DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 2

Digital Humanities and Natural Language Processing: “Je t’aime... Moi non plus”

Barbara McGillivray <bm517_at_cam_dot_ac_dot_uk>, The Alan Turing Institute and University of Cambridge
Thierry Poibeau <thierry_dot_poibeau_at_ens_dot_fr>, CNRS / Ecole normale supérieure — PSL/ Université Sorbonne nouvelle
Pablo Ruiz Fabo <ruizfabo_at_unistra_dot_fr>, Laboratoire LiLPa, Université de Strasbourg

Abstract

In spite of the increasingly large textual datasets humanities researchers are confronted with, and the need for automatic tools to extract information from them, we observe a lack of communication and diverging goals between the communities of Natural Language Processing (NLP) and Digital Humanities (DH). This contrasts with the wealth of potential opportunities that could arise from closer collaborations. We argue that more efforts are needed to make NLP tools work for DH datasets so that that NLP research applied to humanities data receives more attention, leading to the development of evaluation approaches tailored towards relevant research questions. This has the potential to bring methodological advances to NLP, while at the same time confronting DH datasets with powerful state-of-the-art techniques.

Introduction[1]

The recent years have witnessed an increased interest in Digital Humanities (DH) textual datasets within the Natural Language Processing (NLP) community, as several initiatives (such as the Computational Humanities group at Leipzig[2] and the Computational Humanities committee),[3] workshops (such as Computational Humanities 2014,[4] Teach4DH,[5] COMHUM 2018,[6] and the various editions of the LaTeCH-CLfL workshops),[7] and publications testify to [Biemann et al. 2014] [Nyahn and Flinn 2016] [Jenset and McGillivray 2017] [van der Zwaan et al. 2017] [Bamman 2017] [Schulz 2018] [Hinrichs et al. 2019]. Research in this area has focussed on developing new computational techniques to analyse and model humanities data. This interest stems from the methodological and technical challenges posed by these datasets, including those related to non-standard textual or multi-modal input, historical languages, multilinguality, and the need for advanced methods to improve the quality of digitization and for semi-automatic annotation tools. A number of successful results have been achieved, especially in the area of handwriting recognition [Christlein et al. 2018], computational stylistics [Boukhaled 2016], historical natural language processing pipelines [Piotrowski 2012], authorship attribution [Stamatatos 2009], and semantic change detection [Tang 2018], to name a few examples. 1

In spite of this growing activity, there is a real danger that NLP research on DH datasets does not take into account all the complexities of the phenomena and corpora, as others have pointed out [Dubossarsky et al. 2017] [Hellrich and Hahn 2016]. Moreover, NLP work on DH data has often been confined within the limits of the NLP community, which leads to serious methodological limitations for its applicability to DH and humanities research. In spite of the huge potential impact of NLP for DH datasets, NLP activities aimed at applying and adapting NLP research to the needs of the humanities are still marginal. This can be explained by the standard processes that the discipline adopts. Because the emphasis is on developing new computational systems or improving existing ones, it is very important that these are evaluated on standard datasets using reproducible methods. This means that there is an incentive for NLP researchers to work on very restricted sets of datasets and languages, leading to the development of tools which are optimized for those datasets and languages. This drives research towards a very specific direction, away from the idiosyncratic 2

features displayed by historical languages and DH data. Moreover, publication venues dedicated to NLP methods for DH are few and do not set the mainstream agenda of the field. Coupled with the challenges and the effort required to work on DH datasets, this means that engaging with this line of research appears to be a less than attractive option for most scholars.

3 On the other hand, a large part of humanities research involves analysing and interpreting written texts. Over the past few years large digital text collections have become available to the scholarly community, but where DH scholars confront Big Data to answer humanities questions, they often rely on methodologically un-sophisticated tools such Google Books Ngram Viewer [Greenfield 2013]. There is a real danger that these non-scientifically rigorous approaches will become state of the art [Pechenick et al. 2015].

4 In this article we aim to draw attention to the lack of communication between the communities of NLP and DH. In spite of its damaging effect on the progress of the disciplines, we believe this lack of communication and miscommunication are underestimated. We argue that what is needed is to bridge the gap between the highly technical state of the art in NLP and the needs of the DH community. We also offer a solution to this situation, inviting DH researchers to play a more active role in making NLP tools work for their data in order to give new insights into their questions, while at the same time advocating for a higher profile of NLP research applied to humanities data. Institutions also need to play a role in enabling better communication between the two communities, promoting interdisciplinary work and multi-author publications; publication venues welcoming such NLP/DH collaborative research also have an important role to play.

Contexts

5 An informal definition of the scope of DH was given by Fitzpatrick commenting on the DH 2010 conference, as “a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities [...] or who ask traditional kinds of humanities-oriented questions about computing technologies” [Fitzpatrick 2010]. Though informal, this broad characterization agrees with the variety of work described as DH in overviews of the field [Berry 2012, 1–20] [Schreibman et al. 2004].

6 More recently, some authors [Biemann et al. 2014, 87–91] have observed two types of research in the work described as DH in the overviews just cited. First, what Biemann et al. call DH “proper”, which in their characterization focuses on digital resource creation and access. Second, research which these authors call “Computational Humanities”, and which analyzes digital materials with advanced computational techniques, while trying to assess the value of those computational means for addressing humanities questions. They see work in what they term “Computational Humanities” as situated in a continuum between the humanities or the DH (according to their definition of the latter term) and Computer Science. Therefore, should we want to adopt the Digital vs. Computational Humanities terminology sometimes proposed, the work referred to here can be considered within the Computational Humanities. However, in the rest of this paper we will use the more widely adopted term of “DH”. In 2019, a heated debate emerged around the role of computational analysis in the study of literature specifically, after the publication of Nan Da's paper, which questions whether the computational analysis of literary texts can bring any additional insight in literary studies [Da 2019]. Counterarguments to such claims were offered by computational literary studies researchers [Critical Inquiry 2019] [Cultural Analytics 2020]. We do believe that computational methods can contribute to literary text analysis, and cite some examples related to character identification below. We agree, however, with Da's emphasis on the need to use computational tools optimally [Da 2019, 604]. As we argue below, this can involve considerable work in order to adapt to the specifics of a DH-relevant dataset, going beyond the tools' default configuration and requiring at times novel evaluation procedures.

7 Data relevant for social sciences and humanities research often take the shape of large masses of unstructured text, which is impossible to analyze manually. For example, regarding the use of textual evidence in political science, a variety of relevant text types have been identified, such as regulations issued by different organizations, international negotiation documents, and news reports [Grimmer and Stewart 2013]. Grimmer and Brandon conclude that “[t]he primary problem is volume: there are simply too many political texts”. In the case of literary studies, the complete text of thousands of works spanning a literary period [Clement et al. 2008] [Moretti 2005, 3–4] are beyond a scholar’s reading

capacity, and researchers turn to automated analyses that may facilitate the understanding of relevant aspects of those corpora.

Because DH researchers now face volumes of data that cannot be analyzed manually, NLP technologies need to be applied and adapted to specific use cases, integrating them in user interfaces to make the technology more easily usable by domain experts from the humanities and social sciences. Besides, a critical reflection on the computational tools and methods developed must be initiated, based on an evaluation by domain experts who are expected to benefit from those technological means. 8

We argue that researchers in the social sciences and humanities need ways to gain relevant access to large corpora. NLP can help provide an overview of a corpus, by automatically extracting actors, concepts, and the relations between them. However, NLP tools do not perform equally well on all texts and may require adaptation. Furthermore, the connection between these tools' outputs and research questions in a domain expert's field may not be immediately obvious and needs to be made explicit and kept in mind in the development of computational tools. Finally, evaluating the usefulness of an NLP-based tool for a domain expert is not trivial and ways to enable accurate and helpful evaluations need to be devised. 9

Different Datasets

Research in NLP aims to build tools and algorithms for the automatic processing of language [Jurafsky and Martin 2009]. In NLP, such systems are typically evaluated against baseline and existing systems with the aim to improve various measures of accuracy, coverage, and performance. Because the focus is on developing optimal algorithms, it is common practice to build and evaluate them based on existing, standard corpora. This way, it is possible to compare different approaches in a systematic way. 10

In the case of DH research, however, the focus is not so much on the algorithms as on the results that they lead to, which help the researchers answer their research questions. In this context, each study tends to focus on a specific and often unique dataset, with an emphasis on achieving satisfactory and insightful results using or adapting existing methods, if possible. 11

Several differences separate corpora typically used in NLP research from such DH datasets. First of all, size is typically very large in the former case, unless it is a particular aim of the research to optimize algorithms for small datasets. Moreover, with the exception of the cases in which the NLP research is focussed on a specific domain (such as medical, legal, etc.), balanced corpora are typically used. This ensures that the systems developed on such corpora can be generalized to the language as a whole, in line with generally accepted definitions of a corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" [Sinclair 2004]. In DH research, on the other hand, if answers to more general questions are sought, they rarely concern linguistic phenomena per se, and the aim of generalization is instead replaced by the need to describe and further explore the datasets at hand. Sometimes these take the form of archives, which, unlike corpora, are not collected with the aim to represent a language variety. Therefore, it may not be an option to modify the size of the corpus without changing the scope of the research, as the research question is bound to a specific set of texts. In the case of historical languages, for example, unlike current languages, it is often not possible to increase the size of the data because the amount of transmitted texts is limited by particular historical circumstances [McGillivray 2014, 15]. 12

DH studies normally do not involve balanced datasets according to criteria such as genre, style, register, etc. Another important difference concerns the quality and format of the texts, with particular reference to historical texts. Because almost every DH study requires its own dataset, a necessary preliminary step of the research process consists in acquiring the texts, if they are not already available. When Optical Character Recognition (OCR) is chosen, the texts will likely require a significant amount of processing before they reach an acceptable level of quality, otherwise there is a serious risk that standard NLP tools will fail to provide satisfactory results [Piotrowski 2012]. This is sometimes complemented with other challenges inherent to historical texts and which have to do with diachronic variation, including spelling variation and other types of language change. This means that a diachronic dimension needs to be included in 13

the text processing when applying NLP tools developed for current languages or based on synchronic corpora. We argue that DH corpora can usually be used without the privacy concerns of user-generated data. At the same time, the challenges that come with them offer a good test case not only to pursue meaningful DH questions, but also to measure the robustness of state-of-the-art algorithms.

14 Natively digital texts are not exempt from the caveats above. A case illustrating this is the work on the *Earth Negotiations Bulletin* (ENB) corpus [Ruiz et al. 2016] [Ruiz 2017]. ENB's volume 12 consists of reports summarizing participants' statements at international climate negotiation summits, where global climate agreements are handled. The goal of the study was to automatically identify where participants stand with respect to negotiation issues (support or opposition) as well as regarding other participants, thanks to syntactic and semantic role annotations obtained with an NLP toolkit. The corpus covers a period starting in 1995, and it contains what we might call "historical HTML varieties", as well as fixed column-width plain-text formats that required normalization before the content can be input to the NLP pipeline. In addition to this type of normalization, language use in the corpus has several non-standard traits, for which it was necessary to adapt the NLP toolchain (see section 4 below). In other respects, this corpus displays some of the typical features of DH datasets that we described above: it does not seek to represent a linguistic phenomenon, instead it is the set of texts required to answer domain-specific research questions. Venturini et al. argued that the corpus is a good choice to study climate negotiations, given that the corpus editors strive to cover participants in a balanced way, using a neutral and controlled language, although its specific linguistic characteristics need not generalize to political reporting texts [Venturini et al. 2014].

15 When the texts are already available to the DH community, they are often encoded according to the Text Encoding Initiative (TEI) markup, which has become a standard in DH. This markup focusses on editorial aspects and structural properties of the texts themselves, and less on their linguistic features. This contrasts with the state of the art in NLP, where corpora typically have shallow metadata information and are often rich in linguistic annotation [Jenset and McGillivray 2017, 124–125, 137–138]. Some linguistic corpora using TEI are however available, e.g. the National Corpus of Polish [Bański and Przepiórkowski 2009] [Przepiórkowski 2009]. Although traditions have diverged since then, at its inception the TEI was sponsored by both DH-related and NLP-related professional associations [Cummings 2007]. Also, major cultural works encoded in TEI have been annotated with morphosyntactic information, e.g. Dante's works [Dante Search 2018]. Therefore, extra processing and particular attention is required when employing NLP tools on such texts.

NLP Techniques in DH and Their Challenges

16 The DH research process can sometimes include an NLP pipeline, such as sentence segmentation, tokenization, lemmatization or stemming, or part-of-speech tagging. These steps are often required in order to conduct further analyses, either because they contribute to "cleaning" the texts, or because they help the researchers identify individual linguistic elements of interest such as word tokens, or classes such as parts-of-speech. This is often complemented by the use of corpus linguistics techniques, such as collocation analysis or various quantitative analyses of linguistic elements. In other words, the boundary between computational and corpus linguistics blurs in the interest of the DH research questions [Jenset and McGillivray 2017, 101].

17 Partially related to and building on the previous point, another way in which NLP methods are used in DH research concerns those techniques that extract various types of structured information from texts, including keywords, named entities, and relations. Again, these steps usually precede further analysis, in the form of qualitative or quantitative investigations, and can rest on other levels of linguistic processing, such as syntactic or semantic parsing.

18 Semantic processing can also support the identification and analysis of more abstract entities, such as semantic fields or concepts, and their linguistic realization in texts. For example, McGregor and McGillivray report on a methodology for extracting explicit occurrences of smell in historical medical records, the Medical Officer of Health Reports collected in London in the 19th and 20th century [McGregor and McGillivray 2018]. Given the large size of the dataset, automatic detection of instances of smell-related words by drawing on the distributional semantic properties of a small set of seed words enables medical historians to extract relevant passages in the texts, which can then be further analyzed, for

instance geographically or diachronically.

19 Some types of information are generally useful to understand a corpus. These include actors mentioned in it (e.g. people, organizations, characters), core concepts or notions of specific relevance for the corpus domain, as well as the relations between those actors and those concepts. This is a critical element in the analysis of the ENB, the climate diplomacy corpus mentioned in section 3 [Ruiz 2017]. As the corpus covers international negotiation reports, in order to understand its content it is essential to know not only which concepts were discussed in the negotiation, but also who discussed them and with which attitude (in a supporting or oppositional manner). Syntactic dependencies, semantic role annotations and coreference chains obtained with an NLP toolkit were exploited to that end. The agent of a reporting predicate was identified as a negotiation actor, and the message tied to that reporting predicate was considered to express negotiation issues addressed by the agent. The predicate itself (a reporting verb or noun) was seen as expressing the actor's attitude. The corpus has some non-standard linguistic features, and the NLP toolchain had to be adapted to handle them. For instance, personal pronouns "he", "she" can refer to countries in this corpus, so anaphora resolution had to be modified accordingly. Corpus sentences contain information about several participants at the same time (multiple actors mentioned within the agent role, or in adjunct roles), which also required a specific treatment in order to identify as many relations as possible of the type ⟨*actor*, *predicate*, *concept*⟩. A good coverage of such relations is crucial to analyses on negotiation behaviour relevant to domain experts.

20 A widespread approach to gain an overview of a corpus relies on network graphs called concept networks, social networks or socio-technical networks, depending on their content [Diesner 2012, 5, 84]. In such graphs, the nodes represent terms relevant in the corpus (actors and concepts), and the edges represent either a relation between the terms (like "support" or "opposition"), or a notion of proximity between them, based on overlap between their contexts. Creating networks requires a method to identify nodes, as well as a way to extract relations between nodes or to define node proximity, such as a clustering method. Networks have yielded very useful results for social sciences and humanities research. To cite an example, Baya-Laffite et al. and Venturini et al. created concept networks to describe key issues in 30 years of international climate negotiations described in the ENB corpus, providing new insights regarding the evolution of negotiation topics [Baya-Laffite et al. 2016] [Venturini et al. 2014]

21 Established techniques to extract networks from text exist, and networks offer useful corpus navigation possibilities. However, NLP can complement widespread methods for network creation. Sequence labeling and disambiguation techniques like Entity Linking can be exploited to identify the network's nodes: actors and concepts. The automatic definition of network edges is usually based on node co-occurrence, while more detailed information about the relation between actors and concepts is not usually automatically identified for defining edges. Nonetheless, such information can also be obtained via NLP methods. As for corpus navigation, networks do not in themselves provide access to the corpus fragments that were used as evidence to create the networks; but they can be complemented with search workflows that allow researchers to access the contexts for network nodes and the textual evidence for the relations between them.

22 The above-mentioned techniques are, in most cases, considered solved problems in NLP research, and are normally developed and tested on a set of large standard synchronic corpora of current languages. However, when applied to "messy", noisy, and/or historical texts, these tasks prove to be significantly more challenging. This points to the need of more research to bridge the gap between the DH and NLP communities, so that NLP tools can be developed with the requirements of the former in mind. In this context, Perrone et al. [Perrone et al. 2019] develop further a Bayesian learning model for semantic change detection developed by Frermann and Lapata [Frermann and Lapata 2016]. Frermann and Lapata's original system was built with the aim to model the change in meaning of English words in a general corpus covering the time period 1700-2010. Perrone et al. extend this to the case of Ancient Greek. This extension is not trivial, as the differences between the two corpora are substantial. The Ancient Greek corpus covers 13 centuries and its content is highly skewed, as certain centuries are only represented by one author or very few works, and some genres only appear in certain centuries. This lack of balance means that the models need to account for the differences in the texts' features and their effect on word semantics. For example, the sense of a word is highly dependent on the genre of the text it appears in. The Ancient Greek word *mus* can mean "muscle", "mussel", and "mouse", but in medical texts is it more likely to mean 'muscle', independently of the era of the text. As a result, Perrone

et al. modified the original system to incorporate genre information into the statistical model, thus achieving improved accuracy of the system compared to the English case.

23 Applying NLP for text analysis in social sciences and humanities poses some specific challenges. First of all, researchers in these domains work on texts displaying a large thematic and formal variety, whereas NLP tools have been trained on a small range of text types, e.g. newswire [Plank 2016]. Second, the experts' research questions are formulated using constructs relevant to their fields, whereas core tools in an NLP pipeline (e.g. part-of-speech tagging or syntactic parsing) provide information expressed in linguistic terms. Researchers in social sciences, for example, are not interested in automatic syntactic analyses per se, but insofar as they provide evidence relevant for their research questions: e.g. which actors interact with each other in this corpus, or which concepts does an actor mention, and which attitudes are shown towards those concepts? Adapting tools to deal with a large variety of corpora, and exploiting their outputs to make them relevant for the questions of experts in different fields is a challenge in itself.

Evaluation and Usability

24 In this section we first discuss the mismatch between the evaluation carried out in NLP and the needs of DH scholars in terms of tool evaluation and tool performance. Then, we present an example of a DH-relevant evaluation approach.

25 In the same way that exploiting NLP technologies to make them useful to experts in social sciences and humanities is challenging, evaluating the application of NLP tools to those fields also poses difficulties. A vast literature exists about evaluating NLP technologies using NLP-specific measures. However, these NLP measures do not directly answer questions about the usefulness for a domain expert of a tool that applies NLP technologies. Even less do they answer questions about potential biases induced by the technologies (e.g. focusing only on items with certain corpus frequencies), and how these biases affect potential conclusions to draw from the data [Rieder and Röhle 2012, 77] [Marciniak 2016]. As Meeks et al. state, research is needed with "as much of a focus on what the computational techniques obscure as reveal" [Meeks et al. 2012].

26 Let us take the example of a researcher interested in the analysis of characters in different novels. Named-entity recognition is an interesting application for the task, but existing tools are not always very accurate with fiction texts. Moreover, named entities are not enough: the system must probably be coupled with an anaphora recognition and resolution system, and with other modules able to recognize titles and occupations, for example, as some characters may just be mentioned by the title they have. Indeed, some available NLP-based systems for character detection have taken such difficulties into account [Coll Ardanuy and Sporleder 2015] [Bamman et al. 2014] [Vala et al. 2015]. Standard NLP evaluations do not provide enough information to evaluate if a tool will be accurate enough or not on a specific corpus, and evaluating this is a hard and, above all, time-consuming task for computational humanists. The character-detection papers just cited, accordingly, developed task-specific corpora.

27 To take another example, McGregor and McGillivray describe a computational semantics system for information retrieval from historical texts, the Medical Officer of Health reports from London for the years from 1848 to 1972 [McGregor and McGillivray 2018]. The ultimate aim of this research was to answer a specific question in medical history, namely the nature of the relationship between smell and disease in 19th-century London. Therefore, the computational system had to be evaluated in the context of the original research question and according to metrics that assessed its suitability for the specific task at hand, which was the retrieval of smell-related sentences in a specific large collection of historical health reports, rather than on standard NLP metrics and approaches.

28 As we said in section 2, evaluation procedures for NLP tools are typically focussed around the aim to improve the state of the art. In the case of DH research, the objective is connected to a set of research questions, which are typically not methodological or linguistic. A computational system may perform very well according to standard NLP evaluation measures, but it is not usable in DH if it does not help the researchers answer their questions. Moreover, the converse scenario has also been attested: a system that attains low scores in an NLP task may prove useful for a DH application. An example of the latter case is automatic keyphrase extraction. A standard evaluation task took place within the SemEval campaign [Kim et al. 2010]. The best systems reached an F1 measure below 0.3. These scores in themselves may seem unimpressive, the possible range being between 0 and 1. However, keyphrase extraction is routinely applied

in order to get an overview of a corpus in DH research [Moretti et al. 2016] [Rule et al. 2015], which suggests the usefulness of this technology for corpus-based research in the humanities. The way the NLP task was defined at SemEval (a keyphrase extracted by the candidate systems had to exactly match a keyphrase in the reference set annotated by human experts for it to count as correct) does not fully reflect the way the technology is useful for a corpus overview in DH tasks, where inexact matches can still be useful. In other words, a tool's performance in a standard NLP competition like SemEval and the tool's performance with a DH-relevant corpus need not be related.

29 In the same way that we encourage humanities researchers to accept automatic textual analyses as complementary to manual ones, we would also like to insist on the importance of understanding the limits of computational tools. Initiatives in this direction have emerged, like Tool Criticism [Traub and van Ossenbruggen 2015], which seeks to understand biases introduced by tools and digital methods on a task's results.

30 We will now review an example of evaluation relevant to a specific DH task [Ruiz 2017]. That study evaluates a corpus navigation application for the ENB corpus introduced above. The application relies on relation extraction, based on an NLP pipeline adapted to the corpus. Based on the pipeline's results, a user interface (UI) allows users to search separately for actors, their statements in climate negotiations, and their attitudes towards negotiation items or other actors (support or opposition). First of all, an NLP intrinsic evaluation for relation extraction was performed, comparing automatic results with human reference annotations. The difference here is that, additionally, a qualitative evaluation was carried out, based on interviews of over one hour with three domain-experts familiar with the corpus: two of them had previously published research on it, and one of them works at the organization that produces the corpus reports. The goal of the evaluation was threefold. First, to assess to what extent the corpus navigation application developed for the ENB corpus helps experts obtain an overview of its content (i.e. an overview of actors' behaviour in the negotiations). Second, whether the tool can help experts gain new insights on the corpus: facts unknown to them that emerge from using the tool, or new research ideas made apparent to them by using the tool. Finally, another goal was to see if the quality (in F1 terms) of the NLP-based system was sufficient for this specific application.[8] As an evaluation protocol, the experts were first shown the corpus exploration functions available to them on the UI, following the same steps with each expert. The experts were then asked to come up with questions about the corpus, use the interface to obtain information relevant to their questions, and comment on the results. As said above, it was assessed whether the UI helps them gain an overview as well as new insights into the corpus. The assessment was based on verbal evidence (expert comments) or other behavioural evidence (queries and other operations performed on the UI). The sessions were recorded and later transcribed non-verbatim: expert comments were summarized, and a description of operations performed by users on the UI was written up; the reports and session audios are publicly available. Rather than asking experts explicitly for feedback, their spontaneous comments were considered as possible evidence for or against the usefulness of the system in terms of corpus overview and insight gain.

31 The results that emerged from this qualitative evaluation highlight those aspects of the NLP-based system that are useful or pose problems for domain experts. Dynamic extraction of speakers as agents of reporting predicates, without relying on a predefined speaker list, brought to light statements by lesser-studied negotiation participants, that our experts had no data about, like NGOs and interest groups on climate and gender or climate and indigenous peoples. A perceived shortcoming of the system was its incomplete treatment of complex predicates like "express concern", where the informative part of the predicate regarding the speaker's attitude to a negotiation issue is the noun "concern" rather than the verb "express". Such results, obtained not from researcher expectations but from actual user feedback, can help shape improved versions of the system, and even identify generalizable areas of improvement for an NLP technology itself.[9]

32 The main point here is that a simple evaluation task as the one we just described can be informative in ways an intrinsic quantitative NLP evaluation would not be. However, note that tasks like the one just described involve several challenges, and the system just described could be improved. A first difficulty is that, as reported before [Khovanskaya et al. 2015], study participants often form an opinion about the intended contribution of a study, and they believe that it helps the research if the experiment provides evidence for their expected result. These beliefs can bias participants' behaviour. The attempt by Ruiz et al. to reduce this bias relied on avoiding to ask participants for explicit feedback, but

rather on recording their comments and behaviour, although it is debatable whether such bias decreases this way. A second difficulty is that collecting domain-expert feedback in the way described is time-consuming; even finding suitable experts may be difficult, hence the small number of participants in the study. Finally, as the NLP-based system was exploited via a UI, the task could be complemented by an evaluation based on Human Computer Interaction criteria, like usability or user satisfaction [Al-Maskari and Sanderson 2010] [Kelly 2007]. Such an evaluation would involve defining tasks to perform with the UI, employing a larger sample of domain experts. The proportion of tasks completed successfully could be measured, as well as task completion time; a questionnaire could measure user satisfaction. All in all, even the simple small-sample qualitative evaluation task above was informative about concrete aspects of an NLP-based system that helped or were an obstacle to answering specific questions relevant to a DH-research task.

Conclusion

We would like to end this paper with a summary of the status quo and some recommendations for the future. 33

As we have seen, many DH projects are based on large or even very large textual corpora. Sometimes it is not possible for the experts to know all the texts included in their corpus, and in some other cases the corpus is complex and a specific interface would help and quicken the analysis. In such contexts, it makes sense to use advanced NLP techniques in DH research. NLP tools are now mature and accurate enough to provide, in principle, reliable and extended analyses of various corpora in literature, history, and other text-focussed humanities fields. It is possible to annotate entities (people and locations, companies and institutions), semantic concepts, technical terms and domain-specific expressions. It is also possible to extract links between entities, and derive a network of relations from the information included in the texts in an unstructured form. It is even possible to represent relevant information through navigable maps, so that the end user can navigate the corpus and find relevant pieces of information scattered in different texts. 34

However, despite the accuracy and robustness of current NLP techniques, these are not yet widely used in DH. And, on the other hand, even if we see a growing interest for social sciences and cultural heritage in the NLP community, this is still quite marginal. The main issue is probably that the two communities are fundamentally driven by opposite goals. 35

The NLP community is interested in advancing NLP techniques, which means every published experiment must be evaluated and compared to previous work and show some improvement over it. The use of gold standards is thus of paramount importance to perform this comparison. The drawback is a large standardization of the research effort: many researchers explore broadly the same methodological avenues, with the same techniques applied on the same data. In fact, the NLP community tends to focus on one paradigm at a time and to produce a phase of homogeneous research questions and methods, sometimes at the expense of leaving out potentially interesting but less mainstream work. Today it is hard to publish research in an NLP venue that does not use word embeddings, neural networks and deep learning. Deep learning methods have had a huge impact on the field, leading to a clear (and sometimes dramatic) improvement in performances for almost all kinds of NLP tasks. However, we believe diversity of methods and critical approaches to their use can only be beneficial. 36

On the other hand, the DH community, generally speaking, is only secondarily interested in processing techniques. The goal of DH is to shed new light on humanities problems, taking into account the advantages of digital and computerized methods. Evaluating processing techniques is a challenge, since data are by definition always project-specific and there are no gold standards in the same sense intended in NLP. Tools are thus used as they are (“off-the-shelf”), sometimes after a brief evaluation and estimation of their quality and adequacy for a given task, sometimes without any evaluation. The complexity of current tools should also not be underestimated: most tools are difficult to use and, even when they are open source, modifying them is hard for most users and even most projects. Adapting tools to DH corpora is not trivial: most tools nowadays are based on machine learning techniques, which means large quantities of annotated data are necessary to be able to train a new model or, in other words, to adapt the system to the corpus under study. This means that even when the team is skilled enough to adapt a system, this adaptation is not always possible in practice. 37

All this explains the current situation and maybe some of the missed opportunities. At this point, one conclusion could be that the two communities are just different, they have different goals and, even if we can observe some points of 38

contact, these are marginal and not so interesting from a scientific point of view. But in fact we would like to defend the exact opposite view.

39 DH offers new, original and complex challenges to the experts, and these challenges require new, original and complex techniques to tackle them. DH research also offers real-world problems that are needed to prove that NLP techniques can be applied to different languages than English, to different corpora than newspapers and to different periods than just the 21st century. At the same time, NLP tools and methods are often underutilized and a more accurate choice of the techniques to use can have a strong impact on DH research. In fact, we argue that adapting and tuning NLP techniques to the corpus or domain under study is precisely where some of the most challenging, innovatively interesting and impactful research can be.

40 The dialogue between disciplines and the constructive and critical use of methods that we advocate for implies that most projects need a multidisciplinary approach. We acknowledge that more and more institutions support this interdisciplinary and multidisciplinary research and even create new job positions in this space, often spanning over several departments. However, more needs to be done to properly support interdisciplinary careers. An academic culture that favours single-author publications does not sufficiently support multidisciplinary (and multi-authored) research. Institutions should acknowledge this situation, which means promoting multi-author publications, and also publications related to different domains, ranging from computer science to traditional areas of the humanities.

41 There is a growing number of papers mixing DH and NLP presented at conferences such as the annual Digital Humanities conference or in the series of LaTeCH workshops (Language Technology for Cultural Heritage, Social Sciences, and Humanities). This indicates that there is a community, which sometimes calls itself Computational Humanities and includes the community of NLP for DH, but also research around techniques dedicated to image, video, sounds and music, etc. However, the insufficient number of high-profile publication venues where such research can be hosted penalizes scholars active in this space, as they are less likely to be considered “successful” by academic standards. One possible solution to this is that humanities departments recognize publications in computer science venues as equally valuable as publications in venues considered more traditionally humanistic.

42 In addition to these academic cultural changes, we believe that changes to research practices and publication standards can support the work at the interface between NLP and DH. We have stressed that off-the-shelf algorithms are often blindly applied, but it is hard to benchmark an algorithm’s performance on a DH corpus due to its unique features. We propose that reproducibility and algorithmic robustness checks are added to all NLP/DH publications, to strengthen the methodological foundations of the research results. For example, if a DH paper uses LDA topic modeling, a requirement for publication should be that the authors run the analysis using differing values of the parameter for the number of topics (k) and differing pre-processing steps.

43 After stressing the challenges (and some possible solutions to them), we would like to end this paper on a positive note, highlighting the interest of the research at the frontier between these domains and the opportunities it brings. The state of the art in both fields is advanced enough to contemplate combined approaches, research opportunities happen at a global scale, ethical considerations are a priority and the potential negative impact of the increasing use of artificial intelligence in propagating fake news, for example, is widely recognized. NLP for DH offers a unique opportunity to explore complex data, and to show how we can deal with complexity to get a better knowledge of our past.

44 By increasing the recognition and support of computational scholars within DH, NLP scholarship can become an attractive area of DH, thus creating space for scholars who might otherwise hesitate to go into DH due to poor job prospects. Just as the social sciences have successfully created space for these kinds of scholars, which has benefited the social sciences overall, DH could achieve similar results by adopting a similar strategy.

Acknowledgements

45 We would like to thank the anonymous reviewers for their thorough and helpful comments. Thierry Poibeau’s research is partially funded by the PRAIRIE 3IA Institute, part of the “Investissements d'avenir” program, reference ANR-19-P3IA-0001. This work has been mainly done while Thierry Poibeau benefited from a Rutherford fellowship at the Turing

Institute (London and University of Cambridge). This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

Notes

[1] “Je t'aime… moi non plus” (French for “I love you… me neither”) is a 1967 song written by Serge Gainsbourg for Brigitte Bardot. “The song was banned in several countries due to its overtly sexual content” (Wikipedia), but there is nothing sexual in this paper.

[2] https://ch.uni-leipzig.de/about/ (Last accessed on 20/05/2020).

[3] https://www.ehumanities.nl/computational-Humanities/ (Last accessed on 20/05/2020).

[4] https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=14301 (Last accessed on 20/05/2020).

[5] https://teach4dh.github.io (Last accessed on 20/05/2020).

[6] http://wp.unil.ch/llist/en/event/comhum2018/ (Last accessed on 20/05/2020).

[7] https://sighum.wordpress.com/events/latech-clfl-2018/ (Last accessed on 20/05/2020).

[8] F1 was 0.69 based on an exact match across system and reference results of triples containing a negotiating actor, its statement, and the reporting predicate (verbal or nominal) relating both.

[9] Indeed, at the time those evaluations took place, a current concern in computational linguistics [Bonial and Palmer 2016] was how to represent certain complex predicates (light verbs) in the lexical knowledge-base against which we automatically annotated the ENB corpus (PropBank [Palmer et al. 2005]).

Works Cited

Al-Maskari and Sanderson 2010 Al-Maskari, Azzah and Sanderson, Mark. “A review of factors influencing user satisfaction in information retrieval”. *Journal of the American Society for Information Science and Technology* 61. (2010): 859–868.

Bamman 2017 Bamman, David. “Natural Language Processing for the Long Tail”. Digital Humanities 2017 Conference Abstracts, pages 382-384, Montreal, Canada (2017). https://dh2017.adho.org/abstracts/408/408.pdf

Bamman et al. 2014 Bamman, David, Underwood, Ted, Smith, Noah A. “A Bayesian Mixed Effects Model of Literary Character”. *Proceedings of the Association for Computational Linguistics*, pages. 370–379. (2014)

Baya-Laffite et al. 2016 “Mapping Topics in International Climate Negotiations: A Computer-Assisted Semantic Network Approach”. In Kubitschko, S., Kaun, A. (eds.), *Innovative Methods in Media and Communication Research*. Springer International Publishing, Cham, pp. 273–291. https://doi.org/10.1007/978-3-319-40700-5_14 (2016)

Bański and Przepiórkowski 2009 Bański, Piotr and Przepiórkowski, Adam. “Stand-off TEI Annotation: The Case of the National Corpus of Polish”. Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP ’09 (2009): 64–67.

Berry 2012 *Understanding Digital Humanities*. Palgrave Macmillan. (2012)

Biemann et al. 2014 Biemann, C., Crane, G., Fellbaum, C., and Mehler, A. (eds) (2014). “Computational Humanities - bridging the gap between Computer Science and Digital Humanities”. Report from Dagstuhl Seminar 14301.

Bonial and Palmer 2016 Bonial, Claire and Palmer, Martha. “Comprehensive and Consistent PropBank Light Verb Annotation”. *Proceedings of LREC 2016, the 10th Language Resources and Evaluation Conference*. (2016): 3980–3985.

Boukhaled 2016 Boukhaled, Mohamed Amine. *On Computational Stylistics: mining Literary Texts for the Extraction of Characterizing Stylistic Patterns. Document and Text Processing*. Université Pierre et Marie Curie - Paris VI (2016).

Christlein et al. 2018 Christlein, V., Nicolaou, A., Schlauwitz, T., Späth, S., Herbers, K. & Maier, A. “Handwritten Text Recognition Error Rate Reduction in Historical Documents using Naive Transcribers”. In Burghardt, M. and Müller-Birn, C. (eds), *INF-DH-2018*. Bonn (2018).

Clement et al. 2008 Clement, Tanya, Sara Steger, John Unsworth, and Kirsten Uszkalo (2008). “How Not To Read A Million Books”. http://www.people.virginia.edu/~jmu2m/hownot2read.html.

Coll Ardanuy and Sporleder 2015 Coll Ardanuy, Mariona, Sporleder, Caroline. “Clustering of Novels Represented as Social Networks”. *LiLT (Linguistic Issues in Language Technology)* 12. (2015).

Critical Inquiry 2019 *Computational Literary Studies: A Critical Inquiry Online Forum*. https://critinq.wordpress.com/2019/03/31/computational-literary-studies-a-critical-inquiry-online-forum/.

Cultural Analytics 2020 “Debates” section of the *Journal of Cultural Analytics*. https://culturalanalytics.org/section/1580-debates.

Cummings 2007 Cummings, James. “The Text Encoding Initiative and the Study of Literature”. In Siemens, Raymond G., Schreibman, Susan (eds), *A Companion to Digital Literary Studies*. (2007): 451–476.

Da 2019 Da, Nan Z. (2019). “The Computational Case against Computational Literary Studies”. *Critical Inquiry*, 45(3), 601-639. https://doi.org/10.1086/702594.

Dante Search 2018 Dante Search Project. Dante Search. University of Pisa. http://www.perunaenciclopediadantescadigitale.eu/.

Diesner 2012 Diesner, J., 2012. *Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts*. Carnegie Mellon, Pittsburgh, PA.

Dubossarsky et al. 2017 Dubossarsky, H., Grossman, E., & Weinshall, D. “Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models”. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 1147–1156 (2017).

Fitzpatrick 2010 Fitzpatrick, Kathleen. “Reporting from the Digital Humanities 2010 Conference”. *The Chronicle of Higher Education*. https://web.archive.org/web/20190829004943/https://www.chronicle.com/blogs/profhacker/reporting-from-the-digital-humanities-2010-conference/25473 (2010).

Frermann and Lapata 2016 L. and Lapata, M. “A Bayesian Model of Diachronic Meaning Change”. *Transactions of the Association for Computational Linguistics*, 4 (2016).

Greenfield 2013 Greenfield, Patricia M. “The Changing Psychology of Culture From 1800 Through 2000”. *Psychological Science*, vol. 24.9: 1722–1731, doi:10.1177/0956797613479387 (2013).

Grimmer and Stewart 2013 Grimmer, Justin and Stewart, Brandon M., 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. *Political Analysis* 21, 267–297. doi:10.1093/pan/mps028 (2013).

Hellrich and Hahn 2016 Hellrich, Johannes and Udo Hahn. “Bad Company — Neighborhoods in Neural Embedding Spaces Considered Harmful”. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: Technical Papers, pages 2785–2796, Osaka, Japan, December 11-17 (2016).

Hinrichs et al. 2019 Hinrichs, E., Hinrichs, M., Kübler, S. et al. “Language technology for digital humanities: introduction to the special issue”. *Language Resources & Evaluation* 53, 559–563 (2019). https://doi.org/10.1007/s10579-019-09482-4.

Jenset and McGillivray 2017 Jenset, Gard and McGillivray, Barbara. *Quantitative Historical Linguistics. A corpus framework*. Oxford Studies in Diachronic and Historical Linguistics. Oxford University Press, Oxford (2017).

Jurafsky and Martin 2009 Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall (2009).

Kelly 2007 Kelly, Diane. “Methods for Evaluating Interactive Information Retrieval Systems with Users”. *Foundations and Trends® in Information Retrieval* 3. (2007): 1–224. https://doi.org/10.1561/1500000012.

Khovanskaya et al. 2015 Khovanskaya, Vera, Baumer, Eric, and Sengers, Phoebe. “Double binds and double blinds: evaluation tactics in critically oriented HCI”. *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*. (2015): 53–64.

Kim et al. 2010 Kim, Su Nam, Medelyan, Olena., Kan, Min-Yen and Baldwin, Timothy, 2010. “Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles”. *Proceedings of the 5th International Workshop on Semantic Evaluation*. (2010): 21–26.

Marciniak 2016 Marciniak, Daniel. “Computational text analysis: Thoughts on the contingencies of an evolving method”. *Big Data & Society* vol. 3, 1–5 . doi:/10.1177/2053951716670190 (2016).

McGillivray 2014 McGillivray, Barbara. *Methods in Latin Computational Linguistics*. Brill, 2014.

McGregor and McGillivray 2018 McGregor, Stephen and McGillivray, Barbara. “A distributional semantic methodology for detecting implied smell in historical medical records”. In Adrien Barbaresi, Hanno Biber, Friedrich Neubarth, and Rainer Osswald (eds), *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)* – September 19-21, 2018, pages 1–11, Vienna, Austria, 2018.

Meeks et al. 2012 Meeks, Elijah and Weingart, Scott B. “The Digital Humanities Contribution to Topic Modeling”. *Journal of Digital Humanities*, vol. 2:1 http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/.

Moretti 2005 Moretti, Franco. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

Moretti et al. 2016 Moretti, Giovanni, Sprugnoli, Rachele, Menini, Stefano, Tonelli, Sara, 2016. “ALCIDE: Extracting and visualising content from large document collections to support humanities studies”. *Knowledge-Based Systems*. (2016) 111: 100–112. https://doi.org/10.1016/j.knosys.2016.08.003.

Nyahn and Flinn 2016 Nyhan, J., Flinn, A. *Computation and the Humanities: Towards an Oral History of Digital Humanities*. Springer (2016).

Palmer et al. 2005 Palmer, Martha, Gildea, Daniel, Kingsbury, Paul. “The proposition bank: An annotated corpus of semantic roles”. *Computational linguistics*, 31 (2005): 71–106.

Pechenick et al. 2015 Pechenick, Eitan Adam, Danforth, Christopher M., Dodds, Peter Sheridan “Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution”. *PLOS ONE* 10(10): e0137041. https://doi.org/10.1371/journal.pone.0137041 (2015).

Perrone et al. 2019 Perrone, Valerio, Hengchen, Simon, Vatri, Alessandro, Palma, Marco, and McGillivray, Barbara. “GASC: Genre-Aware Semantic Change for Ancient Greek”. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, Florence, Italy, August 2, 2019. Association for Computational Linguistics.

Piotrowski 2012 Piotrowski, Michael. *Natural language processing for historical texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, (2012).

Plank 2016 Plank, Barbara. “What to do about non-standard (or non-canonical) language in NLP”. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)* – September 19-21, 2016, pages 13–20, Bochum, Germany, 2016.

Przepiórkowski 2009 Przepiórkowski, Adam. “TEI P5 as an XML Standard for Treebank Encoding”. *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT)* (2009): 149–160.

Rieder and Röhle 2012 Rieder, Bernhard and Röhle, Theo. “Digital Methods: Five Challenges”. In Berry, David (ed.) *Understanding Digital Humanities*, pages 67–84 (2012).

Ruiz 2017 Ruiz, Pablo, 2017. *Concept-based and Relation-based Corpus Navigation: Applications of Natural Language Processing in Digital Humanities* (PhD Thesis). École normale supérieure, PSL Research University, Paris.

Ruiz et al. 2016 Ruiz, Pablo, Plancq, Clément and Poibeau, Thierry. “More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing”. *Proceedings of LREC 2016, the 10th Language Resources and Conference* (2016): 192–197.

Rule et al. 2015 Rule, Alix, Cointet, Jean-Philippe, and Peter S. Bearman. “Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014”. *Proceedings of the National Academy of Sciences* (2015) 112: 10837–10844. https://doi.org/10.1073/pnas.1512221112.

Schreibman et al. 2004 Schreibman, Susan, Siemens, Ray G., Unsworth, John. (Eds.), *A companion to digital humanities. Blackwell companions to literature and culture*. Blackwell, Malden, MA. (2004).

Schulz 2018 Schulz, Sarah. “The Taming of the Shrew - Non-Standard Text Processing in the Digital Humanities”. Dissertation, University of Stuttgart (2018).

Sinclair 2004 Sinclair, John. “Corpus and text. basic principles”. Martin Wynne, editor, *Developing linguistic corpora: a guide to good practice*: 1–16. Oxbow books, Oxford (2004).

Stamatatos 2009 Stamatatos, E. “A survey of modern authorship attribution methods”. *Journal of the American Society for Information Science and Technology*, 60: 538-556. doi:10.1002/asi.21001 (2009).

Tang 2018 Tang, Xuri. “A State-of-the-Art of Semantic Change Computation”. *Natural Language Engineering* 24: 649-676 (2018).

Traub and van Ossenbruggen 2015 Traub, Myriam C. and van Ossenbruggen, Jacco (eds.) Workshop on Tool Criticism in the Digital Humanities. Centrum Wiskunde & Informatica, KNAW eHumanities, and Amsterdam Data Science Center. http://persistent-identifier.org/?identifier=urn:nbn:nl:ui:18-23500 (2015).

Vala et al. 2015 Vala, Hardik, Jurgens, David, Piper, Andew, Ruths, Derek. “Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts”. *Proceedings of Empirical Methods in Natural Language Processing*. (2015).

Venturini et al. 2014 Venturini, Tommaso, Baya Laffite, Nicolas, Cointet, Jean-Philippe, Gray, Ian, Zabban, Vinciane and De Pryck, Kari. “Three maps and three misunderstandings: A digital mapping of climate diplomacy”. *Big Data & Society* 1 (2014): 1–19 https://doi.org/10.1177/2053951714543804.

van der Zwaan et al. 2017 van der Zwaan, J. M., Smink, W. A. C., Sools, A. M., Westerhof, G. J., Veldkamp, B. P., & Wiegersma, S. “Flexible NLP Pipelines for Digital Humanities Research”. Paper presented at 4th Digital Humanities Benelux Conference 2017, Utrecht, Netherlands (2017).