Lang Resources & Evaluation (2019) 53:559–563
https://doi.org/10.1007/s10579-019-09482-4

EDITORIAL

Language technology for digital humanities: introduction to the special issue

Erhard Hinrichs[1] · Marie Hinrichs[1] · Sandra Kübler[2] · Thorsten Trippel[1]

Published online: 19 November 2019

The use of digital resources and tools across humanities disciplines is steadily increasing, giving rise to new research paradigms and associated methods that are commonly subsumed under the term *digital humanities*. Digital humanities does not constitute a new discipline in itself, but rather a new approach to humanities research that cuts across different existing humanities disciplines. While digital humanities extends well beyond language-based research, textual resources and spoken language materials play a central role in most humanities disciplines.

In order to showcase the use of language tools and resources in digital humanities research, the LT4DH (Language Technology for Digital Humanities) workshop was held at COLING 2016 in Osaka, Japan. Discussions focused mainly on the following topics:

– Case studies of using language technology and/or language resources with the goal of finding new answers to existing research questions in a particular humanities discipline or addressing entirely new research questions.

✉ Marie Hinrichs
marie.hinrichs@uni-tuebingen.de

Erhard Hinrichs
erhard.hinrichs@uni-tuebingen.de

Sandra Kübler
skuebler@indiana.edu

Thorsten Trippel
thorsten.trippel@uni-tuebingen.de

[1] Seminar für Sprachwissenschaft, University of Tübingen, Tübingen, Germany

[2] Department of Linguistics, Indiana University, Bloomington, IN, USA

- Case studies of expanding the functionality of existing language processing tools in order to be able to address research questions in digital humanities.
- The design of new language processing tools as well as annotation tools for spoken and written language, showcasing their use in digital humanities research.
- Domain adaptation of rule-based, statistical, or machine-learning models for language processing tools in digital humanities research.
- Challenges posed for language processing tools when used on diachronic data, language variation data, or literary texts.
- Showcasing the use of language processing tools in humanities disciplines such as anthropology, gender studies, history, literary studies, philosophy, political science, and theology.

The motivation for this special issue was to provide a means for presenting work in these areas in greater detail. Submission was open to anyone working at the interface of computational linguistics, digital humanities, and social science disciplines. Furthermore, all participants of the LT4DH workshop were invited to submit extended versions of their work.

The collection of articles accepted for publication in this special issue covers a wide range of topics relevant to researchers in the digital humanities. Our main objective for this introduction is to relate the individual contributions to larger, on-going research themes in digital humanities and to highlight the role of natural language tools and resources in each article.

The articles can be grouped into three thematic clusters: text analysis, corpus building, and corpus annotation. The order in which the articles appear follows this grouping, starting with the largest cluster of six articles that address the use of language technology for different aspects of text analysis:

- *Computational Text Analysis within the Humanities: How to Combine Working Practices from the Contributing Fields?* by Jonas Kuhn.
- *Dialogue Analysis: A Case Study on the New Testament* by Chak Yan Yeung and John Lee.
- *Vector Space Explorations of Literary Language* by Andreas van Cranenburgh, Karina van Dalen-Oskam, and Joris van Zundert.
- *Geoparsing Historical and Contemporary Literary Text set in the City of Edinburgh* by Beatrice Alex, Claire Grover, Richard Tobin, and Jon Oberlander.
- *Token-based Spelling Variant Detection in Middle Low German Texts* by Fabian Barteld, Chris Biemann, and Heike Zinsmeister.
- *Beyond Lexical Frequencies: Using R for Text Analysis in the Digital Humanities* by Taylor Arnold, Nicolas Ballier, Paula Lissón, and Lauren Tilton.

The cluster of articles on text analysis begins with a position paper, followed by three use-cases in religious studies, linguistics, and literature, and ends with two articles about NLP tool building. The position paper by Kuhn compares and contrasts the traditional hermeneutics paradigm for humanities research with recent

data-driven, computational approaches. On the basis of his own research experience, Kuhn proposes a new methodological framework that tries to bridge these two research paradigms. The article ends with two experimental scenarios that illustrate the use of this new integrative framework. The two scenarios are taken from the domains of corpus linguistics and literary studies.

The contribution by Yeung and Lee presents an automated approach for analyzing dialogues in the New Testament. The authors have developed a machine learning approach for identifying dialogues in three steps: First, they identify speakers and listeners, then they detect chains of quotes with alternating speakers/listeners, and finally they determine the boundaries of complete dialogues. They use automatic POS tagging, dependency parsing, and named entity recognition to create features for the machine learner. Based on these extracted dialogues, Yeung and Lee present a quantitative analysis of the dialogues.

The article by van Cranenburgh, van Dalen-Oskam, and van Zundert investigates to what extent machine learning (ML) approaches can successfully predict the degree of literariness of a novel. More specifically, the authors utilize two types of unsupervised ML models that are widely used in natural language processing: a topic model and a neural vector space model which are trained on different text passages of 2–3 pages in length. They show how different notions of semantic complexity can be derived from these models and investigate how well these complexity measures correlate with the literacy ratings of Dutch novels that were collected by an on-line survey.

Data-driven analysis of literary text is also the topic of the article contributed by Alex, Grover, Tobin, and Oberlander. The authors adapt the Edinburgh Geobrowser, an NLP tool for automatic enrichment of textual materials with geographical information, for use with historical literary texts set in the city of Edinburgh. The tool allows fine-grained annotation of street names, monuments, and other landmarks. The quality of the automatic annotation is evaluated against a gold standard. The tool has a modular architecture and is therefore easily adapted to other geographical locations.

A recurrent theme in the digitization and use of historical text corpora concerns wide-spread spelling variation for the same lemma. Barteld, Biemann, and Zinsmeister offer a new computational approach to dealing with this issue for Middle Low German texts. Contrary to most studies that deal with the phenomenon of spelling variation, the authors of the present article do not attempt to convert different spelling variants to a single, normalized form. Rather, their spelling variant detection approach generates all potential spelling variants for a given lemma and filters the set of potential variants by systematically inspecting the linguistic contexts of the spelling variants that occur in the text.

Due to its focus on data analysis and visualization, its large number of processing packages, and an active user community, the statistical computing language R is well suited to text analysis tasks and is gaining popularity in digital humanities. The article by Arnold, Ballier, Lissón, and Tilton presents a collection of R packages, built around a common text interchange format, to be used in digital humanities workflows. They demonstrate the power and usefulness of the ecosystem, which includes NLP tools, in a digital humanities project.

The second thematic cluster of articles on corpus building begins with an article about a spoken-language corpus, followed by two articles presenting text corpora for Arabic and Hebrew. The articles are as follows:

- *Digitising Swiss German—How to Process and Study a Polycentric Spoken Language* by Yves Scherrer, Tanja Samardžić, and Elvira Glaser.
- *Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus* by Yonatan Belinkov, Alexander Magidow, Alberto Barrón-Cedeño, Avi Shmidman, and Maxim Romanov.
- *Historical Corpora meet the Digital Humanities: The Jerusalem Corpus of Emergent Modern Hebrew* by Aynat Rubinstein.

The contribution by Scherrer, Samardžić, and Glaser reports on the creation of ArchiMob, a spoken language corpus of Swiss German variants. The corpus is based on transcribed spoken text of historical narratives. In the creation of the corpus, it was necessary to adjust tools and training material, originally developed for standard German, to improve the results on the language variants in the corpus. These tools include automatic speech recognition, spelling normalization, and POS tagging. The article concludes with three case studies, showcasing the usefulness of the corpus for a range of digital humanities questions. The lessons learned in creating the ArchiMob corpus can be generalized to other languages.

Belinkov, Magidow, Barrón-Cedeño, Shmidman, and Romanov describe the creation of a large scale diachronic corpus of written Arabic, automatically annotated for sentence boundaries, morphological segments, lemmas, POS tags, and constituent syntax. They then develop a computational methodology to cluster the diachronic texts into periods. In a final analysis of the results of their periodization algorithm on the corpus, Belinkov et al. not only confirm the established periodization of Standard Arabic into Classical and Modern Standard Arabic, but they also find evidence for a more differentiated periodization.

Rubinstein describes the process of creating an open-access corpus of Emergent Modern Hebrew, which includes extensive metadata and linguistic annotations. Throughout the process, care was taken to follow best practices and to comply with standards in the digital humanities. His article shows how the use of NLP tools, in combination with crowd sourcing and collaboration with external partners, made the construction of the resource possible. Use-cases are presented to demonstrate the use of the corpus in diachronic linguistic research.

The third thematic cluster on annotation starts with two articles about POS tagging, and ends with a project note on WSD. The articles are as follows:

- *From 0 to 10 Million Annotated Words— Part-of-Speech Tagging for Middle High German* by Sarah Schulz and Nora Ketschik.
- *Exploiting Languages Proximity for Part-of-Speech Tagging of three French Regional Languages* by Pierre Magistry, Anne-Laure Ligozat, and Sophie Rosset.

– *Approaching Terminological Ambiguity in Cross-disciplinary Communication as a Word Sense Induction Task. A Pilot Study.* by Julie Mennes, Ted Pedersen, and Els Lefever.

The article by Schulz and Ketschik addresses issues similar to the ones addressed by Barteld et al.: Schulz and Ketschik investigate approaches to POS tagging for Middle High German given the non-standard spelling. Their investigation addresses questions of the amount of training data, and methods for integrating information from a lexical database and another corpus, annotated for POS, but using a different tagset. Additionally, they investigate domain adaptation in order to improve the performance of their POS tagger on their region-specific corpora.

Magistry, Ligozat, and Rosset also address the issue of POS tagging, but they investigate methods for POS tagging the regional languages Alsatian, Occitan, and Picard. For these languages, no resources exist, thus Magistry et al. leverage resources from the related, high-resourced languages German, Catalan, and French using delexicalization and transposition of highly frequent words.

Linguistic enrichment and disambiguation of text corpora is by no means limited to morphological and syntactic information. Mennes, Pedersen, and Lefever present a pilot study for sense clustering of ambiguous terms. The authors point out that such ambiguities can make communication difficult among researchers from different scientific disciplines in cross-disciplinary investigations, including in digital humanities. Mennes et al. conduct a pilot study for automatically inducing different sense clusters for ambiguous terms. Their study is based on the NLP software package Sense Clusters, previously developed by Pedersen.

We hope that this brief preview of each article will prove useful for navigating through this special issue and will stimulate readers to consult the individual contributions. We would like to take this opportunity to thank all authors for their contributions to this issue, and all referees who kindly agreed to review the submitted manuscripts for their in-depth comments and helpful suggestions.