Introduction to the Special Issue on Digital Humanities and Computational Linguistics

John Nerbonne*
University of Groningen

Sara Tonelli**
Fondazione Bruno Kessler, Trento

1. Digital Humanities

Digital Humanities (DH) seeks to support research into Humanities disciplines using digital, computational techniques. Its exact definition is discussed often and may even be the subject of interesting debate (Vanhoutte, Nyhan, and Terras 2013), but we do not need to linger too long on definitional issues. At this time, DH invites contribution from all Humanities disciplines, including those where language plays a secondary role, such as anthropology, archeology, fine (visual) arts, film studies, and musicology. These are not the most likely disciplines for computational linguists to get involved in, but linguistics and literature (studies) are also Humanities discipline, where language is central, as are history and philosophy, where language is not of central interest, but where archival material in textual form often plays a central role. There are enormous opportunities for contributions from computational linguistics (CL) from all the disciplines where language and text are important.

Just as in other computational disciplines, the fundamental benefits that DH can bring to its non-computational parent disciplines are the ability to deal with large amounts of data, the speed with which analyses can be tested, assessed, and criticized, and finally, the commitment to well-codified procedures, which can better be tested, replicated and modified. All of these benefits are being realized in some projects even today. Jockers (2013) analyzes 3,500 American, Irish and English novels of the nineteenth century, exploring especially the trends in themes over this period, e.g. when the religious themes of sin and salvation were popular in the different countries. Estimating conservatively, 3,500 novels would require about 100 meters of shelf space and reading — but not yet taking notes and analyzing them — would take over ten years for a disciplined reader, reading a novel a day. As larger amounts of material become available, so too will the scope of projects such as Jocker's. Speed is of course related to the first advantage, that of capacity, since the capacity would be pointless if analyses could not be produced promptly.

Nerbonne et al. (2011) describe Gabmap, a web application for dialectology. Gabmap requires that users input dialect data in the form of a table organized into sites on the one hand and forms that vary on the other. A given cell contains an indication of which form is used at a given site. The data may be categorical such as lexical or syntactic choice, numerical such as the formant frequencies of vowels, or strings such as pronunciation transcriptions. Analyses may be categorical (same vs. different), numerical (Euclidean distance), or, for strings, edit distance. Once data is

* University of Groningen, Oude Kijk in 't Jatstraat 26, Groningen, The Netherlands. E-mail: j.nerbonne@rug.nl
** Fondazione Bruno Kessler - Via Sommarive 18, Trento, Italy. E-mail: satonelli@fbk.eu

uploaded to Gabmap, a range of calculations is executed on a remote server, including measures of the consistency of the data, clustering and multidimensional scaling. If a map is additionally uploaded, results are projected onto the map in a variety of forms, including maps indicating the areas of clusters in the dendrograms, maps where color is determined by a mapping of the most important multidimensional scaling (MDS) dimensions to red, green and blue hues. While this is normally completed within a minute or two on a good server, users definitely appreciate seeing results very quickly and, e.g., being able to compare clustering and MDS results directly (Leinonen, Çöltekin, and Nerbonne 2016).

Finally, analyses that rely on software are transparent in ways that non-computational analyses mostly fail to be — assuming, of course, that the code is openly available for inspection, replication of results and use in modified form. Although DH is still quite young, text analysis has always been a core activity, and there is a longish history of software tools developed for this purpose (Bradley 2004), many of which continue to be available. STYLOMETRY is the subdiscipline devoted to studying authorial style, i.e. what is relatively distinctive about a given work or a given author. It is challenging, particularly since an author's style may develop over time and may also vary depending on the genre on examines (e.g., novels, short stories, journalism, letters, or scholarly essays) but also depending on subject matter and the 'voices' of the characters being portrayed. AUTHORSHIP ATTRIBUTION , sometimes known as 'non-traditional authorship attribution' is a sub-field of stylometry which has flourished in part because it can test its ideas against very clear criteria — whether or not an analysis can identify the author of a text withheld from use in training. Naturally other parameters, such as those mentioned above, but also the number and similarity of alternative candidates, can be significant. `Stylo R` (Eder, Rybicki, and Kestemont 2016) is an R package[1], and since all of R is open source, all of the code in Stylo R is as well. It was written by Maceij Eder, Jan Rybicki and Mike Kestemont, three of the leading practitioners of stylometry active today. A number of papers cite it already, and although it is too early to say for certain, it has the potential to become the sort of 'evolving standard' that attracts criticism but also contributions from others. This sort of package is well known to computational linguistics, who can point to open projects such as the GATE framework, Giza++ for machine translation or the Stanford parser[2] (Cunningham et al. 2002; Och and Ney 2003; Klein and Manning 2003).

2. DH and Computational Linguistics

The examples we have chosen above illustrate some of the opportunities for CL practitioners who wish to experiment in DH. Jocker's work on identifying themes in novels (Jockers 2013) relies on topic modelling with LATENT DIRICHLET ALLOCATION (LDA) (Blei, Ng, and Jordan 2003), a technique regularly used in CL works. Gabmap is an open-source web application that analyzes language variation across different levels (lexical, phonetic, phonological as well as syntactic), and papers on the work Gabmap is based on have also appeared at CL conferences (Nerbonne and Heeringa 1997; Nerbonne 2003). Finally, there is now a regular workshop series on computational linguistics for literature (Elson et al. 2012), which held a fifth annual meeting in 2016 (Feldman,

1 See the R project for statistical computing, https://www.r-project.org/

2 See `https://gate.ac.uk/`, `http://www.statmt.org/moses/giza/GIZA++.html`, `http://nlp.stanford.edu/software/lex-parser.shtml`

Kazantseva, and Szpakowicz 2016). The DH community has also shown substantial interest in collaboration. Van Dalen-Oskam et al. (2014) has experimented with using named entity recognition (NER) to find geographical references in text to compare literary works based on that. And Eder (2015) has added an examination of sequences of part-of-speech tags to the arsenal of techniques used in studying authorship attribution.

All the papers included in this special issue represent well the strong interdisciplinary approach and the broad range of topics mentioned above, which can be observed both in the theoretical/methodological and in the project-oriented articles. The range of disciplines covered from a computational perspective include among others historical linguistics, past and contemporary political studies, translation studies, linguistic resources for Latin. Besides, even if this issue will appear in the 'Italian journal of Computational Linguistics', Italian is just one of the languages covered by the published works, showing an interest in multilinguality and language diversity that is usually less evident in the English-oriented CL community. The list of accepted works shows this variety of topics and perspectives. The article by Clemente and Passarotti deals with the problem of comparing existing lexical resources, and propose a normalised coefficient to compute the degree of overlap between a Latin valency lexicon and Latin WordNet. The issue of exploring parsed historical corpora is instead tackled by Ingason, who presents both a tool and a methodology to analyse syntactic variations in historical corpora. The use case is focused on Icelandic, but the approach can be generalised to historical corpora in different languages. The third methodological paper, by Lauscher et al., has the goal to improve the use of topic modelling, which is widely-used for corpus exploration in humanities studies, by combining entity linking and labeled LDA.

Three other articles in this issue focus on past and ongoing projects using computational linguistics methods for DH studies. The article by Bellandi et al. presents the *Traduco* system, a tool for computer-assisted translation developed to support the translation of the Babylonian Talmud. The project posed several challenges, such as the need for a strictly controlled editing process of different translations and the highly complex nature of the book content. The article by Sprugnoli et al., instead, presents an ongoing project on the computational analysis of Alcide De Gasperi's corpus of public documents. The analyses are the outcome of a two-year collaboration with history scholars, and include ad-hoc visualisations after the corpus was processed with the ALCIDE system (Moretti et al. 2016). A third, more recent project is *Voci della Grande Guerra*, which started in 2016 with the goal to create a corpus of writings issued during World War I (letters, official documents, news articles, etc.) and a suite of CL tools to explore this wealth of extremely different – and often forgotten – voices. Finally, this special issue includes also an overview of the CLARIN initiative to create a European infrastructure of linguistic resources for humanities studies. The invited paper by Monachini and Frontini introduces the European consortium behind this initiative as well as the ongoing efforts aimed at creating the Italian network CLARIN-IT. This contribution represents an invitation for researchers in CL and the humanities to share resources, tools and expertise for the advancement of DH. We expect that the sorts of work we mention here are just scratching the surface of what is possible, and that progress in the coming years is likely and that it will come quickly. The current volume suggests more of what is possible!

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bradley, John. 2004. Text tools. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*. Blackwell Publishing, pages 505–525.
Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. A framework and graphical development environment for robust NLP tools and applications. In *ACL*, pages 168–175.
Eder, Maciej. 2015. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.
Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. Stylometry with R: A package for computational text analysis. *R Journal*, 16(1).
Elson, David, Anna Kazantseva, Rada Mihalcea, and Stan Szpakowicz, editors. 2012. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Montréal, Canada, June.
Feldman, Anna, Anna Kazantseva, and Stan Szpakowicz, editors. 2016. *Proceedings of the NAACL-HLT 2016 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, San Diego.
Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, Champaign.
Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.
Leinonen, Therese, Çağrı Çöltekin, and John Nerbonne. 2016. Using Gabmap. *Lingua*, 178:71–83.
Moretti, Giovanni, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support humanities studies. *Knowledge-Based Systems*, 111(C):100–112, November.
Nerbonne, John. 2003. Linguistic variation and computation. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 3–10. Association for Computational Linguistics.
Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap - A web application for dialectology. *Dialectologia: revista electrònica*, pages 65–89.
Nerbonne, John and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97)*, pages 11–18.
Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.
van Dalen-Oskam, Karina, Jesse de Does, Maarten Marx, Isaac Sijaranamual, Katrien Depuydt, Boukje Verheij, and Valentijn Geirnaert. 2014. Named entity recognition and resolution for literary studies. *Computational Linguistics in the Netherlands Journal*, 4:121–136.
Vanhoutte, Edward, Julianne Nyhan, and Melissa Terras, editors. 2013. *Defining digital humanities: A Reader*. Ashgate Publishing, Ltd., London.