fanfics.

VADEMECUM ON DIGITAL HUMANITIES IN ROMANIA: AVAILABLE RESOURCES AND THEIR APPLICABILITY

Madalina Chitez, Roxana Rogobete & Adrian Cîntar
West University of Timisoara

In recent years, there has been in impetus to develop the interdisciplinary field of Digital Humanities in Romania (e.g. Dinu et al., 2022; Chitez et al, 2021). As in other national contexts, the field has been expanding rather heterogeneously since research has been conducted either in the area of computational linguistics (e.g. RACAI and Human Language Technologies Research Center in Bucharest) or digitization of written resources (MDRR in Sibiu). Other recent initiatives (e.g. CODHUS in Timisoara) have approached the DH field from an applied perspective with projects that use corpus linguistics methods to create education-relevant digital tools. At the same time, the plethora of international resources useful for DH research (www.corpus-analysis.com) and applications in Romanian language has not been analysed and inventoried for further use. In this paper, we propose an overview of existing resources, national and international, which have been analysed, tested and categorized based on several criteria: origin (±national), Romanian-only (±), type (online dataset/ds, Github/git, tool/t, other/x) availability (±), quality (±), language related (±), (field (corpus linguistics/cl, computational linguistics/CL, digitization/d, history/h, geography/g, literature/l, social studies/ss, education/e, applied studies/as, other/x), availability (freely available/fa, commercial/c), user-friendliness (±). For the selection of resources, we have performed a systematic literature (e.g. Tufis, 2022) and web resource review (www.eadh.com) as well as personal experience with some of them. For major resource types, we bring forth examples, such as the platform https://deportatiinbaragan.ro/, which we describe in more detail. We also insist on the didactic impact of the presented resources. The final part of the paper consists of a discussion on the statistical distribution of the resource inventory features, followed by broader conclusions and recommendations.

References

- Chitez, M., Rogobete, R. and Foitos, A. (2020). Digital Humanities as an Incentive for digitalisation strategies in Eastern European HEIs: a case study of Romania. In A. Curaj, L. Deca and R. Pricopie (Eds.), European Higher Education Area: Challenges for a New Decade (pp. 545-564). Cham: Springer.
- Dinu, A., Chitez, M., Dinu, L., & Dobre, M. (Eds.). (2022). Recent Advances in Digital Humanities. Romance Language Applications. Bern: Peter Lang.
- Tufiş, D. (2022). Romanian Language Technology-a view from an academic perspective. International Journal of Computers Communications & Control, 17(1).

Web references
https://www.corpus-analysis.com/ - Tools for Corpus Linguistics
https://codhus.projects.uvt.ro/ - Centre for Corpus Related Digital Approaches to Humanities
https://eadh.org/- The European Association for Digital Humanities
https://deportatiinbaragan.ro/ - Deportaţi în Bărăgan
https://revistatransilvania.ro/mdrr/ - Muzeul Digital al Romanului Românesc
https://nlp.unibuc.ro/ - Human Language Technologies Research Center in Bucharest
https://www.racai.ro/en/ - Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu"

REDEFINING CONVERSATIONAL AI: THE CHISM MODEL'S APPROACH TO CHATBOT-HUMAN COMMUNICATION

Jose Belda-Medina
University of Alicante

The field of Artificial Intelligence (AI) is currently undergoing a significant transformation, thanks to advancements such as neural networks and deep learning models (Huang et al. 2023). These are enhancing the capabilities of modern AI systems to unprecedented heights. This growth is not solely driven by cutting-edge technologies but also by the vast data available through Large Language Models (LLMs). In this evolving landscape, the Digital Humanities industry finds itself at a critical crossroads, grappling with the challenge of how best to integrate Artificial Intelligence into human communication (Sundar &amp; Lee, 2022). Recent studies on chatbot-human interactions have delved into the capabilities of intelligent conversational agents, revealing a multifaceted landscape of opportunities and challenges (Kooli). However, despite these advancements, the lack of a specific model to systematically assess user satisfaction remains an obstacle to optimizing and refining digital conversational experiences. This presentation introduces the CHISM (Chatbot-Human Interaction Satisfaction Model) to analyze the intricacies of digital communication, with a special focus on the complex dynamics of chatbot-human interactions. CHISM offers a comprehensive framework that includes three distinct dimensions: Language Experience (LEX), Design Experience (DEX), and User Experience (UEX). These dimensions address various aspects of chatbot-human interaction, such as response interval, lexical variety, and grammatical complexity. To validate this model, a study was conducted involving 142 college students from Spain and the Czech Republic. Over a month, these participants engaged with three intelligent conversational agents. Using a mixed-methods approach and convenience sampling, interactions were evaluated based on the CHISM criteria, emphasizing features like pragmatic understanding and speech technologies. Data, both quantitative and qualitative, were gathered through surveys and semi-structured interviews and were analyzed using SPSS statistical software and QDA Miner Lite. On the CHISM scale, pragmatic understanding and Speech technologies” in the LEX dimensión and multi-user interaction and customizing options in the DEX dimensión received lower scores compared to the other items assessed, suggesting that further improvements are needed to ensure user engagement and satisfaction. Participants also raised privacy concerns and highlighted the urgent need for ethical considerations and a user-centered approach in the further development and deployment of chatbot technologies.

References

- Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. Educational Technology & Society, 26(1), 112-131.
- Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. Sustainability, 15(7), 5614.
- Sundar, S. S., & Lee, E. J. (2022). Rethinking communication in the era of artificial intelligence. Human Communication Research, 48(3), 379-385.

THE SOCIAL CONSTRUCTION OF CLIMATE CHANGE IN ROMANIAN NEWS: A NOVEL DATASET AND CONTENT ANALYSIS

Denis Iorga (University of Bucharest, **Tudor-Andrei Dumitrașcu** (Pythia Socio-dynamical Technologies) **& Luca-Mircea Mihăilescu** (Department of Physics, Freie Universität Berlin)

Online news stories have an important role in shaping perceptions related to climate change. However, existing efforts to analyze Romanian online news articles related to climate change are limited to case studies conducted on a reduced number of samples. As such, the current study aims to use a large sample of Romanian online news related to climate change in order to answer the following research question: How is the issue of climate change socially constructed in the Romanian online news press? The answer to the research question is arrived at through an analysis of a novel publicly available dataset. The dataset consists of 5,884 Romanian online news articles published between 2008 and 2022 by 13 different online news media outlets, which contain the "climate change/changes" keywords (ro. "schimbarea climatică/schimbările climatice"). Two natural language processing pipelines were applied to the dataset for the purpose of identifying themes/topics related to climate change. The output of the first pipeline was represented by clusters of online news articles based on the TFIDF scores of the most common nous & proper nous in the dataset. The output of the second pipeline was represented by clusters of online news articles resulting from a BERTopic configuration. For each pipeline, the final processing step consisted of associating a list of keywords to the outputted clusters. The two methods were compared in terms of clustering efficiency and interpretability. The BERTopic configuration obtained a better silhouette score (.47 vs. .18), whereas the TFIDF configuration resulted in what can be regarded as higher-order (and thus more interpretable) themes. Based on the analysis of both outputs, the results suggest that Romanian online news stories related to climate change are built on at least nine themes: temporal, scientific, political, economic, grassroots movements, local, global, natural, and anthropomorphic themes. Such themes are often used conjointly to build discourses concerning climate change in the Romanian online news media.

ASSESSING THE READABILITY OF ROMANIAN L1 & ENGLISH L2 UNDERGRADUATE LITERARY ANALYSES. A CONTRASTIVE APPROACH

Alexandru Oraviţan, Mădălina Chitez & Roxana Rogobete
West University of Timisoara

In Romania, the issue of readability in conjunction with student writing has gone largely unexplored. As increasing attention is being paid to developing frameworks for analysing academic writing genres in languages other than English, quantitative and qualitative studies that address the multifaceted nature of student writing are much needed. The particular focus is required for one of the most frequent student academic writing genres in Romania: literary analysis. For this purpose, 208 Romanian L1 and 160 English L2 literary analyses from the bilingual Corpus of Romanian Academic Genres (ROGER) were digitally processed with various software to obtain individual readability scores using established formulas. An aggregate of this data was used to pinpoint similarities and differences and to map out a contrastive perspective on the readability of these texts, which may be helpful for academic writing practitioners. The texts written by the Romanian students in English display a satisfactory linguistic level typical of a "quality" or "standard" readability score. The texts written in Romanian (L1) display a similar readability score, calculated using standard readability tests adapted for the Romanian language (e.g. FRES, LIX) This might be interpreted in two ways: (a) students of the same study level use similar vocabulary because their general linguistic skills manifest similarly in L1 and L2, or (b) students engage in similar writing practice irrespective of the language in which they write, when approaching a similar academic genre. This study treads new ground in building a framework for assessing student writing from the readability perspective.

LOW-RESOURCE MACHINE TRANSLATION FOR COPTIC

Mihaela Antal-Burlacu & Sergiu NisioI
University of Bucharest

In the context of Natural Language Processing, low-resource languages (LRL) constitute a significant challenge that has started to be explored in the past few years (Gu et al., 2018; Weller-Di Marco and Fraser, 2022; Doğruöz and Sitaram, 2022).

Falling into this category, the present work represents a study on Coptic, an extinct language that represents the last stage of Egyptian spoken between the third and twelfth centuries CE in a standardized form, and how various machine translation models can be used in order to translate texts written from Coptic into English.

The contribution brought within this work, which is still a work-in-progress, consists in providing a cleaning pipeline for parallel Coptic - English texts that can be further expanded for a large-scale translation purpose and defining a system capable of performing translation between these two languages. Moreover, a future objective of this project is to implement an open-source model for Coptic, that can become a useful tool for in-depth research. The main question this work answers is whether Machine Translation can be applied in the context of LRL, considering the limitations that data scarcity and the lack of adequate linguistic resources impose, particularly for the case of Coptic, and whether these methods lead to coherent results that can be later interpreted.

Since there is no similar work that provides a corpus to make this task possible, the dataset is built specifically for this project. Even if Coptic has a vast heritage of written texts that were preserved throughout history, the lack of translated scripts from Coptic into English, where the translation matches the original text, was one of the difficulties encountered in the whole process of data gathering. In order to collect the data for building a parallel corpus of texts, the Coptic Scriptorium (Schroeder and Zeldes, 2013; Schroeder and Zeldes, 2016) resources were used and, after performing multiple data pre-processing techniques and removing duplicated or empty entries, the final corpus consisted of 9990 parallel Coptic-English texts, taken from a total of 30 different documents.

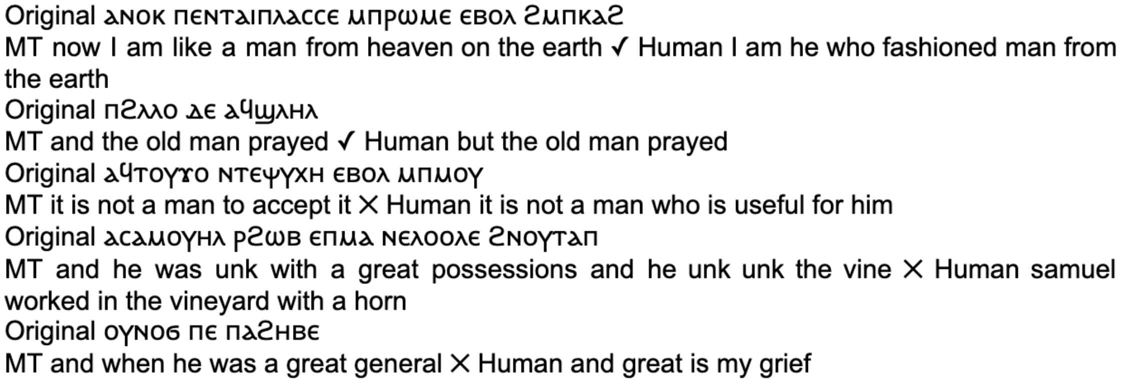
Most of the work attempted in this project concentrates around the use of neural networks to perform translation (Klein et al., 2017; Vaswani et al., 2017). Neural Machine Translation has proved to bring state-of-the-art results not only for the case of high-resource languages, but also when working with low-resource languages, even if it requires the availability of large-scale corpora. Two different architectures were taken into account: a Recurrent Neural Network-based model with Long Short-Term Memory (LSTM) layers for both the encoder and the decoder and a Transformer-based model, with a self-attention mechanism. The results were evaluated by using the BLEU metric: the Transformer model obtained the best score, with a BLEU of 9.6. Even if these results might appear relatively poor for real-life translations, the present project is still a work-in-progress and, also, a human analysis of the error could emphasize the actual quality of the translation. As the task of MT applied on the Coptic-English pair of languages hasn't been approached until this work, the results cannot be compared to other MT systems, but can represent a valuable baseline for future work.

Appendix

Below are several examples of translations produced by the system. Marked with ✓ symbol are the adequate translations and with symbol the erroneous ones. The occurrence of UNK words in the output is a result of the small size of the corpus which prohibits training high-quality word embeddings. To alleviate these errors, we are considering including explicit linguistic information from existing lexicons and the Coptic WordNet.

References

- Doğruöz, A. S. and Sitaram, S. (2022). Language technologies for low resource languages: Sociolinguistic and multilingual insights. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under Resourced Languages, pages 92–97, Marseille, France, June. European Language Resources



Original ⲁⲛⲟⲕ ⲡⲉⲛⲧⲁⲓⲡⲗⲁⲥⲥⲉ ⲙⲡⲣⲱⲙⲉ ⲉⲃⲟⲗ ϩⲙⲡⲕⲁϩ
MT now I am like a man from heaven on the earth ✓ Human I am he who fashioned man from the earth
Original ⲡϩⲗⲗⲟ ⲇⲉ ⲁϥϣⲗⲏⲗ
MT and the old man prayed ✓ Human but the old man prayed
Original ⲁϥⲧⲟⲩϫⲟ ⲛⲧⲉⲯⲩⲭⲏ ⲉⲃⲟⲗ ⲙⲡⲙⲟⲩ
MT it is not a man to accept it ✕ Human it is not a man who is useful for him
Original ⲁⲥⲁⲙⲟⲩⲏⲗ ⲣϩⲱⲃ ⲉⲡⲙⲁ ⲛⲉⲗⲟⲟⲗⲉ ϩⲛⲟⲩⲧⲁⲡ
MT and he was unk with a great possessions and he unk unk the vine ✕ Human samuel worked in the vineyard with a horn
Original ⲟⲩⲛⲟϭ ⲡⲉ ⲡⲁϩⲏⲃⲉ
MT and when he was a great general ✕ Human and great is my grief

Association.

- Feder, F., Kupreyev, M., Manning, E., Schroeder, C. T., and Zeldes, A. (2018). A linked Coptic dictionary online. In Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 12–21, Santa Fe, New Mexico, August. Association for Computational Linguistics.
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9:1460–1474.
- Gu, J., Hassan, H., Devlin, J., and Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 344–354, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.
- Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., and Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In Proceedingsof the Seventh Conference on Machine Translation (WMT), pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.
- Kontogianni, A., Ganetsos, T., Kousoulis, P., and Papakitsos, E. C. (2020). Computer-assisted translation of egyptian-coptic into greek. Journal of Integrated Information Management, http://ejournals. uniwa. gr/index. php/JIIM/article/view/4470. -
- Schroeder, C. T. and Zeldes, A. (2013). Coptic scriptorium.
- Schroeder, C. T. and Zeldes, A. (2016). Raiders of the lost corpus. Digital Humanities Quarterly, 10(2).
- Slaughter, L., Costa, L. M. D., Miyagawa, S., Büchler, M., Zeldes, A., and Behlmer, H. (2019). The making of Coptic Wordnet. In Proceedings of the 10th Global Wordnet Conference, pages 166–175, Wroclaw, Poland, July. Global Wordnet Association.
- Smith, D. and Hulden, M. (2016). Morphological analysis of sahidic coptic for automatic glossing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2584–2588.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

- Weller-Di Marco, M. and Fraser, A. (2022). Findings of the wmt 2022 shared tasks in unsupervised mt and very low resource supervised mt. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 801–805.
- Zeldes, A. and Abrams, M. (2018). The Coptic Universal Dependency treebank. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 192–201, Brussels, Belgium, November. Association for Computational Linguistics.
- Zeldes, A., Martin, L., and Tu, S. (2020). Exhaustive entity recognition for Coptic: Challenges and solutions. In Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 19–28, Online, December. International Committee on Computational Linguistics.

IN SEARCH OF THE LOST AUTHOR: A STYLOMETRIC ANALYSIS OF 185 ROMANIAN NOVELS PUBLISHED BETWEEN 1850-1950

PN-III-P1-1.1-TE-2019-0127

Lucreţia Pascariu, Alexandra Olteanu, Roxana Patraş & Oana Ionescu
A.I. Cuza University of Iasi

The present study relies on novel research methods that are specific to Computational Literary Studies, particularly on the stylometric analysis of a collection of novels by using the StyloR package. Through this proposal, we aim to open new research perspectives centered on the quantitative approach to literary texts. Such an approach involves selecting appropriate methods for research hypotheses on one hand, and evaluating the “specificity” of the analyzed material on the other. Among other functionalities, the StyloR package allows for the assessment of cases of unattributed authorship. Providing a homogeneous corpus (in terms of text size, genre, period, author gender, and the number of texts attributed to a single author),
stylometry has proven to yield remarkable results (Juola 2015). Building on the stylometric principles articulated by the Polish philosopher Wincenty Lutosławski as early as the 19th century, subsequently implemented in tools such as StyloR, this research endeavor seeks to identify and attribute the literary authorship to several anonymous texts included in the Romanian collection of the multilingual literary corpus ELTeC and in other 2 literary corpora (Hai-Ro and Pop-Lite). Among the 11 texts, two represent cases of uncertain authorship, which have thus far been discussed and attributed solely based on specific deductions from historical investigation (Bălăeţ 1982).
Through the application of stylometric analysis, we aim to either confirm or refute a hypothesis already formulated within the framework of literary history (typically based on contextualization). In the first stage of our exploration, we anonymized the targeted texts/authors, specifically, renaming the two novels and assigning a codename (“Rica Venturiano”) to the novelist to generate a visualization of stylistic affinities (or distances) among all the novels in the collection: 187 texts included in three corpora developed within the Digital Humanities Laboratory (ELTeC, Hai-Ro, Pop-Lite). In the second stage, we calibrated the analysis by modifying parameters available in the StyloR package, such as the “number of iterations”, sampling, and comparing sequences of a specific length (e.g., 1000 words) or reorganizing the material by creating microcollections based on chronology (works
from the same period) or the stylistic elements of more prolific authors. In the third stage, we discuss the results of our experiments and the possibility of validating the

authorship of the texts under consideration. In the third stage, namely that of results, our aim is to focus on the research questions.

For this reason, the outcomes of the StyloR analysis will encompass the following:
- Extracting the anonymous hajduk texts (6 out of 12) into a hajduk sub-corpus to discern the affiliations or disparities among the texts, considering the productive poles of the subgenre (N.D. Popescu, Panait Macri, Ilie Ighel).
- Investigating the case of uncertain literary paternity by individually tracing the two texts within a sub-corpus of texts published between 1850 and 1880.
- Conducting a separate analysis of the remaining anonymous texts by creating micro-corpora/micro-collections based on word count (long texts over 60,000 words, medium texts ranging from 20,000 to 60,000 words, short texts under 20,000 words).

References

- Dinu, L, Popescu, M., Dinu, A. (2008): Authorship Identification of Romanian Texts with Controversial Paternity. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).
- Eder, M. (2018): Authorship verification with the package stylo. Computational Stylistics Group Blog.

LEARNER CORPORA, INTERLANGUAGE AND CATALAN AS AN ADDITIONAL LANGUAGE: SOME DESCRIPTIVE CONSIDERATIONS

Elga Cremades
(University of the Balearic Islands

This paper aims at both presenting the creation of the Catalan Learner Corpus and offering a a first approach to the characteristics of the interlanguage of students of Catalan as an additional language whose first languages are Slavic languages, on the one hand, and Romance languages, on the other, using a methodology that combines error analysis, interlanguage studies, and learner corpus research.

As stated by Granger (2002), computer learner corpora, made up of continuous stretches of discourse which contain both erroneous and correct use of language, are real prove of learners' interlanguage and are thus extremely useful from both a linguistic point of view and a pedagogical perspective. Researchers on language acquisition may benefit from corpora when studying how particular features are acquired by different kinds of learners (for texts are classified according to sociolinguistic characteristics of learners, to conditions of the production and to linguistic content of the texts themselves). At the same time, teachers and researchers on language teaching and learner can see real evidence of production of different types of students and consequently design materials and activities for each kind of student.

The Catalan Learner Corpus, which is still not complete nor public, aims to become a computer corpus with 4357 written texts written by learners of Catalan of 42 different L1s (Cremades 2021), ranging from the intermediate low level to the advanced mid (ACTFL). The corpus aims at having a double utility. On the one hand, it aims to be a useful tool for the development of materials for teaching and learning Catalan –such as the interactive Catalan grammar GramCat: Morfologie současné katalánštiny (Cremades & Javorová-Švandová 2019). On the other, it will constitute the first public collection of data on the interlanguage of learners of Catalan as an additional language, which will provide helpful evidence for researchers.

This is the case for the study that will be presented in this paper. More than 200 texts were compiled. The errors were marked and classified according to a grammatical criterion (Rodríguez 2020), having as a result the following categories: orthography and orthotypography, morphosyntax, lexic-semantics or pragmatics-discourse (Abuhakema et al 2008). The paper will show how orthography and morphosyntax are the areas that have the highest percentages of errors, although the lexical-semantic field is where we can detect more clearly the interferences of the first language and, above all, of the other languages spoken by the students.

References

- Abuhakema, G., Faraj, R., Feldman, A., i Fitzpatrick, E. (2008). Annotating an Arabic Learner Corpus for Error. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08). www.lrec-conf.org/proceedings/lrec2008/pdf/343_ paper.pdf
- Cremades, E. (2021, April 21). El Catalan Learner Corpus: una eina per desenvolupar la recerca i l’ensenyament del català com a llengua addicional [oral presentation]. Research seminar at the University of València.
- Cremades, E. & Javorová-Švandová, P. (2019). GramCat: Morfologie současné katalánštiny. Masaryk University.
- Granger, Sylviane. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation, in Aijmer, K. (ed.) Corpora and Language Teaching (pp. 13-32). Benjamins.
- Rodríguez, C. (2020). Anàlisi d'errors en català com a llengua estrangera en l'expressió oral d'estudiants universitaris txecs. Estudis Romànics, 42, 83-99.

II. NLP outreach

COMPUTATIONAL TOOLS AND RESOURCES FOR ROMANCE HISTORICAL LINGUISTICS

Liviu Dinu
University of Bucharest

Natural languages are living eco-systems, they are constantly in contact and, by consequence, they change continuously. Traditionally, the main Historical Linguistics problems (How are languages related? How do languages change across space and time?) have been investigated with comparative linguistics instruments. The main idea of the comparative method is to perform a property-based comparison of multiple sister languages in order to infer properties of their common ancestor. It is a time-consuming manual process that required a large amount of intensive work.

We propose here computer-assisted methods for identifying cognates, for discriminating between cognates and borrowings, and for protoword reconstruction.

The identification of cognates is a fundamental process in historical linguistics, on which any further research is based. Even though there are several cognate databases for Romance languages, they are rather scattered, incomplete, noisy, contain unreliable information, or have uncertain availability. We introduced a comprehensive database of Romance cognates and borrowings based on the etymological information provided by the dictionaries (the largest known database of this kind, in our best knowledge). We extracted pairs of cognates between any two Romance languages by parsing electronic dictionaries of Romanian, Italian, Spanish, Portuguese and French. Based on this resource, we proposed a strong benchmark for the automatic detection of cognates, by applying machine learning and deep learning based methods on any two pairs of Romance languages. Beside the largest database of this kind, we find also that automatic identification of cognates is possible with accuracy averaging around 94% for the more difficult task formulations. We also proposed a computational approach for discriminating between cognates and borrowings, one of the most difficult tasks in historical linguistics. We compared the discriminative power of graphic and phonetic features and we analyze the underlying linguistic factors that prove relevant in the classification task. We performed experiments for pairs of languages in the Romance language family (French, Italian, Spanish, Portuguese, and Romanian), based on RoBoCoP. To our knowledge, this is one of the first attempts of this kind and the most comprehensive in terms of covered languages.

Further we developed a methodology for protoword reconstruction and missing Romanian cognates reconstruction. Given words in Romance modern languages, the task is to automatically reconstruct the Latin proto-words from which the modern words evolved. We applied the method for producing related words based on sequence labeling, aiming to fill in the gaps in incomplete cognate sets in Romance languages with Latin etymology (producing Romanian cognates that are missing)

References:

- Alina Ciobanu, Liviu P. Dinu, 2019. Automatic Identification and Production of Related Words for Historical Linguistics. Computational Linguistics, vol. 45, No. 4, 667-704.
- Alina Maria Cristea, Liviu P. Dinu, Simona Georgescu, Mihnea-Lucian Mihai and Ana Sabina Uban, 2021. Automatic Discrimination between Inherited and Borrowed Latin Words in Romance Languages In: Proc. EMNLP 2021(Findings), Punta Cana, 2021
- Alina Maria Ciobanu and Liviu P Dinu, (2018). Ab initio: Automatic latin proto-word reconstruction. In Proc. COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018
- Alina Maria Ciobanu and Liviu P Dinu, 2014. Automatic detection of cognates using orthographic alignment. In Proc. ACL 2014, June 22-27, 2014, Baltimore, MD, USA
- Alina Ciobanu and Liviu P Dinu, 2015. Automatic discrimination between cognates and borrowings. In Proc. ACL 2015, July 26-31, 2015, Beijing, China
- Alina Ciobanu, Liviu P. Dinu, Laurentiu Zoicas, 2020. Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words. In Proc. LREC 2020 Marseille, France, 2020
- Ana Uban, Alina Ciobanu, Liviu P Dinu. Cross-lingual laws of semantic change. In: Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, Simon Hengchen (eds). Computational Approaches to Semantic Change. Berlin: Language Science Press., p. 219-260, 2021.
- Alina Maria Ciobanu, Liviu P. Dinu, 2017. Romanian Word Production: an Orthographic Approach Based on Sequence Labeling, In Proc. CICLing 2017, Budapest, Hungary, 2017
- Liviu P. Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Bogdan Iordache, Simona Georgescu, Laurentiu Zoicas, 2023. RoBoCoP: A Comprehensive ROmance BOrrowing COgnate Package and Benchmark for Multilingual Cognate Identification. In Proc EMNLP 2023

MACHINE TRANSLATION – FRIEND OF FOE FOR ACADEMICS?

Alina Radoi & Loredana Punga
West University of Timisoara

As free and fast-working tools, public machine translation platforms are used more and more often, in very diverse contexts, by a growing number of people, academics, constantly under the “publish (in English!) or perish” threat, included. However, frequency of MT use does not guarantee an appropriate translation quality, as, quite often, target texts contain errors - from minor to quite serious ones.

This paper reports on the results of a small-scale study that takes a comparative perspective on the quality of the translation of an academic text (in the area of linguistics) from Romanian into English, using three public MT translation platforms - Bing Microsoft Translator, DeepL and Google Translate. Errors, annotated using the CATMA digital tool, are considered from the point of view of their influence on the target text fluency, accuracy and fitness for purpose and analysed in terms of frequency and gravity. It hopefully comes to the aid of translators, MT systems developers and academics alike, by providing them with data to understand what kind of errors to expect, be ready to correct and, ideally, eliminate in the case of academic texts translated by the three MT systems considered, for the Romanian - English pair of languages.

BEYOND THE HEADLINES: INVESTIGATING LINGUISTIC VARIATION IN ROMANIAN FAKE NEWS AND NON-FAKE NEWS CORPORA

Iulia Arion, Anca Dinu & Livia Măgureanu
University of Bucharest

Research Question and Relevance:

This paper brings into discussion the topic of linguistic variation within a Romanian fake news corpus compared to a corresponding non-fake news corpus, exploring the applicability of a framework primarily grounded in functional theories of language use and research on register variation – as proposed by Grieve, J. and Woodfield, H. in “The Language of Fake News” (2023). The paper demonstrates that a linguistics-based approach to the study of Romanian fake news is not only promising, but necessary for the advancement of Romanian fake news research, especially since the existing literature on Romanian fake news in general is very limited – and almost non-existent when it comes to the study of the linguistic features of fake news. By applying an adapted version of the framework proposed by Grieve & Woodfield (2023) and analyzing a Romanian fake news corpus in relation to a corresponding non-fake news corpus, our primary goal is to uncover insights into the linguistic features and differences between the two corpora, showing that linguistic variation within Romanian fake news is a topic worthy of more attention.

Approach, Data and Methods:

Our research begins with a review of the existing literature on Romanian fake news, which shows that there is a significant research gap in linguistic exploration of fake news in Romania compared to English-speaking countries: the existing literature is mainly focused on non-linguistic aspects, with limited works focused on linguistic features. To address this gap, we started by compiling a Romanian fake news corpus and a corresponding non-fake news corpus using subjective selection criteria, given the absence of reliable detection systems for Romanian. We then adapt the framework proposed by Grieve & Woodfield (2023) - originally designed for English - to analyze linguistic variation in Romanian fake news: our study performs a linguistic analysis of fake news based on the theory of register variation and disinformation, while defining fake news in relation to the communicative purpose of the author(s) and using methods specific to register analysis.

A set of selected grammatical features which are chosen based on their relevance to the Romanian language is tested using the spaCy POS tagger and code written specifically for this tasks (in the absence of a Romanian alternative to the MAT program used in the original study).

Some of the tested features are: third person pronouns, average word length, adjectives, nouns, WH relatives, demonstrative pronouns, prepositions, subordinating conjunctions, verbs, adverbs, first person pronouns, time adverbs, suasive verbs, place adverbs and second person pronouns. To identify the differences in the values of these features we computed Cliff's delta, as the original study did. Additionally, our data is tested using a Random Forest Classifier to assess the potential of the approach for future research.

Last but not least, our paper provides a comparison between both the fake news features and the real news features identified in the Romanian fake news corpus and the Grieve & Woodfield (2023) paper.

Main Results and Interpretation:

The in-depth analysis of the 15 grammatical features examined so far reveals that 10 of them exhibit non-negligible differences between the non-fake and fake news, suggesting the need for more research on linguistic features in Romanian fake news. More linguistic features are currently being evaluated for this paper to better explain our approach both experimentally and theoretically.

One prominent example of linguistic variation within fake news so far is the presence of the third-person pronouns in the fake news corpus, indicating their use to create a sense of neutrality and fabricate situations or events. Based on the Cliff's delta, we notice that the relative order of the features as well as their magnitude changed (in comparison with the original study which we tried toreplicate for the Romanian fake news). This could mean that some of the features are more relevant to the Romanian language - and to our corpus, subsequently - than to the English language, and vice versa.

Our paper also identifies consistent trends across topics within the corpora, highlighting the potential value of content-topic-based analysis. The Random Forest Classifier achieved a reasonable accuracy score of 0.7684 - a result which is expected given the limited training data, showing the approach we used for the present study should be further explored and refined because the results can be promising. Currently, we explore how the Random Forest Classifier performs on various train / test splits of the corpora.

AUTOMATIC PROCESSING OF REAL-TIME RECORDED WRITING: PAUSAL SEGMENTATION VERSUS CHUNKING

Georgeta Cîșlaru (Université Paris Nanterre), **Iris Eshkol-Taravella** (Université Paris Nanterre) **& Sarah Almeida Barreto** (Université Sorbonne Nouvelle)

Due to keystroke-logging software, real-time recording of the writing process has become a valuable resource for psycholinguistics, linguistics and NLP, allowing for a better understanding of writing as a technology and as a socio-cognitive practice. While psycholinguistics is interested in the behavioral dimension related to cognitive functioning and linguistics seeks to understand the linguistic principles underlying writing processes, NLP approaches are confronted with a series of methodological questions related to the automatic processing of logging data. In this presentation we discuss some new applications for process analysis, based on a corpus or short texts produced by university students.

The writing process is globally organized into periods of production generating linguistic sequences called bursts, and pauses, a significant length of which – usually above 2 s – is considered to reveal complex cognitive processes. This perspective on pauses tends to justify the interpretation of bursts as process-significant segmentation units. In recent years bursts have seen increased interest from a linguistic perspective, searching for their qualitative and quantitative description. While some regularities have been observed between the boundaries of linguistic levels (word, sentence, paragraph) and sentence length, several studies have highlighted the syntactic inconsistency of sizeable percentages of bursts (eg. Cislaru & Olive 2018).

Chunking is an interesting entry, as it is defined both as a cognitive process for processing information (Johnson 1970) and as a unit of linguistic segmentation (Abney 1991). To meet the challenge of linguistic burst analysis, we are testing automatic chunking tools on textual data. So far, these methods have been applied to text and some of them have been developed for applications to oral texts (Eshkol-Taravella et al. 2020) which characteristics are similar to the writing process.

We understand the writing process as a flow of linguistic data interspersed with pauses. The idea was to align pausal segmentation (= bursts) and automatic chunking of the language stream. To do so, we isolated chunks and words within chunks into different classes and marked by an inserted symbol pause location (either in between chunks if the pause occurs at two chunks' frontiers, or inside a chunk if the pause occurs within a

chunk's frontiers). We distinguished three types of bursts: production, revision, immediate revision. The data were annotated (SEM) in order to identify the linguistic nature of the units potentially favoring pausal segmentation within a chunk.
Preliminary results show some regularities where chunk boundaries and burst boundaries align. 75 % of pauses occur between chunks' boundaries, with little variation when comparing pauses filtered by bursts' type. Among all grammatical categories, the noun seems to be the unit that favors a higher proportion of pausal breaks than the other categories (ex., pauses occurring before nominal phrases make up to 12 % of all nominal phrases). Also, strong punctuation constitutes separate chunks segmented by pauses.

This research has enabled us to settle some important steps for processing data recorded in real time to align behavioral and linguistic data.

References

- Abney S. (1991). Parsing by chunks. In R. Berwick, R. Abney, and C. Tenny (Eds.), Principle based Parsing. Kluwer Academic Publisher.
- Cislaru G., Olive T. (2018). Le processus de textualisation. De Boeck.
- Eshkol-Taravella I., Maarouf M., Badin F., Skrovec M., Tellier I. (2020). Chunk Different Kind of Spoken Discourse: Challenges for Machine Learning. Language Resources and Evaluation Conference, May 2020, Marseille, France, 5164-5168.
- Johnson N. F. 1970. The role of chunking and organization in the process of recall. Psychology of Learning and Motivation, 4, 171–247.

III. Culture, History and Language

THE MARKETING OF "BREATH"

Eugen Istodor
University of Bucharest

This paper analyzes the Romanian word “respiraţie” (”breathing”) in the virtual environment, in Google and YouTube contexts. I obtained approximately 2.200.000 results (0.27 seconds, quoted on Sept. 15th, 2023). The term “respiraţie” is defined as such (a fundamental physiological process through which organisms exchange oxygen and carbon dioxide with the environment) in a limited number of contexts, in the tens.

“Breathing” thus surpasses the limits of a simple notion, the notion that names the natural gesture of survival. “Breathing” receives the connotation of survival in a hostile world. Every day’s stress, the difficulties of daily confrontations find a remedy in the simple miracle of breathing. The present research will discuss the diversity of breathing techniques, but also the need to adapt them to nowadays’ “oppressive” society. Posthumanity (Braidotti, 2013) and its nomad ethics (Braidotti, 2006) estranges the community from the religious and philosophic sense of Christian and European “breathing” (Skof, 2015) and introduces it in the Oriental circuit, transforming it into a technique. “Breath” becomes an algorythm that is easy to get considering its significant details.

An algorithm which, at the same time, as we may notice from the Google and YouTube search results, becomes a form of “breath” marketing. The digital approach of “breath” includes the main transformational processes. “Breath” becomes an inner transformation process for making self-understanding and also understanding others more efficient. “Breath” is included in Business model transformation, domain transformation, cultural/organizational transformation. It becomes a form of learning and survival through the means of algorythmic methods presented virtually.

We thus find out that the popularity of a word is determined by the global connection practice and has consequences in social behaviour. Breathing becomes a form of success recommended in the cultural and business approach to existence. We also discover the solution for the oppression of a social proximity from the globalist society.

Last but not least, breathing is also the solution when faced with “paradoxical happiness” oppression (Lipovetsky, 2007). “Breathing” becomes the need to once again find the lost identity in a world of unfulfilled or wrongly interpreted promises (Fukuyama, 2022).

Yielding to a nomad ethics, to a nomad social content and relations (Braidotti, 2006), the individual is in a permanent state of confusion. Even when pausing to breathe, trying to save himself, he doesn't seem able to escape. The human organism feeds itself with breathing from a toxic environment and it loses its supremacy, becoming one of the integrated mechanisms, equal to the environment and to the hi-tech universe. It is in a permanent state of fluidity and transformation (Braidotti, 2006), that always imposes the kind of breathing that is not redeeming, that doesn't find its way to identity.

DIGITAL HISTORIES OF PHILOSOPHY AND SCIENCE: THE EARLY MODERN PERIOD

Mihnea Dobre
University of Bucharest

In 2019 a special issue of one of the leading journals in the history of science, Isis, was dedicated to the topic of computational history, a discipline that was presented as complementing traditional scholarly perspectives with the help of digital humanities approaches in order to enrich historians' perspectives upon scientific change and practices. The early modern period is constantly depicted as an illustrative episode for the intricate dialogue between philosophy, science, religion, and society at large, leading to what was called "The Emergence of a Scientific Culture" (Gaukroger 2006) or "The Rise of the Modern Science" (Cohen 2015). Hence, in recent years, it has becoming more urgent to address the following question: how and why do histories of philosophy and of sciences in the early modern period use digital humanities approaches? This is a broad and complex question, not only due to the large variety within the field of early modern studies, but also due to the diversity and the lack of standardization of the digital tools and approaches. The paper aims to explore some of the recent trends, moving beyond the digitization of the sources. In particular, three topics will be examined: (1) textual analysis of large – and often unread – corpora; (2) the role of visualizations in advancing (new) research questions; (3) the use of network science to expand the scope of traditional investigations. Examples derived from recent projects in the field will be discussed (e.g., The Newton Project; Cartesian Cosmological Illustrations; the project "The Sphere. Knowledge System Evolution and the Shared Scientific Identity of Europe"; or the project "The normalisation of natural philosophy. How teaching practices shaped the evolution of early modern science"), while raising questions about data collection and the problem of selecting early modern resources; the relation between traditional scholarship and digital approaches; or the format of research results.

THE DIGITAL TURN IN TEACHING COMPARATIVE LITERATURE

Gabriela Glăvan
West University of Timisoara

As it has been recently argued (Li, 2022), there is a solid background for establishing connections and similarities between Comparative Literature and Digital Humanities. I propose a double analysis in my paper - a commentary on the increasing relevance of the theorizing efforts concerning the intersection between the two domains and a brief overlook of a potential case study in teaching comparative literature with the aid of strategies, methods and technologies pertaining to Digital Humanities.

The COVID pandemic marked a shift in the ways literature has been taught in recent years, as students of all academic levels were obliged to conform to an online regime of teaching and learning. If digitalization was a necessary and ultimately unavoidable phenomenon in teaching Humanities, the two-year period in which many schools and universities offered courses exclusively online meant an accelerated turn towards digital teaching methods and means. What was gained and what was lost? How could such a generously positioned domain, as Comparative Literature is, in the framework of the Humanities, benefit further from employing digital elements? Can we speak of an enrichment of Digital Humanities due to Comparative Literature? What concrete strategies have educators and specialists involved in their experience of teaching Comparative Literature using digital means? Are there evident drawbacks and limitations of this digital turn in a domain that has long fought for scientific legitimation?

My investigation will be a Full Paper and it is a simultaneously theoretical and personal approach, as I believe the articulation of the digital dimension in the Humanities should preserve a solid amount of vitality and involvement that should transcend the often cold and impersonal nature of technology.

References

- Li, Q. Comparative literature and the digital humanities: disciplinary issues and theoretical construction. Humanit Soc Sci Commun 9, 437 (2022).

LATE ANTIQUE LETTERS AND NETWORK ANALYSIS – APPROACHES AND CHALLENGES. CASE STUDIES: THE SOCIAL NETWORKS OF THE CAPPADOCIAN FATHERS AND JEROME OF STRIDON

Andra Jugănaru
University of Bucharest

The aim of this paper is to compare two different approaches used in researching two different epistolographical corpora written at the end of the fourth century AD using both quantitative and qualitative tools.

The first corpus consists in the letter collections of the Cappadocian Fathers. Basil of Caesarea (ca. 330 - 378), his younger brother, Gregory of Nyssa (ca. 335 – ca. 395), and their friend, Gregory of Nazianzus (329 - 390) were not only sending and receiving letters, but also organized them in collections. These three distinct, but tangled collections pose several challenges to a researcher attempting to investigate them. First, their sizes (i.e. the number of letters in each collection) are very different, mainly due to the fact that most letters authored by Gregory of Nyssa were lost during Gregory's exile. Another challenge occurs in attempting to establish the authorship of certain letters, as in some cases copists mixed up Gregory of Nyssa and Gregory of Nazianzus, or assigned to Basil some letters of the two Gregories. In addition, identifying the names mentioned in the letters could prove to be a difficult task, due to the repetition of certain names. Besides, some letters transmitted in the collections, although not authentic, are still valuable for investigating the social networks of the Cappadocians, as they are relevant for the intellectual environment of their authors.

The second corpus consists in the letter collection of Eusebius Sophronius Hieronymus of Stridon (ca. 342/347 – 420), known in English-speaking countries as Jerome. Jerome also organized his letters in a collection. Researching this particular collection poses mostly problems of dating and establishing the identity of the characters mentioned.

By using both qualitative and quantitative methods, I aim at overviewing the ties established either between individuals, or between individuals and communities. Assessing the strength of these ties and the positions of each actor within a network (or several networks) is another objective of the research.

In the final part of the paper, I will present two tools used for collecting data (e. g. actors' attributes, relational data, data about the letters), calculating measures related to actors, and visualizing the networks: UCI.NET and ORA-LITE.

The objective of this comparison is to answer the following questions: Which is the best tool for determining the roles played by each actor in a network? How strong are the links between these actors and what is the hierarchy of these connections? How do such networks reflect the dynamics of the ecclesiastical arena at the end of the fourth century?

EMOTIONS, LANGUAGE, ECONOMY: HOW DO WE EXPRESS BELONGING THROUGH EMOJI AND ABBREVIATIONS IN THE DIGITAL WORLD?

Mira Bekar & Andrijana Kjose
Ss Cyril and Methodius University in Skopje

Online communication has outclassed face-to-face communication, becoming a common and preferred form of social interaction. This has been a result of personal choice and of Covid-19 pandemic isolation. The processes of establishing and maintaining social relationships for both personal and professional goals have been transferred significantly to online venues and modes. Our research focuses on the use of emoji and abbreviations in English and Macedonian in social media as a tool for achieving economy of language and showing sense of belonging. Research has shown that emoji are in fact evolving into a separate language which is specific for its graphic features (Ge & Herring, 2018; Monti et al., 2016) and will be soon universally used and understood (Ai et al. 2017). The relevance of our study is that close discourse analysis of everyday social activities such as online chatting can help us understand how we develop personally and professionally as human beings, i.e., how we use language as social action (Bekar, 2015).

Two major aspects were addressed in the research: 1) emoji and abbreviations being used as a type of cryptic language which enables belonging to a certain online community and 2) the use of emoji and abbreviations as a means for economy of language – that is, whether abbreviations maintain and strengthen their primary use. In this mixed-method research, 15 participants with 10 conversations were included, which means 150 conversations that lasted 10 to 15 minutes, which created a large corpus for discourse and conversion analysis. We analyzed conversations (chats) and reactions on major social media by English and Macedonian speakers. These significantly long texts in an uninterrupted time sequence were taken from any application such as Viber, Reddit, Instagram, Telegram, Facebook Messenger and Twitch. Since this is work-in progress, we are going to present the preliminary findings. They show that 1) economy of language is the primary reason for using abbreviations and emoji since people have less and less time; 2) the emotional state of participants and their mutual relationship in real life dictates the conversation; 3) cryptic and context-bound communication defines the sense of belonging. Ideas for future research and strategies for avoiding the sense of fear that young generations are destroying standard variations of languages will be presented.

References

- Bekar, M. (2015). Language, writing, and social (inter)action: An analysis of text-based chats in Macedonian and English. (Doctoral dissertation). Purdue University, West Lafayette, IN.
- Ge, J., & and Herring, S. C. (2018). Communicative Functions of Emoji Sequences on Sina Weibo. First Monday 23(11).
  https://firstmonday.org/ojs/index.php/fm/article/view/9413/761
- Monti, J. et al. (2016). Emojitalianobot and EmojiWorldBot. New Online Tools and Digital Environments for Translation into Emoji. Paper presented at the Third Italian Conference on Computational Linguistics, Naples, Italy, December 5-6.

TEXTUAL REPRESENTATIONS OF ARTEFACTS AND ARTEFACT INSPIRED TERMINOLOGY IN CLASSICAL ANTIQUITY. CASE STUDY: GLASS AND GLASS ARTEFACTS

Alexandra Litu
University of Bucharest

Many texts in Classical Antiquity include descriptions of artefacts, either as an aim in itself (the so-called ekphraseis), either in narratives with a different focus. On the other hand, through archaeological research have been uncovered big quantities of artifacts made of various materials: metals, glass, clay etc. Various types of artefacts were produced in particular periods and their circulation had highs and lows; also their centers and techniques of production, composition, shapes, uses, quality and characteristics changed over time as well as the lifestyle choices of their users. We are interested to investigate if the changes visible in the archaeological record go hand in hand (or not) with potential changes discernible in the literary discourse about artifacts. It seems to us an interesting question since there are objects, present in texts, whose existence is entirely imaginary as far as we can say. In this respect, we have chosen to investigate first glass and glass artefacts in the literary discourse and in the archaeological record mostly during the Roman Empire. We concern ourselves with the way glass and glass artefacts are described, with their collocations and especially with the uses for the adjectives derived from the terminology of glass or implying comparisons with glass. To give an example, archaeology shows that the characteristics of glass are changing over time, it can be transparent, translucent or opaque. It would be interesting to see if the adjectives follow these changes or if there is a literary constructed image of how glass and everything associated with glass should be. In order to facilitate the digital analysis of texts we will use two text collections, the LATIN ISE on Sketch Engine (McGillivray, B. and Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, SilkeScheible, Richard J. Whitt (eds.), New Methods in Historical Corpus Linguistics. Tübingen: Narr) and the Perseus Digital Library.

A DIGITAL APPROACH TO THE ANCIENT GREEK LEXICON: WHAT'S NEW?

Constantin Georgescu, Simona Georgescu & Theodor Georgescu
University of Bucharest

The Greek-Romanian dictionary written by Constantin Georgescu, Simona Georgescu and Theodor Georgescu (Dicţionar Grec-Român, work-in-progress), unlike many others published so far in different languages, has the chance to make use of the latest developments in information technology. We aim to show how we can beneficially combine digital research methods with traditional philological acumen.

The first part of the project consisted in finding a software program to host this lexicon. We chose to adapt the needs of an ancient Greek dictionary to the tlTerm software platform: Terminology Management Software, which provides a very flexible starting point for various types of scientific work. Basically, starting from a computer matrix applicable to any lexicographical research, we have built a special program, which has the capacity on the one hand to incorporate a huge amount of information in the long term, with the possibility of increasing the number of authors and works, and on the other hand, to be a fast and flexible tool for editing work. Digitizing the information in this format will allow the dictionary to be used both in the classic printed form and in a virtual (online) environment, or in the form of compatible applications on personal computers, tablets, mobile phones, etc.

Once the host program for this dictionary was set up, we established the working method for the actual development of the headings. In this case, too, electronic means proved to be extremely useful. Since the aim of the dictionary is to treat a number of authors exhaustively, it was necessary at first to inventory the entry words. In addition to the lists provided by already published dictionaries, which we will mention below, there are now several programs that can index texts (relatively) automatically, providing alphabetical lists of words. In the specific case of the Greek language, the situation is made more difficult by morphemes (augment, reduplication) which are placed before the lexical root and whose removal is obligatory in order to reach the lemma.

The inventory of entries was followed by the analysis of each lexeme individually. As a fundamental working principle, examples were selected and meanings were established from the primary source, the texts of (so far) 30 authors in the most recent editions accessible in digitized form. Thus, each lemma was searched by electronic methods in the authors' corpus. For this purpose we used the constantly updated TLG database, in particular with the help of the Diogenes and Workplace Pack administration and search programs, with variants tailored to the needs of this dictionary. However, words were not analysed in isolation but in context, so that their most appropriate meaning could be determined.

ADVANCES IN CULTURAL HERITAGE THROUGH THE LENS OF INTELLECTUAL PROPERTY: A PATENT TREND ANALYSIS

Rajesh Kumar Das
Noakhali Science and Technology University

Background: With the mass progression of science and cutting-edge technologies like artificial intelligence, augmented reality, virtual reality etc., patent has now been considered as great source of technological innovation in todays' world. In the field of cultural heritage (CH), it offers constitutional protection and legal rights of art, culture and heritage through democratization of cultural policies from an individual to international level. But in recent years, "nationalism" or "retentionism" approach of the cultural heritage law challenges the universal access and protection of cultural treasures or common goods [1]. Therefore, a comprehensive analysis and visualization into the nature and trend of the patent is needed to determine the subtlety and overlaps between intellectual property and cultural heritage.

Rationale and Objectives: Patent as an intellectual property provides legal rights aiming to protect the fruits of the human intellect in the field of industry, literature, art and science. Thus, patent analysis is a unique way to study of intellectual and technological innovation to help the scientific community in getting an overview on the patent state of the art on any particular field. As evidenced in the literature, comprehensive measurement of research progress around patents in this research field is still lacking. Therefore, this study is aimed to- i) trace the growth trend of patents in CH field on the basis of their families, jurisdictions, applicants, inventors, owners, and classification scheme; ii) visualize network map of inventors, owners, countries, keywords etc.; iii) prediction future innovation pathway of patents in the cultural heritage field.

Methodology: To achieve its objectives, this study employs a quantitative bibliometric and data mining approach. Data were collected from lens patent database (https://www.lens.org/) on August 7, 2023 using the search string TITLE, ABSTRACT OR CLAIMS: (cultur*) AND (heritage) and YEAR: All. After screening, a total of 738 patents were selected and analyzed using statistical methods, citation analysis, co-occurrence and cluster mapping methodologies.

Results:

Preliminary results indicate that the highest number of patent is filed (96) and also granted (59) in 2021 on the area of CH. Also, there are 242 active and 182 pending patents, while other 189, 81 and 43 patents are discontinued, inactive and expired respectively during the studied period. According to the findings, the top 10 jurisdictions containing the applicants and inventors are: China, Korea, United States, WO–WIPO, Romania, European Patents, Japan, Bulgaria, Russia and Australia. From the citation analysis, it is found that the highly cited patent is titled “CULTURAL INTERACTIVE PANORAMA (CULTURAMA)” holding the CPC code G03B37/04. The assignees institutions applying highest in this field of study are: Elwha LLC, Centre National De La Recherche Scientifique, Elwha LLC a Limited Liability Company of the State of Delaware, President and Fellows of Harvard College, Artglass USA LLC, Consiglio Nazionale Delle Ricerche etc. Cluster analysis shows that the most dominant cooperative patent classification (CPC) schemes are: G06Q50/10, A47F3/002, A47F3/125, A47F3/001, G06Q50/26 etc. This work-in-progress study seeks to accomplish rest of the findings to get more compendious result.

References

- Stamatoudi, I. (Ed.). (2022). Research handbook on intellectual property and cultural heritage. Edward Elgar Publishing.

A NEUTROSOPHIC-BASED APPROACH TO IDENTIFY YOUNG PEOPLE’S EDUCATIONAL ATTITUDES AND BEHAVIOURTOWARDS ACTIVE ENGAGEMENT

Mihaela Colhon, Monica Tilea & Alina Resceanu
University of Craiova

In this paper, we apply a neutrosophic model in order to identify certain attitudinal and behavioural patterns of actively engaged young people based on answers given in questionnaire-based surveys. To this aim, we use the results obtained from administering a questionnaire drawn up by a group of researchers from 6 European countries in the European project Mindchangers: Regions and Youth for Planet and People and investigate what set of characteristics could influence or not young people’s likelihood to become mindchangers. The questionnaire was designed to determine youths’ awareness about the Sustainable Development Goals and their engagement as active agents of development and change at regional level. Two types of variables were used: nominal and ordinal, in the form of Likert-type questions with four and five-point rating scales. The significant level of uncertainty corresponding to certain answers offered in this questionnaire, such as rather agree/rather disagree, led us to implement this model for data representation and processing. In short, the model allows us to accurately process the respondents’ answers in order to extract the possible sets of the respondents’ characteristics selected from the questionnaire answers and, based on these, to corelate them to the target question (in our case, the Mindchanger question) thus establishing a connection with the respondents’ likelihood to become Mindchangers.

RADH
2023

The Second
International Conference
on Recent Advances
in Digital Humanities

www.radh2023.uvt.ro