DH Benelux Journal

Volume 1:
Integrating Digital Humanities

Autumn 2019

General Editors
Wout Dillen, Marijn Koolen, and Marieke van Erp

Guest Editors
Julie Birkholz and Gerben Zaagsma

digital humanities Benelux Journal

DH Benelux Journal 1. Integrating Digital Humanities.

ISSN: 2666-6952

This journal was typeset in LaTeX by Marijn Koolen using VariantX — a reusable template for journals in the Humanities, developed by Wout Dillen. VariantX is open source, available on GitHub, and deposited in the Zenodo Open Science Repository. DOI: 10.5281/zenodo.3484652.

Contents

Editors' Preface

As the DH Benelux conference enters its sixth edition, and its research is maturing, we wanted to offer participants the opportunity to publish more elaborate accounts of their work. Since the digital humanities community does not have many journals yet, we thought that a dedicated DH Benelux Journal could fill a gap here, and allow us to showcase some of the best work that is happening in our region.

Taking inspiration from our colleagues at Computational Linguistics in the Netherlands (CLIN) in the general setup of the journal, we invited authors of accepted conference abstracts to submit full papers to the journal, that were then subjected to a more stringent review process. For each new issue, our goal is to complete the submission, review, and publication processes before the next edition of the conference. Naturally, we will evaluate and tweak this cycle to best serve the interests of the DH Benelux community as our journal keeps maturing as well.

In line with the ideals of open science and open source software, we decided to make the journal available in Open Access – publishing all of the contributions in this issue under a Creative Commons "Attribution 4.0 International" license. Taking this one step further, we decided to only offer authors open formats for submitting the final versions of their articles: namely LaTeX and Markdown – formats that are increasingly gaining currency in the field. We are convinced that by sharing some of the formatting workload between authors and editors as such, we can make our journal's publication workflow more sustainable in the long run. Aware of the fact that this may introduce somewhat of a learning curve for our authors, we are committed to lowering the threshold as much as possible, by offering detailed formatting instructions on the journal's website, and helping authors where necessary.[1] We thank Folgert Karsdorp, Lars Wieneke and Joris van Zundert for preparing the publishing pipelines based on these formats, and for helping us set up and style the journal site.

We hope you enjoy this first issue, and we look forward to continuing this exciting new chapter for digital humanities research in the Benelux.

September 9, 2019
Amsterdam

Wout Dillen
Marijn Koolen
Marieke van Erp

[1] For our introductions to formatting final submissions for the journal in LaTeX or Markdown, see: http://journal.dhbenelux.org/submission/preparing-the-final-version-of-your-manuscript/.

Introduction: Integrating Digital Humanities

Julie Birkholz[1] and Gerben Zaagsma[2]

[1]Ghent University, GhentCDH
[2]Université du Luxembourg, C²DH

Much ink has been spent, and occasionally spilled, trying to define the Digital Humanities and its place among the academic disciplines. Yet whether it is seen as a field of its own, a sub- or inter-discipline, or a set of practices, most proponents agree on some basic characteristics, with interdisciplinarity probably topping the list. As early as two decades ago, Willard McCarty was among the first to assert that DH constituted an *interdiscipline*, due to its "common ground of method [which] makes it possible to teach applied computing to a class of humanists from widely varying disciplines" (McCarty, 1999). At the same time, DH challenges existing and ingrained research practices (perhaps sometimes more imagined than real), according to which humanities research questions must always derive from domain knowledge, by proposing new data- and method-driven approaches to research in the humanities.

In practice, Digital Humanities projects typically involve, and bring together, a variety of practitioners from different backgrounds: academics from various fields and disciplines, librarians, archivists and museum experts. All of this could easily be construed as providing evidence of the existence of some sort of shared field; yet the influence of the digital on the various phases of our research practice (whether information gathering, processing, analysis and dissemination) comes in many forms: sometimes it is obvious, sometimes it is tacit and implicit, and sometimes aspirational. It is, however, precisely this observed (potential for) intersection that can also cut both ways: "[the danger is] that digital humanities may [...] become ghettoised rather than further integrated into scholarship" (Warwick et al., 2007). That might sound almost absurd in an age when many countries and regions, especially in Europe and North America, hold annual DH conferences, with the most recent international DH conference of the Alliance of Digital Humanities Organizations in Utrecht (the Netherlands) attracting over 1,000 participants.[1] Thus we would like to argue here that, both intellectually and practically, integration is not only Digital Humanities' most defining feature but also its most pressing imperative.

Yet for those working in the field it might be all too easy to forget that much work remains to be done to truly integrate digital approaches into the humanities, in both teaching and research. Integration, then, implies bringing together and encouraging productive collaboration between humanities and computer science researchers, as

[1] http://adho.org/

well as heritage professionals. It also means encouraging people to acquire expertise beyond their own professional field, and recognising that the answer to the question of how much expertise and cross-disciplinary knowledge is necessary depends, among other things, on the project(s) at hand, the profile of the participants involved, and the distribution of tasks among them. Integration also means expanding one's own methodological repertoire and established ways of argumentation. It means integrating new practices and/or materials in research and teaching. Finally, integration means consciously working towards a situation where digital and humanities go hand-in-hand, instead of one being promoted at the expense of the other. A blind emphasis on digital methods that loses sight of what contribution these methods actually make to humanistic knowledge misses the point, and inhibits their uptake. Conversely, promoting a humanities in which 'digital' is seen as tainting its seemingly unique character, ignoring the latter's methodological value and the fundamental ways in which our engagement with the human record is changing, is similarly harmful.

Today, it is as common to find a misplaced ignorance of the digital among some 'traditional' humanists ("misplaced," since the digital affects every human and humanist) as it is to find a misplaced, condescending attitude and/or naive ignorance of the humanities among some digital humanists. In order for DH to be(come) integrated as a field, it also means putting aside preconceptions and assumptions and recognising the fields represented by those working in DH. Some digital humanists might be tempted to despair of 'pesky Luddites' who refuse to see the digital light, but it is high time for DH practitioners to frame their work in terms of its broader contribution to and integration into humanities and heritage work, whether that contribution is about domain knowledge, method development, software and code work, or data and tool development. Only by deliberately emphasising both the digital and the humanities can we hope to achieve this.

One may, of course, ask if there is an imperative to do so. The answer to that question is two-fold and relates both to our current state of affairs and to future possibilities. To begin with, the humanities are already touched by 'the digital' in manifold ways. All phases of the humanities research process are somehow impacted by the digital; yet how, and to what extent, that is the case, needs to be questioned. Furthermore, DH is often equated with data and tools, with scale and technology; 'big data' especially seems to define digital humanities in the eyes of many humanists, to the detriment of paying attention to the changes taking place in the research practices of humanities scholars in general. But technology equals methodology and thus directly influences the way in which we as scholars conduct our research. Scholars and DH practitioners need to be able to make informed choices as to the affordances and pitfalls of implementing digital approaches, tools and methods. If separating digital from humanities is already rather nonsensical in light of the above, a look at the future of the human record provides ample proof of the need to consciously engage with the digital (Brügger and Milligan, 2018). A prime example here is the shift from paper to web archives and the fundamental changes this will bring, and is already bringing, to conducting research into human history and culture.

With this in mind, we, as the scientific chairs and guest editors of this inaugural issue, decided upon the theme of "Integrating Digital Humanities" for the DH Benelux 2018 Conference, and we asked our peers and colleagues to reflect, in a critical and self-reflexive way, on how the digital turn affects knowledge production and dissemination in the humanities and heritage sectors. This inaugural issue of the Digital Humanities Benelux Journal features a selection of papers that were presented at the

fifth annual DH Benelux Conference[2], held at the International Institute for Social History in Amsterdam, The Netherlands. Submissions ranged from history, linguistics, literature and cultural heritage to spatial humanities, digital born data, media and DH infrastructure to reflections and debate on DH, resulting in 59 short papers, 36 long papers, 9 round tables, 9 demos and 20 posters. The conference, like this first issue of the DH Benelux Journal, seeks to be a reflection of the diverse and wide community of DHers and DH research not only in and of the Benelux – Belgium, Netherlands and Luxembourg – but also beyond these borders.

The four articles selected for this issue highlight different aspects of the broader question of integration in the digital humanities. The question of integration in the context of the encounter between different disciplines in DH is addressed by Max Kemman in his essay on "boundary practices of digital humanities collaborations". Based upon his recent study of trading zones in digital history, and the observation that collaboration across disciplines is inherent to the digital humanities, Kemman questions how this plays out on the intersection of the humanities and computer science. With the aim to "provide empirical grounding for discussions of digital humanities as a meeting between the computational domains and the humanities", Kemman analyzed an online survey, which was answered by 173 scholars, and found that there is often little disciplinary diversity of digital humanities collaborations, with humanities scholars dominating the collaboration, while there is often a large physical distance between the collaborating partners.

Concluding that "digital humanities collaborations are biased towards the humanities, rather than a balancing of the digital and the humanities", Kemman proposes that this is due to the fact that humanities scholars, not computer scientists, are setting the agenda for research collaborations. As a result, it seems that many scholars retain their disciplinary 'home' culture, instead of going interdisciplinary and entering a distinct 'third, in-between space' of digital humanities. Kemman's work provides those working in DH not only with a comprehensive analysis of how collaboration in DH works, but also with a set of propositions for advancing distinct cross-disciplinary practices.

The second article by Krista Murchison and Ben Companjen highlights the necessary integration of data preparation and curation practices in DH research in general, and the collaboration of librarians and researchers in a specific project in particular. In their article "Manuscripts, Metadata, and Medieval Multilingualism: Using a Manuscript Dataset to Analyze Language Use and Distribution in Medieval England", Murchison and Companjen focus on identifying multilingualism in medieval society through text. Their essay entails the first large-scale quantitative analysis of the distribution of French texts in medieval England. In providing a framework for quantitative manuscript-based analysis, the authors reflect on the methodology and digital approach rather than the project's speci?c sociolinguistic ?ndings, though some analytical results are discussed. Analyzing 958 French manuscripts from medieval England, Murschison and Companjen detail a semi-automatic approach to cataloguing the manuscripts: manually categorising manuscripts and combining this manuscript description data with a machine-actionable, reusable, and interoperable format to calculate the distribution of languages in each manuscript. Their work provides evidence for the persistence of French among both lay and clerical audiences and challenges the master narrative of the 'triumph' of English, while highlighting medieval England's multifaceted intercultural exchanges. This paper should be seen as a gold standard for

[2] http://2018.dhbenelux.org

future DH papers and the quest for integration – combining great clarity in writing, excellent documentation of the approach ensuring valid and reliable reuse, and clear and explicitly stated contributions to the domain knowledge and methods.

The third paper, by Roel Smeets, Eric Sanders and Antal van den Bosch, on "Ranking Characters in Present-Day Dutch Literary Fiction" seeks to integrate data and domain knowledge driven approaches as well as qualitative and quantitative analysis. Combining network analysis with narratology, the authors assessed the demographic metadata of 2,137 characters from a corpus of 170 contemporary Dutch novels, extracting the social networks of characters from each novel and ranking the characters' relations on five centrality metrics. Next, they assessed if there is a relationship between demographic variables and a character's position in the generated network. This resulted in the finding that immigrant and female characters score higher on a number of measures, suggesting that this approach to character centrality, compared to traditional narrative approaches, enhances our understanding of the relations between characters in novels. Smeets, Sanders and van den Bosch's work builds on a trend to automate character relations in text, as well as on the use of network measures to explain narratives in new ways. Their work also contributes to Digital Literary Studies by integrating data-driven approaches to networks into analyses of literary texts.

Finally, in the fourth article in this issue, Sergio Peignier and Patricia Zapata seek to integrate data mining and semio-pragmatic discourse analysis into the traditionally small-data-based study of political rhetoric. In their "Analysis of Fidel Castro Speeches Enhanced by Data Mining", they propose a data mining technique for examining speeches and show how a hybrid discourse analysis methodology provides a more comprehensive way of understanding possible discursive strategies, compared to previous work that has largely been limited by corpus size. In conclusion, they argue that the framework presented in their paper could be integrated as a valuable complementary analysis tool into the rapidly growing and highly relevant field of study of populist rhetoric.

The papers presented in this first issue of the DH Benelux Journal reflect and highlight various aspects of the state of integration of the Digital Humanities, by adapting and developing tools and approaches around specific domain-centred as well as data-driven research questions, and by developing and reflecting upon more specific (digital) pipelines that reflect specific methods. Moreover, the last three papers are excellent examples of the collaboration practices that characterize the field, as documented by Kemman in the first paper in this issue. These essays and the ongoing research that will be documented in future issues of the DHBenelux Journal speak to currently evolving practices of integration in the Digital Humanities. They confirm that the work of integration is indeed already being done on a daily basis, and has been done so for years. It is high time to consciously engage with the question of integration, not only as practice between collaborators, but as a vision and program for how the Digital Humanities (can) complement our understanding of the human condition.

Guest editors of DH Benelux Journal

Julie M. Birkholz, Ghent Centre for Digital Humanities, Department of Literary Studies, Ghent University, Belgium
Gerben Zaagsma, Centre for Contemporary and Digital History (C^2DH), University of Luxembourg

References

Brügger, N. and I. Milligan
2018. *The SAGE Handbook of Web History*. SAGE Publications Limited.

McCarty, W.
1999. Humanities computing as interdiscipline. http://www.iath.virginia.edu/hcs/mccarty.html. [Accessed: 4 Sept. 2019].

Warwick, C., M. Terras, P. Huntington, and N. Pappa
2007. If you build it will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. *Literary and linguistic computing*, 23(1):85–102.

Boundary Practices of Digital Humanities Collaborations

Max Kemman[1,2]

[1]University of Luxembourg, Centre for Contemporary and Digital History
[2]Dialogic

One of the defining characteristics of digital humanities is its emphasis on interdisciplinary collaboration. In order to coordinate across disciplinary boundaries, the development of common ground is necessary to negotiate goals and practices. Yet how such common ground can be established, and whether the adoption of interdisciplinary practices and vocabularies results in participants drifting apart from their disciplinary cultures is underexplored. In this paper I investigate the boundary practices of digital humanities, referring to the interactions of scholars with cross-disciplinary collaborators and disciplinary peers, and how these are affected by disciplinary diversity and physical distance within collaborations. With an online questionnaire, which received 173 responses, I have found that there is often little disciplinary diversity in digital humanities collaborations, with participants and leadership coming mostly from the humanities. The physical distance is often great, and communication increasingly relies on email. I have not found that these dimensions affect the respondents' frequency of communication with collaborators or peers. My conclusion is that physical distance and disciplinary diversity cannot be confirmed to affect the frequency of boundary interactions of digital humanities. I furthermore conclude that digital humanities collaborations are biased towards the humanities rather than a balancing of the digital and the humanities. This paper thus provides an empirical grounding for discussions of digital humanities as a meeting place between the computational domains and the humanities.

Keywords: collaboration, common ground, communities of practice, survey, boundary practices, interdisciplinarity

1 Introduction

One of the defining characteristics of digital humanities is the emphasis on interdisciplinary collaboration (Klein, 2014, Spiro, 2012). The different facets of digital humanities research, such as computer technology, data management and humanistic

inquiry, call for collaborations among experts from different backgrounds. While it is possible for individuals to develop interdisciplinary expertise, collaborations between disciplinary experts may be more effective at providing access to the full range of expertise and practices of disciplines (Siemens et al., 2011a, Stokols, 2006, Wilson, 1996). As a result, it has been argued that diversity of backgrounds in collaborations is a prerequisite for the generation of new knowledge in the digital humanities (Edmond, 2016). Yet how collaborations are conducted in the digital humanities is regularly hidden from view and therefore poorly understood (Griffin and Hayler, 2018).

It is most commonly assumed that digital humanities collaborations consist of a digital and a humanities side, a collaboration between computational experts and humanities scholars (Edmond, 2005). Digital humanities then is a 'meeting place' of these two sides (Svensson, 2011). McCarty (2012) has argued that this meeting should constitute a 'level ground' of mutual collaboration, truly working together rather than computational experts working in the service of humanities scholars. In order to achieve this, it has been argued that collaborators should develop a common ground of shared practices and vocabularies to coordinate goals and practices within the collaboration (Siemens, 2009, Siemens et al., 2012). As a result, scholars may learn to translate their humanistic research questions into computational terms (McCarty, 2012). To illustrate, through their continued interactions with computational experts, historians might start to think of their research as testing hypotheses against a historical dataset with the use of algorithms. This will make it easier for those historians to discuss their research with computational experts, but might make it more difficult to discuss it with other historians who are not part of a digital humanities collaboration.

Yet how such common ground can be established between different disciplinary discourses, and whether this leads to a drifting apart from one's original disciplinary background remain open questions. Therefore, it is necessary to study the practices and consequences of collaborations that develop common ground to better understand the digital humanities as a collaborative practice both in itself and within the humanities at large.

A common approach to this question, on a *micro* scale, is to interview or observe practices of collaboration as they are performed and experienced by individuals or small groups. Yet whether these individual cases are typical or atypical of the digital humanities more generally requires contextualisation on the *meso* scale, by quantitatively collecting dimensions of collaborations in a multitude of groups or institutions (for a discussion of the micro and meso scales, see Edwards, 2002). Focusing on the meso scale, this paper is thus concerned with a quantitative exploration of the composition of collaborations and the question of how participants interact with cross-disciplinary collaborators and disciplinary peers. This exploration is based on the findings of the online questionnaire described in Section 2.

This paper builds upon the theory of *communities of practice*, defined as the binding together of individuals through a sharing of practices, as experienced in mutual engagement, a joint enterprise and a shared repertoire of resources (Wenger, 1998, p. 73). In this paper I focus on mutual engagement, the meeting between collaborators, as configured through two underlying dimensions that I will elaborate below: the shared history of learning and the geography of practice (Wenger, 1998). Using these two dimensions I explore how they lead to different boundary practices between communities of practice, i.e. the interactions between members of different communities. For the purposes of this paper, the relevant communities are the humanities, the computational domains, and collaborations as communities consisting of participants

from both the humanistic and computational communities.

The next section elaborates on the two above-mentioned dimensions, followed by a presentation of the questionnaire and a discussion of the results. The paper concludes with lessons learned from the questionnaire and an exploration of research questions that will further our understanding of how digital humanities collaborations work in practice.

1.1 Dimensions of communities of practice

The establishment of communities of practice can be described as contingent on two underlying dimensions: the shared history of learning and the geography of practice.

The first dimension, *shared history of learning*, refers to the joint adoption of practices and vocabularies through participation and reification of meanings. As a shared history of learning how to conduct humanities scholarship, and reification through educating students to be future scholars, the humanities can be thought of as a community of practice. The same can be said about the computational domains. These communities of practice demarcate what type of problems are of interest, what types of approaches are suitable and how research should be communicated, thereby defining the boundaries of disciplinary communities (Becher and Trowler, 2001, Gieryn, 1983). The collaboration between the computational domains and the humanities can therefore be seen as a meeting of different histories of learning. The meeting of shared histories of learning is therefore considered to be the *disciplinary diversity* of a collaboration and is measured by the extent to which collaborations consist of both humanities scholars and computational experts.

The development of common ground can be described as the creation of a new shared history of learning. Over time, a common ground might develop into a shared history of learning of a new community of practice that is the union of the digital and the humanities. In this context, several authors have suggested that the digital humanities can indeed be thought of as a community of practice (Siemens and Burr, 2013, Siemens, 2016).

The second dimension, *geography of practice*, refers to the physical locations of collaborators and how that configures interactions. The geography of practice stands in a bidirectional relationship with the first dimension, as it influences the likelihood of shaping a shared history of learning. The most significant aspect is the *physical distance* between collaborators. First, distance affects communication. When collaborators are closer together, communication has higher quality, is more frequent and more often face-to-face (Kiesler and Cummings, 2002, Kraut and Egido, 1988). Second, distance affects the mutual awareness of collaborators, following the 'out of sight, out of mind' adage (Olson et al., 2002). Finally, distance affects the formation of group identity, leading to collaborators speaking in terms of 'us' and 'them' (Armstrong and Cole, 2002). While disciplinary diversity by itself necessitates coordination to align collaborators, these disciplinary differences, unlike physical distance, have not been found to increase *problems* of coordination (Cummings and Kiesler, 2005, Walsh and Maloney, 2007). Problems of coordination have been found with respect to differences in language and cultural habits, access to technology and conflicting requirements from funders (Siemens and Burr, 2013).

This is not to say that physical distance is only a negative aspect, nor does physical proximity guarantee a better collaboration. Extending physical distance in a collaboration means that one can work with the most fitting collaborators rather than being dependent on who is available nearby (Siemens and Burr, 2013). Previous research

found that for collaborations within a university, an increase in the number of collaborators correlated with an increase in negative collaborative experiences. Yet this correlation was not found for collaborations between universities (Tsai et al., 2016). Moreover, physical distance in collaborations can be a strategy for the dissemination of knowledge beyond one's own local network (Poole, 2013). Furthermore, 'virtual teams' that collaborate mainly through communication technologies such as email or teleconferencing have proven successful, although the formation of mutual trust remains an issue (Purvanova, 2014). Face-to-face communication was found to be more strongly related to team performance than virtual communication (Marlow et al., 2018). Yet Marlow et al. point to the advantage of what they call 'hybrid teams', where complex problems are coordinated face-to-face, while clearer tasks may be coordinated via communication technologies such as email. Establishing trust and addressing ill-defined problems, which are common in the digital humanities, therefore benefit from face-to-face meetings throughout a collaboration (Siemens and Burr, 2013).

Physical distance is thus strongly interrelated with communication, a necessary aspect of negotiating common ground. This brings me to the next aspect of my study.

1.2 Boundary practices

As a meeting place of different disciplinary communities, digital humanities collaborations can be characterised as *boundary practices*. Within a collaboration, the development of common ground requires participants to look beyond their own disciplinary identity and engage with collaborators from different backgrounds (Siemens et al., 2011b). Participants must be willing to learn about the practices and vocabularies employed by collaborators and integrate them into their own practices for mutual coordination. As such, interdisciplinary collaborations can be described as practices of crossing one's own disciplinary boundaries. Collaborations therefore potentially constitute what I term *interdisciplinary boundary crossing*.

The question remains how this (re-)configures the relationship with a participant's disciplinary community. It has been argued that participants drift away from their disciplinary culture and into a new shared history of learning following the adoption of new vocabularies and practices (Wenger, 1998, p. 103). If scholars wish to discuss their research with disciplinary peers who are not part of the collaboration, they now find themselves confronted with a boundary of different practices and vocabularies that they did not experience before. As such, collaborations potentially constitute what I term *intradisciplinary boundary construction*. For a simplified overview of the boundary practices between the communities of practice, see Figure 1.

The interactions between the communities of practice that form part of a digital humanities collaboration can therefore be characterised as a duality of interdisciplinary boundary crossing and intradisciplinary boundary construction. In the rest of the paper, these practices are referred to simply as boundary crossing and boundary construction. This paper explores the boundary practices, disciplinary diversity and physical distance of digital humanities collaborations. Specifically, the research question is as follows: *How do disciplinary diversity and physical distance affect boundary practices of digital humanities collaborations?*

As increased disciplinary diversity is expected to lead to more opportunities for cross-disciplinary interactions within a collaboration, the hypothesis underlying disciplinary diversity is that little disciplinary diversity, e.g. a collaboration consisting only of historians, will result in less boundary crossing and less boundary construction. In contrast, large disciplinary diversity, i.e. a collaboration consisting of an equal number

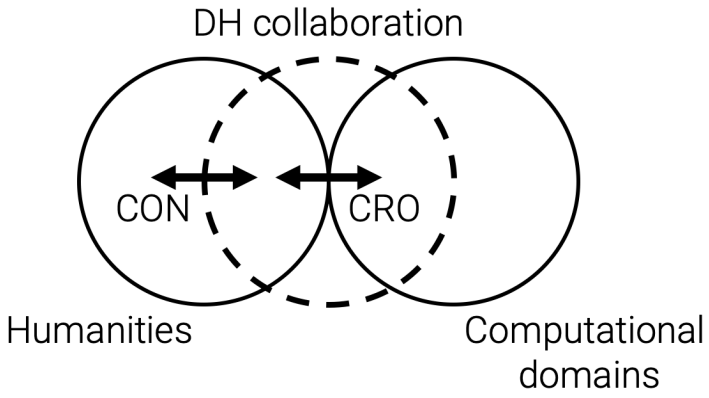

Figure 1: Model of a digital humanities collaboration, including intradisciplinary boundary construction (CON) and interdisciplinary boundary crossing (CRO).

of humanities scholars and computational experts, will result in increased boundary crossing as well as increased boundary construction. Furthermore, increased physical distance is expected to make interactions and coordination between cross-disciplinary collaborators more difficult. In cases of a large physical distance between cross-disciplinary collaborators, it is moreover to be expected that participants will have a small physical distance to their disciplinary peers, since in these cases participants are likely to be situated in their disciplinary departments or institutes. Similarly, a small physical distance to collaborators might mean a large physical distance to disciplinary peers. The hypothesis underlying physical distance is therefore that a small physical distance, e.g. humanities scholars sharing an office with computational experts, facilitates boundary crossing and boundary construction. In contrast, a large physical distance is expected to reduce both boundary crossing and boundary construction.

In order to test these hypotheses, it is necessary to acquire systematic and comparable figures related to these dimensions of digital humanities collaborations. In the next section I therefore introduce an online questionnaire on collaborative practices in the digital humanities.

2 Method

2.1 Online questionnaire

The questionnaire was distributed between November 2017 and April 2018 via social media and email. The distribution on social media included tweets with hashtags of conferences that took place during the period that the questionnaire was open for input. The questionnaire was furthermore distributed through blog posts and mailing lists. Finally, it was emailed to 128 collaborations based on their affiliation to CenterNet[1] or my awareness of them. Invitations for participation were written in English, Dutch, French, German, Spanish, Portuguese and Italian; the questionnaire itself was in English.

[1] CenterNet is an international network of digital humanities centres (Walter, 2012). I scraped the CenterNet list on 20 November 2017. For a version of the list archived in November 2017 see https://web.archive.org/web/20171029201211/http://dhcenternet.org/centers

The questionnaire was hosted on a Qualtrics account of the University of Luxembourg. Respondents were not asked for personally identifiable information in order to preserve anonymity, although the questionnaire did include a question about the name or title of the collaboration for which the participant was filling out the questionnaire. The main reason this question was included is that many scholars active in digital humanities do not participate in just one collaboration. In trial runs I noticed that participants became confused about which collaboration they were describing in the questionnaire. For example, one participant that worked at a digital humanities centre on a project switched back and forth between answers related to the centre and answers related to the project. By asking participants to give the name of the collaboration, be it the centre or the project, this name was included in questions so that they were reminded throughout the questionnaire on which collaboration they were providing information. The name of the collaboration was not used for further analysis.

It is therefore possible that several respondents described the same collaboration. Some statistics that are related to collaborations specifically, such a physical distance and disciplinary diversity, may therefore contain duplicates. However, the inclusion of duplicates need not imply that results are skewed. Furthermore, the boundary practices reported by respondents are individual. It is possible that within the same collaborations, different participants experienced different boundary practices. Insofar as this paper aims to investigate whether specific organisations of collaborations lead to specific boundary practices, the inclusion of duplicates therefore does not cause problems for analysis.

The questionnaire did not provide a definition of digital humanities or digital history, since both terms are contested in the literature (see e.g. Antonijević, 2015, Robertson, 2016, Terras et al., 2013). Likewise, it is not easy to define what constitutes a collaboration or who is part of a collaboration (Katz and Martin, 1997). Whether a certain interaction is considered a collaboration varies between disciplines and institutes (Burroughs, 2017). For example, CenterNet lists all affiliates as 'centres', yet contains a variety of terms such as 'lab', 'centre', 'initiative', 'team', 'department', 'institute' and 'group', without a clear delineation between any of these concepts. Rather than defining types of collaborations, collaborations are defined to exist in mutual recognition as collaborators (Wenger, 1998, p. 56). The questionnaire therefore did not provide a definition of the types of collaboration, nor who should be thought of as collaborators. The questionnaire was thereby designed as a bottom-up approach to investigate the boundary practices of collaborations, rather than a top-down approach of defining types of collaborations and describing the boundary practices for each.

2.2 Main units of analysis

The questionnaire can be consulted in Appendix A. For a structured overview of variables central to the discussion of the research question, see Table 1. For all variables, respondents could choose a single answer. The range of options and outcomes are described in the next section. In this section I describe how these variables relate to the model of digital humanities collaborations as shown in Figure 1.

As described above, *physical distance* concerns the distance between collaborators within the collaboration. In Figure 1, this refers to anyone falling within the 'DH collaboration' circle. *Disciplinary diversity* refers to the number of participants from the 'Humanities" circle and participants from the 'Computational domains' circle that fall within the 'DH collaboration' circle.

Table 1: Main variables for discussing the research question. The numbers in the column 'Question' refer to the numbered questions in Appendix A.

Variable	Question	Type	Response rate (# / %)
Physical distance	9	Ordinal	168 / 97
Disciplinary diversity	5	Ordinal (Likert)	173 / 100
Intradisciplinary interactions	13	Ordinal	170 / 98
Cross-disciplinary interactions	12	Ordinal	168 / 97
Main means of communication	11	Categorical	170 / 98

With respect to boundary practices, *intradisciplinary interactions* is represented by the arrow marked 'CON' in Figure 1, to measure the interactions between participants of the collaboration and their disciplinary peers outside the collaboration. *Cross-disciplinary interactions* is represented by the arrow marked 'CRO' to measure the interactions within the collaboration circle between participants of different disciplinary circles. Finally, *main means of communication* refers to the means of communication between collaborators within the 'DH collaboration' circle.

None of the variables were interval, as shown in Table 1, or normally distributed, see Appendix B. Therefore, for all statistical tests non-parametric tests will be used.

3 Results

The questionnaire received 173 responses.[2] The replies were analysed using SPSS. Since none of the questions were mandatory, some questions received fewer than 173 replies. As this constitutes a relatively small sample size, the analysis below reports both the frequencies of answers as well as percentages. Statistical tests are considered significant when $p<0.05$, following common practices of statistical analysis, and are reported in footnotes accompanying interpretations of results.

The questionnaire collected responses from all continents except Africa. Whether this reflects a lack of African digital humanities collaborations or simply a lack of responses is unclear. To geographically contextualise responses, the questionnaire inquired where collaborators were located, allowing multiple answers. 121 collaborations (70%) included European partners. European countries with over ten responses were France (22), Germany (24), Italy (19), Luxembourg (15), the Netherlands (28), Switzerland (10) and the UK (27). 47 collaborations included North American partners (27%), including Canada (14) and the USA (40). Finally, 32 collaborations included partners from the rest of the world (18%), with no country reporting more than eight responses.

To further contextualise the type of collaborations, the questionnaire included questions about the time-frame (single choice) and the source of funding (multiple choices). With respect to the time-frame, respondents could choose between short-term (deadline within four years), semi-long term (longer than four years or without a concrete deadline) or long-term (no deadline). Although four years is already relatively long, especially for computational projects, this captures projects tied to PhD positions and most fixed-term contracts. The majority of collaborations were short-term, see Table 2. With respect to funding, there was a fairly equal split between university funding and

[2] The data and SPSS code are available open access via Kemman, M. (2019) Boundary practices of digital humanities collaborations. figshare. Dataset. https://doi.org/10.6084/m9.figshare.7813571

Table 2: Descriptive aspects of collaborations, reporting both raw frequencies of answers (#) and the percentage related to all 173 responses (%).

Time-frame	#	%	Funding	#	%	Reason	#	%	Success	#	%
Short-term	95	55	University	77	45	Topic	45	26	Yes	89	52
Semi-long	33	19	National	84	49	PI	33	19	No	5	3
Long-term	44	25	European	23	13	Professional	19	11	Too early	31	18
			Government	22	13	Collaboration	14	8	Partially	12	7
			Infrastructure	9	5	Methods	13	8	Can't say	4	2
			Other	43	25	Novelty	6	3			
						Other	10	6			

funding from a national funding agency.

The questionnaire furthermore inquired about the contractual responsibilities of collaborators, where respondents could choose multiple answers. For the majority of collaborations, participants worked on multiple research projects (129; 75%). Only 17 respondents (10%) indicated that participants were contractually tied to one collaboration only.

Finally, the questionnaire included open-ended questions about the respondents' individual reason(s) to join the collaboration, and whether they considered the collaboration a success. Both questions were manually coded for analysis by grouping similar answers under a single category. The most frequently cited reason for joining a collaboration was interest in its topic, see Table 2. Two other types of responses are particularly interesting. The second most common response was that the respondent was the PI or founder of a collaboration. This is hardly an explanation for why a respondent chose to participate in a collaboration but suggests that there is a different incentive between PIs and others who join the collaboration later on. It appears that PIs do not join a collaboration but establish it, and consequently have no other reasons for joining. Another interesting response is that several respondents indicated that they joined as part of their professional responsibilities, either because the collaboration was tied to a position they had applied for, demands of available funding, or was part of their job duties. This suggests that participation in digital humanities collaborations may in some cases follow top-down decisions rather than intrinsic interests.

With respect to the success of a collaboration, the majority of respondents considered their collaboration a success, see Table 2. Only five respondents answered negatively. I cannot assume or conclude that this is representative of the success of digital humanities collaborations in general. It could be a result of self-selection bias for the question, as several respondents did not answer the question. Furthermore, this finding could show self-selection bias for the questionnaire, with scholars in successful collaborations more likely to respond to a questionnaire about their collaboration. The questionnaire did request no further explanation why a collaboration was deemed successful or not.

Having contextualised the responses as such, the following sections focus on the boundary practices central to this paper, and how they are configured by disciplinary diversity and physical distance.

Table 3: Disciplinary diversity. Frequencies (#) and percentages of all 173 responses (%).

	Leadership		**Participants**	
	#	%	#	%
History	88	51	102	59
Other humanities	82	47	109	63
Cultural heritage	22	13	50	29
Humanities total	**157**	**91**	**150**	**87**
Computer science	23	13	83	48
Computational linguistics	18	10	43	25
Software development	13	8	75	43
Computational total	**40**	**23**	**127**	**73**
Library	21	12	51	31
Other	35	20	54	31

3.1 Disciplinary diversity

With respect to the first dimension, disciplinary diversity, I consider three types of disciplinary backgrounds. First, *humanities backgrounds*, consisting of options 'history', 'other humanities', or 'cultural heritage'.[3] Second, *computational backgrounds*, consisting of the options 'computer science', 'computational linguistics' and 'software development'. Finally, other options were 'library' and 'other'. Three questions inquired about disciplinary diversity.

The first two questions concerned the disciplinary background of the leadership of the collaboration and of other participants. Respondents could choose multiple answers for both questions. For an overview of responses, see Table 3. With respect to the disciplinary diversity of leadership, *collaborations were mostly led by humanities scholars*: 91% included humanities scholars as leaders, compared to 23% that included computational experts. Furthermore, the majority of collaborations were led exclusively by scholars with a humanities background (126; 73%). A much smaller number of collaborations were led exclusively by people with a computational background (9; 5%). For 31 collaborations (18%), leadership involved scholars from both the humanities and a computational background.

With respect to the other participants of collaborations, disciplinary diversity appears more balanced. *Most collaborations included both participants with a humanities background and participants with a computational background*: 87% of collaborations included humanities scholars, and 73% included computational experts.

The third question concerned the ratio of humanities scholars to computational experts on a 5-point Likert scale. *For the vast majority of collaborations, humanities scholars outnumbered computational experts* (for an overview see Figure 2). Collaborations with only humanities scholars or mostly humanities scholars together comprise three-quarters of all responses. No collaboration included computational experts only.

In sum, *there was little disciplinary diversity in digital humanities collaborations*: leadership consisted mostly of humanities scholars, and humanities scholars also outnumbered collaborators with a computational background as participants.

[3] The division between history and other humanities is a result of the questionnaire being conducted as part of my PhD research on digital history specifically (Kemman, 2019).

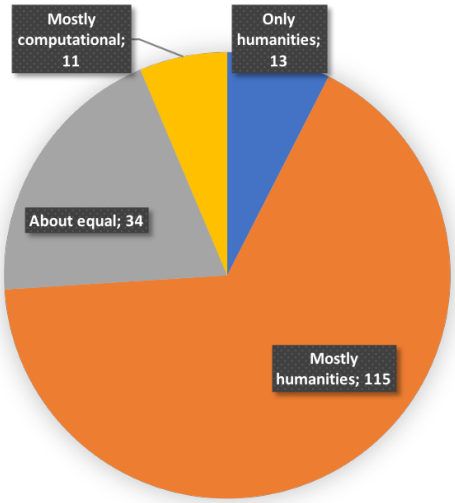

Figure 2: Ratio of disciplinary backgrounds.

3.2 Physical distance

With respect to the second dimension, physical distance, three questions inquired about where collaborators conducted their work and how they communicated with one another.

First, respondents were asked where the main participants of the collaboration worked, allowing a single answer. From my own observations, collaborations are often officially led by professors who have their own offices, but mainly conducted by researchers in PhD or postdoc positions who might or might not be sharing an office together. It is the interactions of these main participants that are of particular interest for the development of common ground. The frequencies of responses to this question can be seen in Figure 3. A total of 34 collaborations (20%) were conducted at a very short distance in a single space, either a lab or an office. 42 collaborations (24%) were conducted across a wider distance, but still within the same institute. Finally, 92 collaborations (53%) were conducted between multiple institutes in national or international contexts. Thus, *the majority of collaborations were conducted at a great physical distance*.

A second question concerned the institutional buildings where these spaces were located, allowing multiple answers. The majority of collaborations were located in the humanities building of an institute (112; 65%). Far fewer collaborations had spaces in the computer science building (29; 17%) or the library building (32; 18%). This corresponds to the earlier finding that the majority of collaborations consisted mainly of scholars from the humanities.

The final question inquired about the main means of communication within the collaboration, allowing a single response. In Figure 4 it can be seen that physical distance affected communication within a collaboration:[4] *as physical distance increased,*

[4] The category 'offices on multiple floors in a single building' received only six responses, and all six respondents communicated face-to-face. This distribution of answers is different from the other categories, which I assume is a result of the small sample size rather than a meaningful consequence of this category of physical distance. This category is therefore excluded from subsequent analyses.

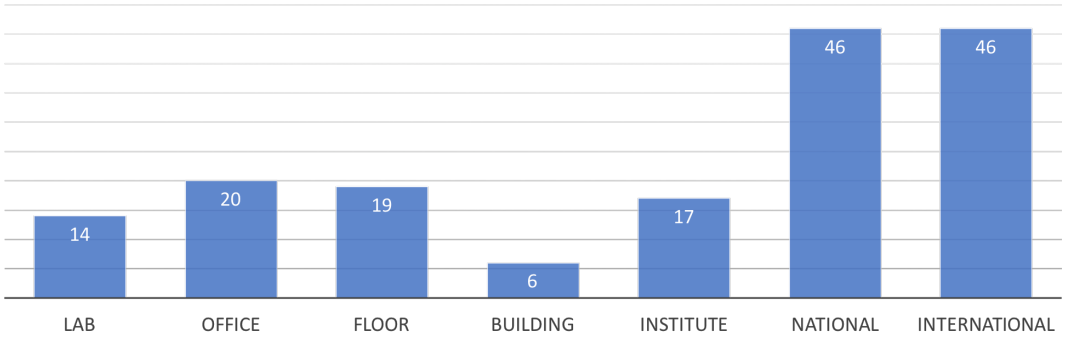

Figure 3: Physical distance between main collaborators, showing the frequencies of responses.

the use of face-to-face communication decreased, and the use of email increased,[5] which is in agreement with the literature. While both are within a single space, communication within a lab appears differently from that within a single office, indicating different styles of collaboration in different spaces. Moreover, when the collaboration was spread out over multiple buildings in a single institute, the use of email increased to a level similar to that for multiple institutes or international collaborations. This seems to indicate that inter-departmental collaborations experience similar obstacles to face-to-face communication as inter-institutional collaborations.

In sum, *the physical distance between participants in digital humanities collaborations is often great*: the majority of collaborations were conducted between different institutes and increasingly depended on email rather than face-to-face communication.

3.3 Boundary practices

As described above, two forms of boundary practices are central to the current study: interdisciplinary boundary crossing and intradisciplinary boundary construction. As a proxy for these boundary practices, the questionnaire inquired about the frequency of research-related communication with both cross-disciplinary collaborators and disciplinary peers. Figure 5 shows the frequency of responses to both questions in comparison. From this figure it can be seen that both disciplinary and interdisciplinary communication were quite frequent. Two-thirds of respondents spoke at least weekly with interdisciplinary collaborators. Three-quarters of respondents spoke at least weekly with disciplinary peers. Interaction with disciplinary peers outside the collaboration was significantly more frequent than interdisciplinary communication within

[5] I tested the relation between physical distance and communication via email or face-to-face with Kendall's tau correlation, a non-parametric test for ordinal data with small sample sizes (Field, 2009, pp. 181-182). Considering both variables are ordinal, tau-c was used to control for so-called 'tied ranks' where multiple respondents chose the same answers. Only cases that mainly communicated via email or face-to-face were selected, and the six responses in the category 'offices on multiple floors in a single building' were left out (see footnote 4). For the remaining 129 cases larger physical distance was found to be significantly correlated to more email instead of face-to-face communication τ-c=0.445, p(two-tailed)<0.001, N=129.

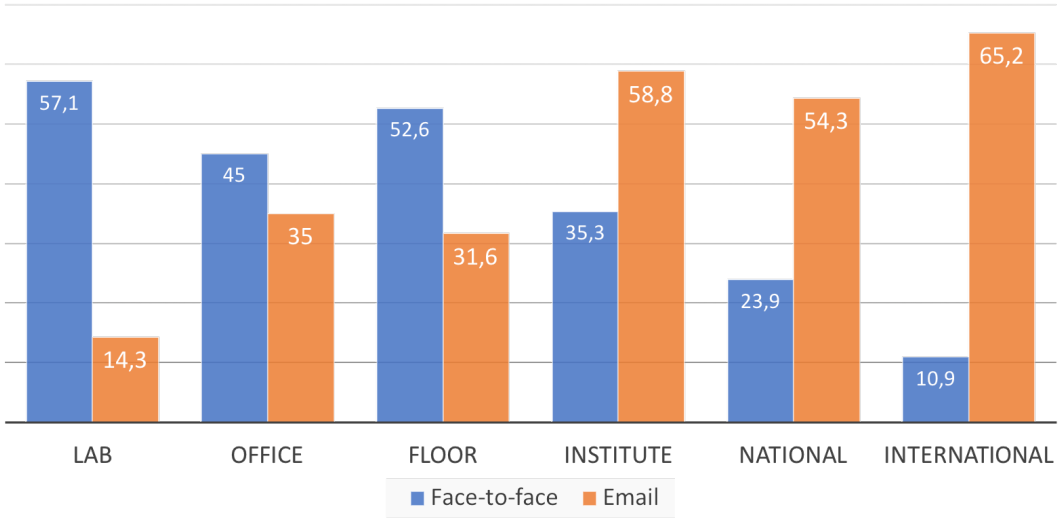

Figure 4: Main means of communication within the collaboration per distance, showing the percentage of responses.

the collaboration.[6]

Tests on the relation between disciplinary diversity or physical distance and boundary practices all returned non-significant results.[7] In other words, *I have found no signs that disciplinary diversity or physical distance affected the frequency of intradisciplinary or interdisciplinary boundary interactions*.

4 Discussion

In sum, I have found, with the aid of the questionnaire, that there was often little disciplinary diversity in digital humanities collaborations: most participants came from the humanities and most collaborations were led by humanities scholars. The majority of collaborations were conducted across a great physical distance, which correlated with an increased dependence on distant communication via email rather than face-to-face.

I hypothesised that small disciplinary diversity and great physical distance would lead to both a decreased opportunity for interdisciplinary boundary crossing and decreased intradisciplinary boundary construction. However, the analyses did not confirm that disciplinary diversity or physical distance affected the frequency of boundary interactions. The hypothesis therefore cannot be confirmed. Yet a number of findings give further insights into the boundary practices that may affect the development of

[6] I tested the difference with a Wilcoxon signed-rank test, a non-parametric test of differences between two sets of answers originating from the same respondents (Field, 2009, pp. 552-558). Respondents communicated significantly less with cross-disciplinary collaborators with z=-2.301, p(two-tailed)<0.05, N=168, r=-0.13.

[7] I tested these relations with Fisher's exact test, a test to compare the relationship between non-interval variables with small sample sizes (Field, 2009, p. 690). This concerns four independent statistical tests, none of which showed a statistically significant relation: 1) disciplinary diversity related to interdisciplinary communication, p(two-tailed)=0.08, 2) disciplinary diversity related to disciplinary communication, p(two-tailed)=0.52 3) physical distance related to interdisciplinary communication, p(two-tailed)=0.12 and 4) physical distance related to disciplinary communication, p(two-tailed)=0.25.

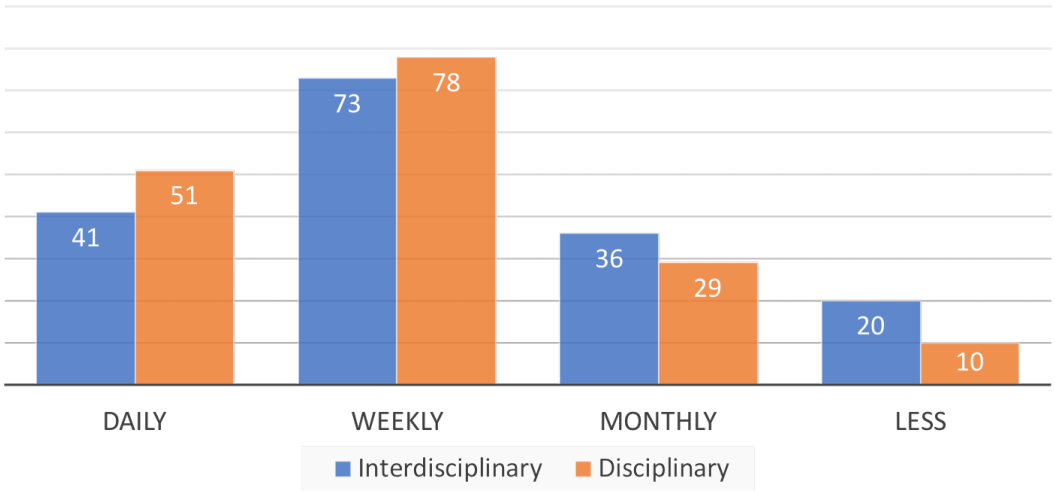

Figure 5: Frequencies of interactions with cross-disciplinary collaborators or disciplinary peers, showing the frequency of responses.

common ground in digital humanities collaborations.

First, while physical distance did not affect the *frequency* of interactions, it did affect the *nature* of communication in the collaborations, as increased physical distance was related to an increased reliance on email. However, previous research found digital communication insufficient for the development of common ground (Siemens, 2009). Especially insofar as collaborations need to establish mutual trust and coordinate ill-defined goals, it has been questioned whether this can be facilitated through distant communication (Sonderegger, 2009). In collaborations, goals emerge through continuous negotiations instead of being established prior to the collaboration (Haythornthwaite et al., 2006). Therefore, collocation, such as sharing an office, has been found to facilitate the development of common ground most effectively (Olson et al., 2002).

Second, for the majority of collaborations participants worked on multiple projects. This suggests that even if a common ground is established in a collaboration, these shared practices and vocabularies are limited to the collaboration. For other collaborations, it is to be expected that participants need to negotiate other common grounds. The opportunity for a common ground to develop into a shared history of learning thus appears limited.

Third, the majority of collaborations were embedded in the humanities buildings of institutes. This suggests smaller physical distance to disciplinary peers for participants from the humanities.

Finally, respondents communicated significantly more with disciplinary peers outside the collaboration than with cross-disciplinary collaborators. These last two findings suggest that humanities scholars remained aligned with their humanities background rather than forming a new alignment with computational collaborators. It thus appears unlikely that the humanities scholars involved in the digital humanities drift apart from their disciplinary community. Considering the finding that collaborations were furthermore predominantly conducted and led by humanities scholars, my findings agree with Svensson (2011) who characterisesd 'the digital humanities as

a humanities project'. Digital humanities thus appears less diverse than is generally assumed (cf. Edmond, 2016).

5 Conclusions and future outlook

In conclusion, let me return to the research question underlying this paper: *how do disciplinary diversity and physical distance affect boundary practices of digital humanities collaborations?* Answering this question is not a straightforward matter. Disciplinary diversity and physical distance do not affect the frequency of boundary interactions. I therefore have not found effects of these dimensions on boundary crossing or boundary construction. Yet physical distance does affect the nature of communication, with increased physical distance related to increased reliance on email. The questionnaire moreover found that the majority of collaborations were conducted at a great distance with predominant participation from humanities scholars. I therefore conclude that digital humanities collaborations are biased towards the humanities rather than balancing the digital and the humanities. The main contribution of this paper then is to provide an empirical grounding for discussions of digital humanities as a meeting place between the computational domains and the humanities. The results of this paper furthermore facilitate the contextualisation of future case studies of digital humanities collaborations, making it possible to position their organisation as typical or atypical compared to the results of the questionnaire.

The approach described in this paper thereby provides a quantitative meso perspective on collaborations, yet has a number of limitations that affect interpretability. A first limitation is that, as a result of my distribution methods, most respondents came from the humanities. This could have skewed findings insofar as humanities scholars outnumbered other disciplinary backgrounds. For example, collaborations between computational experts and cultural heritage, which may well be considered digital humanities practices, are probably underrepresented. Second, by focusing on disciplinary diversity, physical distance and boundary practices, the questionnaire sought to investigate collaborations through a quantitative approach. However, this does not provide in-depth insights into the development of common ground, the establishment of boundary practices or the nature of practices in the digital humanities. For example, since the majority of respondents indicated that participants worked in multiple collaborations, a much more detailed look at how individuals move between collaborations and perform boundary practices is necessary. Therefore, following the results of this questionnaire, a number of questions for future research require further exploration.

First, considering the apparent dominance of humanities scholars in digital humanities collaborations, both among participants and leaders, a question is how this affects possible power relations in interdisciplinary collaborations. Whereas McCarty (2012) has argued for a level ground where computational experts work with rather than in the service of humanities scholars, it is possible that humanities scholars effectively set the agenda for digital humanities collaborations. The question of how such power relations affect the development of common ground and the coordination of practices requires deeper observations of the interactions between humanities scholars and computational experts. For example, one question might be whether commonly negotiated vocabularies are closer to the disciplinary discourses of the humanities or to that of computational domains.

Second, considering the dependence on communication technology such as email,

one question is how this affects the development of common ground. Since communication technology has been found lacking in previous research, long-distance collaborations may need to plan for face-to-face meetings at partners' locations in order to develop common ground (Siemens and Burr, 2013, Sonderegger, 2009). Yet whether such meetings at intervals are sufficient is underexplored. In the case of communication technology, email, rather than synchronous distant communication, has been found to be the best alternative to face-to-face. The reason is that email allows collaborators to contemplate and elaborate on what they intend to communicate, and it also provides a stable backlog that can facilitate common ground (Sonderegger, 2009). More research is therefore needed on how communication technologies are used to develop common ground in the digital humanities.

Finally, the relationship between the digital humanities and the humanities at large deserves more attention. One question is whether scholars ultimately remain part of their disciplinary culture or whether the digital humanities indeed constitute a distinct community of practice (Siemens and Burr, 2013, Siemens, 2016). It has been suggested that digital humanities constitute a "dual citizenship" (Svensson, 2012) or a third culture (Hunter, 2014) between the digital and the humanities. Yet the results from the questionnaire suggest otherwise. Scholars remained in close contact with disciplinary peers. Most scholars were part of multiple collaborations, making it likely that scholars developed and negotiated different practices in different settings. This raises the question of whether vocabularies and practices shared within a single collaboration are able to extend beyond that collaboration into a wider community of practice of digital humanities.

On the meso scale adopted in this paper, the digital humanities therefore appear to maintain the duality of the humanities and the computational domains. Yet the results suggest that the digital humanities may be more heavily oriented towards the humanities than a balancing of the digital and the humanities. While on a micro scale scholars may act in-between the two domains, possessing both humanistic and computational skills, these cases seem atypical in the wider context of digital humanities collaborations. This is not to deny the possibility of the digital humanities emerging as a third space, but it poses questions about how scholars can develop into members of such a third space, how exactly this third space relates to the humanities and the computational domains and what kind of boundary practices are consequently introduced.

Acknowledgements

I would like to thank a number of people for their help and feedback that helped improve this paper: Anita Lucchesi, Benjamin Zenner, Lucas Duane, Stef Scagliola and Vitus Sproten for translating the invitations to fill out the questionnaire; Andreas Heinz for providing feedback on the statistical tests; Christopher Morse for proof-reading the article; and the editors and anonymous reviewers for providing feedback on an earlier version of this paper. The questionnaire reported in this paper is part of my PhD thesis on trading zones of digital history, supervised by Andreas Fickers, Benoît Majerus and Pelle Snickars. My thanks also to them for their feedback on an earlier version of this paper.

References

Antonijević, S.
2015. *Amongst Digital Humanists: An Ethnographic Study of Digital Knowledge Production*, pre-print edition. Basingstoke New York, NY: Palgrave Macmillan.

Armstrong, D. J. and P. Cole
2002. Managing Distances and Differences in Geographically Distributed Work Groups. In *Distributed Work*, P. Hinds and S. Kiesler, eds., Pp. 167–186. MIT Press.

Becher, T. and P. R. Trowler
2001. *Academic Tribes and Territories: Intellectual Enquire and the Culture of Disciplines*, 2nd edition. The Society for Research into Higher Education & Open University Press.

Burroughs, J. M.
2017. No Uniform Culture: Patterns of Collaborative Research in the Humanities. *portal: Libraries and the Academy*, 17(3):507–527.

Cummings, J. N. and S. Kiesler
2005. Collaborative Research Across Disciplinary and Organizational Boundaries. *Social Studies of Science*, 35(5):703–722.

Edmond, J.
2005. The Role of the Professional Intermediary in Expanding the Humanities Computing Base. *Literary and Linguistic Computing*, 20(3):367–380.

Edmond, J.
2016. Collaboration and infrastructure. In *A New Companion to Digital Humanities*, S. Schreibman, R. Siemens, and J. Unsworth, eds. Wiley-Blackwell.

Edwards, P. N.
2002. Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems. In *Technology and Modernity: The Empirical Turn*, P. Brey, A. Rip, and A. Feenberg, eds. MIT Press.

Field, A.
2009. *Discovering Statistics Using SPSS*, 2nd edition. Sage Publications Ltd.

Gieryn, T. F.
1983. Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, 48(6):781–795.

Griffin, G. and M. S. Hayler
2018. Collaboration in Digital Humanities Research – Persisting Silences. *Digital Humanities Quarterly*, 12(1).

Haythornthwaite, C., K. J. Lunsford, G. C. Bowker, and B. C. Bruce
2006. Challenges for Research and Practice in Distributed, Interdisciplinary Collaboration. In *New Infrastructures for Knowledge Production: Understanding e-Science*, C. Hine, ed., Pp. 143–166. IGI Global.

Hunter, A.
2014. Digital humanities as third culture. *MedieKultur: Journal of Media and Communication Research*, 30(57):18–33.

Katz, J. and B. R. Martin
1997. What is research collaboration? *Research Policy*, 26(1):1–18.

Kemman, M.
2019. *Trading Zones of Digital History*. PhD thesis, Université du Luxembourg, Esch-Belval.

Kiesler, S. and J. N. Cummings
2002. What Do We Know about Proximity and Distance in Work Groups? A Legacy of Research. In *Distributed Work*, P. Hinds and S. Kiesler, eds., Pp. 57–82. MIT Press.

Klein, J. T.
2014. *Interdisciplining Digital Humanities: Boundary Work in an Emerging Field*, online edition. University of Michigan Press.

Kraut, R. and C. Egido
1988. Patterns of contact and communication in Scientific Research Collaboration. In *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*, Pp. 1–12. ACM.

Marlow, S. L., C. N. Lacerenza, J. Paoletti, C. S. Burke, and E. Salas
2018. Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. *Organizational Behavior and Human Decision Processes*, 144:145–170.

McCarty, W.
2012. Collaborative Research in the Digital Humanities. In *Collaborative Research in the Digital Humanities*, M. Deegan and W. McCarty, eds., Pp. 1–10. Ashgate.

Olson, J. S., S. Teasly, L. Covi, and G. Olson
2002. The (Currently) Unique Advantages of Collocated Work. In *Distributed Work*, P. Hinds and S. Kiesler, eds., Pp. 113–136. MIT Press.

Poole, A. H.
2013. Now is the Future Now ? The Urgency of Digital Curation in the Digital Humanities. *Digital Humanities Quarterly*, 7(2).

Purvanova, R. K.
2014. Face-to-face versus virtual teams: What have we really learned? *The Psychologist-Manager Journal*, 17(1):2–29.

Robertson, S.
2016. The Differences between Digital History and Digital Humanities. In *Debates in the Digital Humanities*. University of Minnesota Press.

Siemens, L.
2009. 'It's a team if you use "reply all"': An exploration of research teams in digital humanities environments. *Literary and Linguistic Computing*, 24(2):225–233.

Siemens, L. and E. Burr
2013. A trip around the world: Accommodating geographical, linguistic and cultural diversity in academic research teams. *Literary and Linguistic Computing*, 28(2):331–343.

Siemens, L., R. Cunningham, W. Duff, and C. Warwick
2011a. "More Minds are Brought to Bear on a Problem": Methods of Interaction and Collaboration within Digital Humanities Research Teams. *Digital Studies / Le champ numérique*, 2(2).

Siemens, L., R. Cunningham, W. Duff, and C. Warwick
2011b. A tale of two cities: Implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities. *Literary and Linguistic Computing*, 26(3):335–348.

Siemens, R.
2016. Communities of practice, the methodological commons, and digital self-determination in the Humanities. *Digital Studies/Le champ numérique*.

Siemens, R., T. Dobson, S. Ruecker, R. Cunningham, A. Galey, C. Warwick, and L. Siemens
2012. Human-computer interface/interaction and the book: A consultation-derived perspective on foundational e-book research. In *Collaborative Research in the Digital Humanities*, M. Deegan and W. McCarty, eds., Pp. 163–189. Ashgate Burlington, VT.

Sonderegger, P.
2009. Creating Shared Understanding in Research Across Distance: Distance Collaboration across Cultures in R&D. In *E-Research: Transformation in Scholarly Practice*, N. W. Jankowski, ed. Routledge.

Spiro, L.
2012. "This Is Why We Fight": Defining the Values of the Digital Humanities. In *Debates in Digital Humanities*, M. K. Gold, ed. University of Minnesota Press.

Stokols, D.
2006. Toward a Science of Transdisciplinary Action Research. *American Journal of Community Psychology*, 38(1):63–77.

Svensson, P.
2011. The digital humanities as a humanities project. *Arts and Humanities in Higher Education*, 11(1-2):42–60.

Svensson, P.
2012. Envisioning the Digital Humanities. *Digital Humanities Quarterly*, 6(1).

Terras, M., J. Nyhan, and E. Vanhoutte, eds.
2013. *Defining Digital Humanities*. Ashgate.

Tsai, C.-C., E. A. Corley, and B. Bozeman
2016. Collaboration experiences across scientific disciplines and cohorts. *Scientometrics*, 108(2):505–529.

Walsh, J. P. and N. G. Maloney
2007. Collaboration structure, communication media, and problems in scientific work teams. *Journal of Computer-Mediated Communication*, 12(2):378–398.

Walter, K. L.
2012. centerNet: Cyberinfrastructure for the Digital Humanities. Technical report, University of Nebraska, Lincoln.

Wenger, E.
1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

Wilson, P.
1996. Interdisciplinary research and information overload. *Library Trends*, 45(2):192–203.

Appendix A - Questionnaire

1. What is the name or title of your collaboration? (text input)
2. Where are people participating in `Q1-answer` located? (multiple choices)
3. From what backgrounds does leadership (director, PI, supervisor, or otherwise) come? (multiple options)
    - History
    - Other humanities disciplines
    - Computational linguistics
    - Computer science
    - Cultural heritage
    - Library
    - Software development
    - Other (text input)
4. From what backgrounds do participants other than leadership come? (multiple choices)
    - History
    - Other humanities disciplines
    - Computational linguistics
    - Computer science
    - Cultural heritage
    - Library
    - Software development
    - Other (text input)
5. Are participants mostly from a humanities background or mostly from a computational background? (single choice)
    - Only from a humanities background
    - Mostly from a humanities background
    - About equal
    - Mostly from a computational background
    - Only from a computational background
6. Which of the following statements is true about `Q1-answer`? (multiple choices)
    - Participants are all contractually tied solely to `Q1-answer`
    - Participants are all contractually tied to the same organisational unit (not necessarily `Q1-answer`)
    - Participants are contractually tied to other organisational units than `Q1-answer`
    - Participants have a dual position in their contract, `Q1-answer` and another organisational unit

- Participants are contractually tied to different institutions
- Participants work only on the `Q1-answer`
- Participants work on multiple research projects
- Participants perform user research engaging humanities scholars not part of `Q1-answer`

7. What is the time frame of `Q1-answer`? (single choice)
   - Short term (working towards deadline within 4 years)
   - Semi-long term (no concrete deadline or more than 4 years)
   - Long term (no deadline)
8. What is the source of funding? (multiple choices)
   - University
   - National funding agency
   - European funding agency
   - Government (other than national funding agency)
   - Another infrastructure project (e.g. DARIAH, CLARIN)
   - Other (text input)
9. What is the physical space where the main participants of `Q1-answer` work? (single choice)
   - Lab space
   - Single office
   - Multiple offices on a single floor
   - Multiple offices on multiple floors in a single building
   - Offices in multiple buildings of a single institution
   - Offices in multiple institutions in a single country
   - Offices in multiple institutions in multiple countries
10. Where is the space or are spaces located (multiple choices)
    - In the library building
    - In the humanities building
    - In the computer science building
    - Other (text input)
11. How do you mainly communicate with other participants? (single choice)
    - Face to face
    - Video conferencing
    - Telephone conferencing
    - Email
    - Slack
    - Other communication platform(s) (text input)

12. How often do you communicate about research-related matters with participants from a different disciplinary background than your own? (single choice)
    - Daily
    - Weekly
    - Monthly
    - Every 2-6 months
    - Annually
    - Never
13. How often do you communicate about research-related matters with people from your own disciplinary background, that are not part of `Q1-answer`?
    - Daily
    - Weekly
    - Monthly
    - Every 2-6 months
    - Annually
    - Never
14. Of what organisational unit(s) is the lab a part? (multiple choices)
    - Entirely independent
    - The digital history/digital humanities group
    - The humanities faculty
    - The computer department
    - The library
    - Other (text input)
15. Which of the following does the lab provide? (multiple choices)
    - Computers
    - Scanners
    - Printers
    - Personnel (such as software developers)
    - Other (text input)
16. Which of the following statements is true about the lab? (single choice)
    - The lab is meant for people contractually tied to the same organisational unit as the lab
    - The lab is open to anyone from the humanities in the university
    - The lab is open to anyone from history in the university
    - The lab is open to anyone in the university
    - Other (text input)
17. Please describe in short the goal of `Q1-answer`? (text input)

18. What is your individual reason for joining `Q1-answer`? (text input)
19. Would you say `Q1-answer` is successful? (text input)

Appendix B - Kolmogorov-Smirnov tests

None of the variables used in this paper were normally distributed, as found through Kolmogorov-Smirnov tests, a test for normal distribution (Field, 2009, pp. 144-148):

- Disciplinary diversity $D(166)=0.38$, $p<0.01$
- Physical distance $D(166)=0.25$, $p<0.01$
- Intradisciplinary interactions $D(166)=0.28$, $p<0.01$
- Cross-disciplinary interactions $D(166)=0.27$, $p<0.01$
- Main means of communication $D(166)=0.31$, $p<0.01$

Manuscripts, Metadata, and Medieval Multilingualism: Using a Manuscript Dataset to Analyze Language Use and Distribution in Medieval England

Krista A. Murchison[1] and Ben Companjen[2]

[1]Leiden University Centre for the Arts in Society
[2]Leiden University Libraries, Centre for Digital Scholarship

1 Introduction

In the traditional linguistic model of medieval England, the Norman Conquest of 1066 caused English, which had previously been an acceptable language for literary and cultural production, to be displaced by French and sidelined in aristocratic and courtly domains. In this traditional model, English only regained its status within these 'high domains' after about two hundred and fifty years. A growing body of research has pointed to the significant structural problems with this traditional linguistic model, and it is now generally accepted that French persisted as an important domestic and aristocratic language in England for much of the late medieval period.[1] In light of this increasingly important body of research, the status of French in this period, including the contexts and implications of its use, is being re-examined.[2]

To date, studies of the status of French in medieval England have been focused on isolated examples—either of individual cases of sociolinguistic interest, or of the interplay of languages within single manuscripts or texts.[3] This example-based approach has allowed the field to productively challenge the traditional 'grand narrative' of England's linguistic situation, but since examples are, by nature, highly specific, they provide only limited insight into patterns of language use across social groups, gender, and temporal periods.

The goal of investigating these language patterns on a broader scale lay behind this project: the creation of a digital database of manuscripts containing French literature that were copied in medieval England.[4] Manuscripts—hand-written collections of

[1] See, for example, Butterfield (2009), Stein (2007) and Waters (2015).

[2] For studying examining the contexts in which French was used in the medieval period, see, for example, (Baswell, 2007, Ormrod, 2003, Waters, 2015, Watson, 2009).

[3] Studies that explore individual sociolinguistic test cases include (Butterfield, 2009, Clanchy, 1979, Waters, 2015); those focused on the interplay of languages within single manuscripts include (Stein, 2007); studies that explore the use of Anglo-Norman within individual texts are numerous and include (Baswell, 2007, Ormrod, 2003, Postlewate, 2007).

[4] Sincere thanks are due to the Europeana Foundation for the financial support that enabled this project, to the Leiden University Centre for the Arts in Society (LUCAS) for research travel support, and to Leiden University's Centre for Digital Scholarship for technical expertise and support.

texts—were chosen as the focus of this quantitative project because they provide unmatched insight into language use for a period in which no audio or spoken evidence is available. Since manuscripts are, by definition, handmade objects, they are distinct witnesses to the social contexts, patrons, and copyists that produced them.

Manuscripts functioned for their medieval users much like a binder does for modern ones: as a compilation of material to be consulted later. A medieval individual or group would select which text or texts should go in the manuscript and then either copy them out by hand or assign this task to one or more scribes. Texts in a manuscript, much like those in a binder, could be removed or added after the manuscript was originally compiled, either by the original compiler or by a later user. Manuscripts therefore represent the deliberate and conscious choices of one or more medieval users, and each manuscript, and each of its unique stages of compilation, can therefore serve as an information-dense data point about medieval language use. Taken together, these data points can be plotted to identify language patterns.

Manuscripts are a particularly good source of evidence for tracking language use in this context since they survive in far larger quantities than any other medieval textual witnesses. In particular, manuscripts can provide insight into the languages that were considered suitable for literary and documentary culture, and about who owned texts in certain languages and in which contexts. But there are also limitations to the kinds of sociolinguistic insight that manuscript data can provide. First, manuscripts, as written artifacts, cannot be taken as representatives of spoken language use in any straightforward way, so the evidence they provide, while valuable for understanding England's written culture, must be approached with caution when exploring broader sociolinguistic questions. And the division between written culture and everyday life in the medieval period could be significant; producing a medieval manuscript was an expensive and time-consuming process, which means that manuscripts were usually owned by people with social status, such as those within the Church or those with financial means. Manuscripts and their ownership patterns, then, typically provide information about the literary tastes of an elite subsection of medieval society. Nevertheless, the texts in manuscripts did reach a more diverse audience beyond their original owners, and in the absence of audio recordings or other, more populist forms of language data, manuscripts are an unmatched source of information about medieval language use.

The approach taken here has been made possible through two important developments: digital tools that enrich and assist quantitative analysis and the increased availability of digitised manuscripts and their catalogues. Medieval manuscripts have traditionally been studied almost exclusively through qualitative methods—either in isolation or through smaller-scale comparative approaches. These methods, which remain central to manuscript studies, can yield valuable information about medieval reading communities, book production and textual exchange. Over the past few decades, though, scholars have increasingly used digital technology to look for large-scale patterns in medieval manuscript datasets. Among the promising developments in this area are studies aimed at exploring the circulation and production of manuscripts in the medieval world. So, for example, Michael Sargent has explored a dataset of medieval 'bestsellers' and identified a correspondence between the number of surviving manuscript copies of a given text and its circulation numbers.[5]

Quantitative approaches—not all of them digital—have also yielded exciting results

[5] See (Sargent, 2008). Another example of using a quantitative approach to explore medieval circulation numbers is (Buringh, 2011).

in the field of medieval paleography–that is, the study of handwriting in medieval manuscripts. The new horizons in this area are suggested by the project undertaken by a team of researchers headed by Carla Bozzolo and Ezio Ornato, which analysed a large dataset of Psalm manuscripts to gain statistical information about abbreviation practices in the medieval world.[6] Quantitative approaches have also offered new insights into changes in how manuscripts were copied; these include Erik Kwakkel's analysis of approximately 350 manuscripts, which revealed several key changes in letter forms that took place in the twelfth century—changes that can now be used to support the dating of other manuscripts.[7] Although new quantitative approaches often face significant barriers—some of which are discussed below—the developments in this domain of the past few decades point to the value of such approaches for illuminating patterns that would otherwise have gone unnoticed.

In order to identify language use patterns in this quantitative way, this project gathered information about the contents and contexts of manuscripts that would be useful for understanding how French literature circulated in medieval England. For this reason, information was gathered about the number of pages (sides of a manuscript folio) dedicated to French in a given manuscript relative to Latin and English. To identify and track changes in language use and distribution, information was also gathered about a given manuscript's date of composition, and, where possible, about its medieval owners and areas of circulation. This information was compiled in a digital format so that it could be analysed in a quantitative way—a traditional humanities approach that can be greatly enhanced by the predictability, efficiency, and accessibility offered by digital technology.

With the goal of providing a set of blueprints for those seeking to undertake or support similar projects, this article describes the methodology used for compiling and analysing linguistic data from manuscripts and describes the patterns of textual transmission that can be traced particularly effectively using this methodology. Since its aim is sharing a particular framework for this kind of quantitative manuscript-based analysis, this article is focused on the methodology and digital approach of the project rather than its specific sociolinguistic findings, although some discussion of these findings has been included where relevant.[8] Quantitative approaches to medieval manuscript data are still in their infancy, so we discuss some of the current barriers to this kind of research and propose ways of structuring manuscript metadata that will facilitate future linguistic-based manuscript research.

2 Method

2.1 Preparing Manuscript Data for Quantitative Analysis

The dataset for this project was based on a list compiled by Ruth Dean and Maureen Boulton that was aimed at recording every known work of French literature from medieval England. The source list, produced in 1999, remains the most comprehensive list to date of French manuscripts from medieval England, and contains 958 items.[9] For the goal of this study, this list had to be modified somewhat, and two categories of

[6] See (Bozzolo and Ornato, 1980); for an in-depth discussion of quantitative approaches to paleography, see (Derolez, 2003).

[7] See (Kwakkel, 2012).

[8] The specific sociolinguistic findings of this project will be explored further in a future article.

[9] See (Dean and Boulton, 1999). The figure of 958 was arrived at through counting the manuscripts on the list.

manuscripts on the list had to be omitted: those copied on the continent (indicated by a * in Dean and Boulton's list) and those that contain no French but were included on the list for other reasons (marked with an 'r' in Dean and Boulton's list). The former group had to be omitted since, as continental productions, they do not reflect textual production and linguistic exchange in the area under investigation. The latter category had to be omitted because the project was aimed at gathering data about manuscripts containing French literature, so those without it did not belong in the dataset.[10] The remaining manuscripts were assigned an ID number.

In the first stage of the project, traditional catalogue data were gathered about the manuscripts, including their contents, date, linguistic profile, and owner marks—such as library tags, inscriptions and patron illustrations. This traditional information was structured in a relational model to ensure that it would be machine-readable and ready for digital analysis.[11] Manuscript identification and metadata information were therefore entered into a main 'Manuscript' table. This metadata information includes current holding information, period and place of production, explanatory notes and references for the information.

For each manuscript entry—either an individual manuscript or, where necessary, a part of a composite manuscript—data about texts and owners were entered into separate 'Text' and 'Ownership' tables. For each text in a manuscript, information recorded includes the title or incipit, the language, the start folio and the end folio. For each known owner, recorded information includes a name or reference of the owner, the period that they were in possession of the manuscript, source references and where available, whether they held clerical status at time of ownership and their gender.

The structuring of these files was aimed at upholding the internationally-recognized framework for data management known as the FAIR principles (Findable, Accessible, Interoperable and Reusable).[12] All data were stored in CSV files. CSV was chosen for several reasons; first, CSV, as a format that is compatible with existing and readily available tools like Microsoft Excel, allowed data entry to begin quickly and efficiently. Using CSV files as input and output also made archiving the results for reproducibility and reuse straightforward.

The columns in the table files and the relations between the files were described using JSON schema files that follow the CSV on the Web standard. With these schemas the CSV files can be checked for consistency and converted into RDF using available tools like COW and the RDF.rb suite.[13] This use of schemas gives the project data greater machine actionability, which is one goals of the FAIR principles.

For this project, planning for the reusability of the data was particularly key. So, with the goal of ensuring that data, in keeping with the FAIR principle of reusability, would be 'richly described with a plurality of accurate and relevant attributes' (R1) and 'meet domain-relevant community standards' (R1.3), the project gathered some traditional manuscript data that, while not needed for the particular goals of this project, would benefit those working with the dataset for other purposes.[14] This extra data included manuscript headnotes and old foliation system information. A complete

---

10 In a few cases, all noted in the online catalogue, a manuscript that was mistakenly included in Dean and Boulton's list was omitted, or a manuscript was added that had not been identified when the earlier list was compiled.

11 See `https://en.wikipedia.org/wiki/Relational_model`.

12 On the FAIR principles, see (Wilkinson et al., 2016).

13 COW: `https://github.com/CLARIAH/COW`; RDF.rb can read CSVW-described CSV files with the rdf-tabular plugin: `https://ruby-rdf.github.io/rdf-tabular/`.

14 See (Wilkinson et al., 2016).

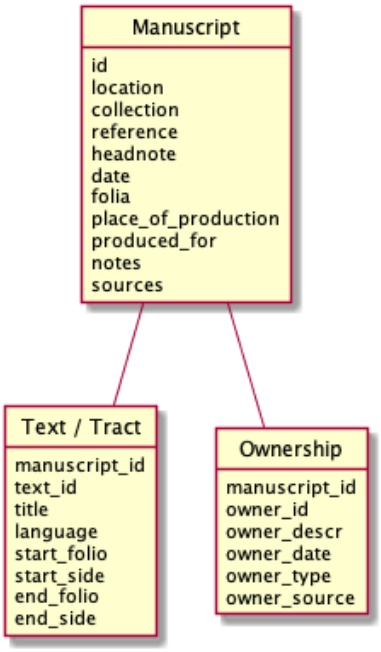

Figure 1: Class diagram showing table properties and relations

list of columns for each table is presented in Figure 1.

2.1.1 Issues in manuscript metadata collection

The manuscript description data for this project was gathered through traditional research methods since at present an automated gathering method is not practical. Historical and structural challenges faced by manuscript collections mean that many descriptions of medieval manuscripts have not been updated since they were first written during the wave of enthusiastic—but not necessarily purely academic—wave of interest in medieval sources that marked the late nineteenth and early twentieth centuries. While these centuries-old manuscript catalogues have the advantage of being free of copyright restrictions and therefore well suited to open scholarship projects such as this one, they have the disadvantage of being out of date, and relying on them too heavily could therefore lead to unnecessary mistakes in a dataset. For this project, the potential pitfalls of using out-of-copyright catalogues were mitigated through the use of more recent dating information, which was gathered wherever possible from Ruth Dean and Maureen Boulton's *Guide*.[15]

Converting traditional manuscript description data to a machine-actionable, reusable, and interoperable format also presents several notable challenges. Manuscript cataloguing, as a practice that began in the medieval period itself, evolved differently to meet the needs of the various contexts and regions in which it was used, and has historically been a widely variable practice. Armando Petrucci, in his *La descrizione del manoscritto*, distinguishes between shorter summary catalogues and longer analytical catalogues, but within these categories there exist marked differences

[15] Ideally the most recent dating information for each manuscript would be established through an in-depth analysis of every manuscript but the—likely relatively small—increase in accuracy enabled by this approach would be greatly offset by the significant time required to gather the required amount of data.

in the contents, layout, and languages of descriptions between regions, libraries, and even within individual collections.[16] The significant variation between the ways in which catalogues described manuscript contents posed a particular challenge for this project. Many catalogues—and most notably those produced by M.R. James which describe the Cambridge manuscript collections—provide only the starting points of texts within manuscripts, omitting their ending points. While these starting points are sufficient for someone looking for a particular text in a manuscript, they are unfortunately insufficient for a quantitative study such as this one, since the end point of one text within a manuscript cannot necessarily be inferred from the starting point of the text that follows it.

Aside from causing gaps in information, the inconsistencies between manuscript descriptions pose a significant barrier to quantitative analysis more generally. Digitization of nineteenth and early twentieth-century catalogues has made these more accessible for analysis, and it is theoretically possible to create a program that could gather and process description information from diverse catalogues in a large-scale way, but at present, such an endeavor would be hindered by the significant differences between description structures and would undoubtedly prove less reliable than collecting such data manually.

There have been various attempts to establish a standard for manuscript description, including the widely used method given in Raymond Clemens and Timothy Graham's *Introduction to Manuscript Studies*, but none have caught on.[17] Within the digital realm, while there have been attempts to design online manuscript catalogues with interoperability in mind, there is currently no established metadata standard for manuscript description.[18] Nor is such a standard likely to be established in the near future, given the piecemeal ways in which online manuscript catalogues have been developed, and the sheer number of institutions that provide online descriptions of their manuscript collections–two challenges which may be counted among the broader barriers to semantic web development. Moreover, libraries and archives are often prevented from adopting the metadata structures that have been developed for manuscripts due to funding or structural staffing issues.[19]

For this project, the most prevalent and significant omission in manuscript metadata was information about the languages of individual manuscript texts; this information is not included consistently in most existing manuscript catalogues and databases.[20] This is due, in part, to a lack of clear guidelines for recording this information. The influential *Descriptive Cataloging of Ancient, Medieval, Renaissance, and Early Modern Manuscripts (AMREMM)*, a guide for adapting the MARC21 record structure of library catalogues for manuscript description, is largely open-ended in its guidelines for recording language use; it encourages cataloguers to note the 'language or languages employed in an item' and to 'provide more detailed notes in the records for individually

---

16 See (Petrucci, 1984).

17 See (Graham and Clemens, 2007, pp. 129-135).

18 For attempts to establish a clear and consistent framework for manuscript metadata, see, for example, the Dublin Core Application Profile proposed by Bair and Steuer (2013); see also the guidelines for adapting MARC21 record structure for manuscript description in (Pass, 2002), available here: `https://rbms.info/dcrm/amremm/`.

19 For the ways in which funding and other structural issues pose barriers to cataloguing special collections, see (Bair and Steuer, 2013, pp. 2-3).

20 Of the catalogues used for this project, the British Library's online catalogue alone contained consistent data about the language of individual texts within manuscripts, but this information is not available for some of its collections—including the Sloane collection—and where it is available it unfortunately contains errors.

analyzed works' but only 'if desired.'[21] It is perhaps not surprising, in light of the well-documented funding and staffing issues encountered by many special collections, that including such detailed information in manuscript metadata has often not been a priority. Moreover, not all libraries have opted to follow the AMREMM guidelines and library cataloging systems may not support the description of manuscripts in a way that makes their contents information suitably findable.[22]

On a positive note, the new electronic Bibliotheca Neerlandica Manuscripta (eBNM+) is an example of a database with detailed contents information; it treats individual texts within a manuscript as separate entities and includes language information as a property of each separate entity, although the catalogue does not provide information about all the texts within each manuscript.[23] This database contains medieval manuscripts produced in the Low Countries and has a focus on Middle Dutch manuscripts, so it could not be used as a source in this project. Nevertheless, the structure of the catalogue shows that there is interest in, and initiatives in support of, this level of cataloguing.

For this project, missing catalogue information could on occasion be inferred from the titles of the constituent works included in the catalogues, but these titles are not always reliable or helpful indicators of language— especially for medieval French texts, which often appear under Latin titles. This, and other limitations of existing catalogue descriptions, led to gaps in the project dataset; where possible, these have been remedied through work with digital manuscript facsimiles or, in some cases, archival work, but of course neither approach is practical for large-scale data collection.

Aside from the limitations posed by existing catalogue records, any attempt to render manuscript data machine readable is met with additional challenges posed by the complex, multivalent and often enigmatic nature of the manuscripts themselves. To give a simple example, the so-called Black Book of Christ Church College Dublin was copied in a series of stages, which adds complexity to the process of assigning a single date to the manuscript in its current form, or to analyse it as a single language use data point. To represent such multi-stage manuscripts accurately, these manuscripts were entered into the catalogue as a series of individual stages, with each stage assigned a unique manuscript ID.

Adding to the challenges of recording manuscript data in a machine-readable format is the difficulty involved in arranging medieval heritage data into the kind of clean categories that facilitate machine-assisted data analysis. Among the difficulties involved in this project was the process of classifying manuscript owners as man or woman and as lay or clerical, which was done in order to identify linguistic patterns among these owners.[24] The process of compiling and sorting this data revealed significant limitations with the established classification structure; first, determining the gender of medieval people generally required guesswork based on their first names. And some medieval lives, such as those of anchorites or students, resisted the lay/clerical binary required for this classification. These challenges can be counted among the broader

---

21 AMREMM (Pass, 2002, p. 54).

22 AMREMM preceded the introduction of the RDA guidelines for description that are in use in some libraries.

23 The database is accessible via `https://bnm-i.huygens.knaw.nl/`; information about its history is available at `https://www.huygens.knaw.nl/ebnm/?lang=en`.

24 For the sake of this project, 'clerical' was used in the broadest possible sense to refer to someone whose vocation fell primarily within the bounds of the established Church; nuns, therefore, were counted among the clergy although they were not, strictly speaking, members of the clergy in the eyes of the medieval Church.

complications that arise when working with humanities data, which is often found deeply embedded in social and historical contexts. But rather than signaling inherent problems with digital humanities approaches, these challenges point to the importance of approaching the results of quantitative humanities projects with a sensitivity to, and awareness of, its multivalent social and historical contexts.

2.2 Calculating Language Distribution

With the aim of calculating the distribution of languages in each manuscript, the number of folia occupied by each text was recorded in each manuscript's 'Text' table. This table noted the specific language of each text if it was in one of the languages under investigation (French, Latin or English), and described the language as 'other' in the rare cases of other languages—including Hebrew and Greek. For the sake of consistency, texts were recorded as such only when they occupied four or more ruled manuscript lines; some interesting or noteworthy writing under four lines was also recorded, but this information was placed in the 'Notes' field. An exception to this approach had been made for glosses, which were recorded in the 'Text' tables; although these often consist of fewer than four 'ruled manuscript lines' they operate as part of a larger textual framework and therefore were considered noteworthy linguistic data.

In this system, pages containing multiple texts in different languages presented a challenge to data representation. A page (one folio side) that contained more than one text was assigned partially to each of its constituent texts; so a page containing both a French and an English text, for example, was counted as .5 of a page of French and .5 of English. Greater accuracy for page ranges may have been achieved by counting the number of lines or words of a given text on each page but these methods, which prioritize either length of words on one hand or wordiness of a text on the other, misrepresent data to some extent in their own way and any extra accuracy they would have offered would not have been worth the enormous time commitment involved in counting individual lines of text for each manuscript.

Calculating the number of pages occupied by a given text was done using a Python script and the Pandas code library, which is excellent for working with tabular data.[25] The code works by calculating the number of pages covered by each individual text and assigning them to their associated language category. To allow a mathematical calculation of page ranges, the collected folio numbers for the start and end points of each text (e.g. 1r, 3v) were converted to their ordinal variants (1, 6, respectively). This method worked well for most folia, but not for those assigned roman numerals, which are generally used by manuscript cataloguers to indicate the flyleaves or endleaves. Roman numerals were therefore initially converted in the same way, but after conversion, the code increased them by 100000 in order to avoid reusing the same ordinal page numbers in calculations. So, for example, IIIr was converted to 100005 for its page range calculation. Since no manuscript has page numbers even close to that number, this method prevented the unintended reuse of page numbers.

Once this method of conversion is in place, the number of pages per language for a whole manuscript can be calculated by grouping the texts per language and summing the number of pages in each group. The effect is that, once the individual tracts are

[25] Because of variations in character encoding between CSV files edited with Excel 2010, the script checks character encodings with the chardet library before Pandas loads the file. Other preprocessing included removing empty rows and normalising cell contents by trimming whitespace and converting the casing of values.

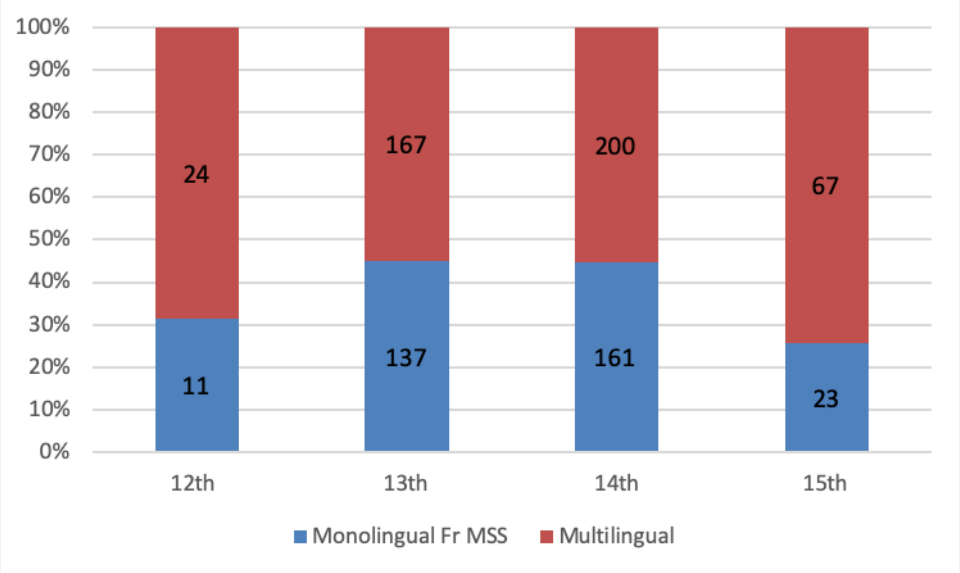

Figure 2: French texts in monolingual and multilingual manuscripts

described in tables, the script can identify the percentage of each language in each manuscript.

2.3 Presenting the Results in HTML

While the main output for this project was the archivable CSV files and their analysis, we also wanted to present the project data in a format that was more user friendly, so the project data were also presented as a website.[26] With the goal of sustainability in mind, the data and results of the language analysis were presented as a static, rather than dynamic, website. A static website was chosen for this project since maintaining the code for a dynamic website after the end of the project was deemed unfeasible and, as others have observed, static websites have the advantage over dynamic ones of offering greater flexibility in terms of preserving and moving websites that consist of text files alone.[27]

3 Results: Preliminary Findings

The statistical analysis of language use in these manuscripts has yielded valuable results about the linguistic situation in England during the centuries following the Norman Conquest. Most notably, this analysis has revealed that French literature circulated on its own relatively infrequently.

As shown in Figure 2, from the twelfth to fifteenth centuries, monolingual manuscripts containing French literature remained in the minority. Only 11 of the 35 manuscripts dated to the twelfth century are monolingual (31%) and the proportion of monolingual French manuscripts in the dataset remains relatively consistent throughout the period under investigation; it increases only slightly in the thirteenth and fourteenth centuries and then decreases in the fifteenth. The extent and nature of

[26] See https://leidenuniversitylibrary.github.io/manuscript-stats/.
[27] See Visconti (2016).

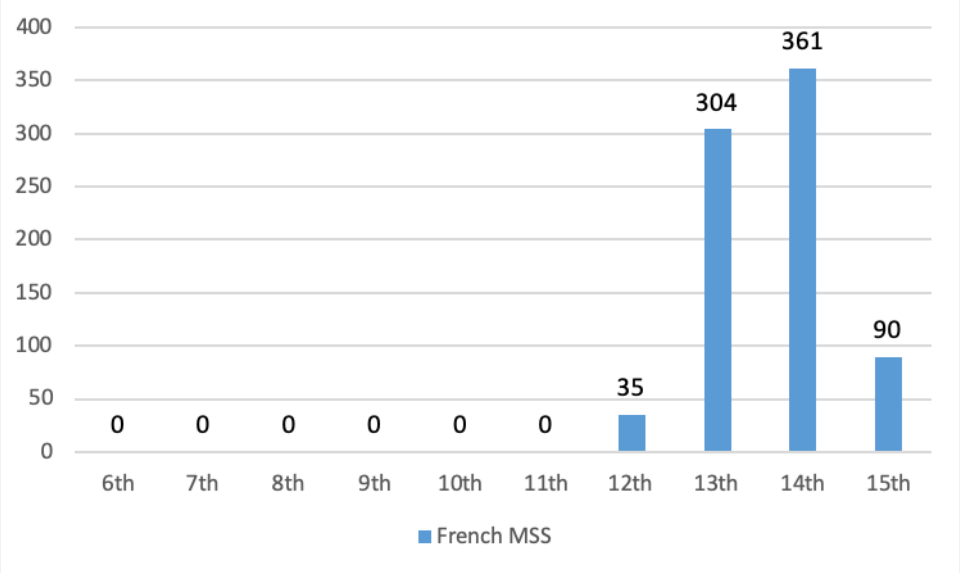

Figure 3: Figure 3. French Manuscripts by Century Copied

the multilingual contexts of England's French writing has not been commented on before and represents a significant result of this project.

Given that manuscripts were often designed with a purpose, and can often provide insight into the literary tastes of their patrons, these findings suggest that French literature was most often read in multilingual contexts and by multilingual patrons.[28] The multilingual contexts of French seem to have remained relatively consistent throughout the period under investigation, a finding that challenges the traditional narrative in which French was increasingly sidelined by English in the later period. This is significant because it suggests that the use of French in the centuries following the Norman Conquest was not usually dictated by a patron's linguistic limitations and speaks to a high level of multilingualism of the patrons of French literature in medieval England.

The dating data provided by the manuscript catalogue has also provided valuable information about the use of French in written contexts in medieval England. In particular, it appears that there was an increase in the copying of French literature in the thirteenth and fourteenth centuries, as seen in Figure 3.

The increase in the thirteenth and fourteenth centuries is remarkable; it suggests that the period in which, in the traditional 'grand narrative', French was supposedly on the decline in England was, in fact, the most significant period for the copying of French literature.

But of course, these data on their own tell us very little, since the apparent increase in copying in the thirteenth and fourteenth centuries could, in theory, be skewed by the contingencies of manuscript survival or by broader patterns of manuscript copying. It is therefore necessary to contextualize this data within broader patterns of manuscript production and survival in England.

Ideally, this would be done using a large-scale dataset for the temporal distribution of manuscripts produced in England as a whole, but at present there is no such dataset available, and given the challenges that current manuscript catalogues present with

[28] For the idea that manuscripts can be read for signs of deliberate compilation—what Seth Lerer terms 'anthologistic moments'—see, for example, Lerer (2003), Nichols (2015).

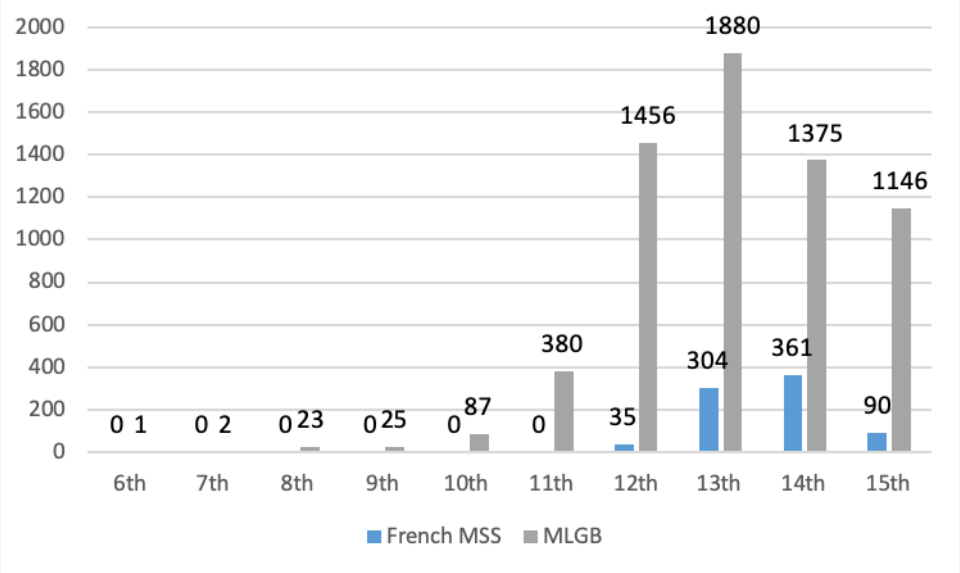

Figure 4: Figure 4. Total Manuscripts with French Literature and Total Manuscripts in Ker (1964)

respect to interoperability and findability already described, it is not currently feasible to produce one. However, some insight can be gleaned by comparing the temporal distribution of French manuscripts to one available catalogue of medieval manuscripts: Neil Ker's *Medieval Libraries of Great Britain* (see Figure 4).[29] This catalogue, which was recently digitized and therefore available for quantitative analysis, contains information about all known manuscripts produced in England in any language that can be traced back to specific medieval libraries. It therefore provides some insight into manuscript production in England as a whole, although since institutional libraries were most often tied to monasteries and other clerical organizations in the medieval period, the dataset is skewed toward clerically-owned manuscripts and so may not accurately reflect manuscript production in England as a whole.[30] Nevertheless, in the absence of other evidence, this dataset provides a valuable point of comparison.

The lack of French manuscripts prior to eleventh century is of course unremarkable given the relatively few French speakers in England prior to 1066. But the difference in the temporal distribution of manuscripts after the eleventh century is striking. A relatively high number of manuscripts on Ker's list were produced in the twelfth century, providing quantitative support for Ker's qualitative observation that the century after the Norman Conquest was the most significant for book production in England in general.[31] But the number of surviving manuscripts containing French literature produced in this period was comparatively quite limited. This stands in powerful contradiction with the traditionally held view that the efflorescence of French literature emerged as a direct response to the Norman Conquest of 1066.[32]

The comparison reveals that the first wave of manuscripts containing French literature, which dates to the thirteenth century, appears to have been part of a broader

[29] See Ker (1964).

[30] Using Ker's catalogue as a model for exploring trends in manuscript production is in keeping with the methodology of Buringh and Luiten Van Zanden (2009), who explore manuscript production in Europe more generally.

[31] The century after the Conquest has been described as 'the greatest in the history of English book production' Ker (1964).

[32] For this traditional view, see the introduction above.

increase in written production in medieval England. Generally speaking, increased book production was undoubtedly both a result of, and contributed to, the rise of commercial centres, the growth of universities, the proliferation of monasteries, and the development new technologies for fitting text on the page.

Why did the first wave of French literary production fall during this period and not, as might be expected, in the century following the Norman Conquest? Following the hypothesis put forth by Michael Clanchy, this particularity may be at least partially explained by the growing role of French in the thirteenth century as a language for international affairs, mercantile exchange and business transactions.[33] Within England, French was also gaining a foothold within a legal context in this period; in the last quarter of the thirteenth century, the dominant language for written statutes changed from Latin to French.[34] We should not be surprised to find that an increase in legal and administrative writing in French would come accompanied by a new interest in literary production.

These findings, which will be explored in greater depth within a sociolinguistic framework in the future, provide the first quantitative evidence for the persistence of French writing in England in the centuries following the Norman Conquest and suggest that the role of French in medieval England was not, as was once thought, an immediate result of the Conquest itself, but rather of a network of complex economic, social and international developments that took place in the thirteenth century. More broadly, the data suggest that England's literary culture remained multilingual throughout much of the medieval period and support an ongoing challenge to the traditional 'grand narrative' of England's linguistic history.

4 Discussion: Facilitating Future Manuscript-Based Research

The script that was used to calculate language distribution in this project is available online and comes accompanied by documentation intended to facilitate reuse.[35] Its function in this project—calculating the language distribution in a set of manuscripts, could also be deployed for exploring other manuscript datasets in a quantitative way. For example, comparing the linguistic distribution of surviving manuscripts between different medieval libraries would undoubtedly reveal valuable quantitative information about language use in these communities. In the future, if more consistently structured manuscript catalogue metadata becomes available—at the level of individual texts in manuscripts—it would also be possible to explore the linguistic distribution within manuscripts produced in a given region. For example, one could identify whether English was used more commonly in manuscripts produced in the West Midlands—typically considered a conservative linguistic area—than in those produced in the East.

Future work could also incorporate the growing knowledge of manuscripts' provenance that various projects and databases are publishing. The Schoenberg Database of Manuscripts, for example, provides a large user-edited knowledge base of manuscript

[33] See (Clanchy, 1979, p. 214), which describes 'the advance of French as an international literary and cultural language, particularly in the thirteenth century' and its increased use in mercantile and business contexts.

[34] Spence (2013) writes that French was 'used more frequently in legal and administrative documentation from the second half of the thirteenth century' (3); in particular, he finds that 'Statutes were made in Anglo-Norman instead of Latin from 1275' (4).

[35] See Code and Data Availability below.

transactions.[36] This database will, over time, likely offer more complete and more rigidly structured data on owners that those that were collected for this project.

On a broader scale, the code produced for this project, by calculating the number of pages occupied by a manuscript's constituent parts, could be beneficial to those wanting to compare the manuscript contexts of various texts. For example, the code could help provide quantitative data about changes in the makeup of manuscripts containing the *Canterbury Tales* over time. With more consistent manuscript metadata, the code could also be adapted to enable largescale comparisons of how various types of texts were copied. For example, data could be gathered into what percentage of thirteenth-century literary production was dedicated to chronicles or to poetic texts. While scholars make claims about the importance of various types of writing in a qualitative way, preparing quantitative data could yield new insights into medieval tastes, interests, and desires.

Given the significant and documented divergences among current approaches to manuscript description, however, quantitative studies such as this project require a great deal of preparation in order to render data machine-actionable, and therefore face barriers. So while it would theoretically be possible to very quickly analyze patterns in, for example, the dating of all surviving manuscripts produced in England, at present the lack of established and adopted metadata standards makes such a task prohibitively labor intensive. This project therefore highlights the need for more consistently structured manuscript data. Greater structural and financial support for this type of structured cataloguing would enable more efficient investigation into the textual culture of the medieval period and, in so doing, shed new light on other emerging cultural history questions.

5 Code and Data Availability

The code for analysing and converting the data to HTML is hosted at GitHub (`https://github.com/LeidenUniversityLibrary/manuscript-stats`) and archived in Zenodo: `https://doi.org/10.5281/zenodo.1472267`.

The normalised input files and results have been archived in EASY: `https://doi.org/10.17026%2Fdans-zxr-juar`.

References

Bair, S. A. and S. M. B. Steuer

2013. Developing a Premodern Manuscript Application Profile Using Dublin Core. *Journal of Library Metadata*, 13(1):1–16.

Baswell, C.

2007. Multilingualism on the Page. In *Middle English: Oxford Twenty-First Century Approaches to Literature*, P. Strohm, ed., Pp. 38–50. Oxford: Oxford University Press.

Bozzolo, C. and E. Ornato

1980. *Pour une histoire du livre manuscrit au Moyen Âge: trois essais de codicologie quantitative*, number 2 in Textes et études. Paris: Centre national de la Recherche scientifique.

[36] See `https://sdbm.library.upenn.edu/`.

Buringh, E.
2011. *Medieval manuscript production in the Latin West: explorations with a global database*, volume 6 of *Global economic history series*. Leiden: Brill.

Buringh, E. and J. Luiten Van Zanden
2009. Charting the 'Rise of the West': Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries. *The Journal of Economic History*, 69(2):409–445.

Butterfield, A.
2009. *The Familiar Enemy: Chaucer, Language and Nation in the Hundred Years' War*. Oxford: Oxford University Press.

Clanchy, M.
1979. *From memory to written record: England 1066-1307*, 2nd edition edition. Oxford: Blackwell.

Dean, R. J. and M. B. Boulton
1999. *Anglo-Norman Literature: A Guide to Texts and Manuscripts*. London: Anglo-Norman Text Society from Birkbeck College.

Derolez, A.
2003. Possibilités et limites d'une paléographie quantitative. In *Hommages à Carl Deroux IV. Archéologie et Histoire de l'Art*, P. Defosse, ed., Pp. 98–102. Brussels: Latomus.

Graham, T. and R. Clemens
2007. *Introduction to Manuscript Studies*. Ithaca, NY: Cornell University Press.

Ker, N. R.
1964. *Medieval libraries of Great Britain: a list of surviving books*, number 3 in Royal Historical Society guides and handbooks ; no. 3; 15, 2nd ed. edition. London: Offices of the Royal Historical Society.

Kwakkel, E.
2012. Biting, Kissing and the Treatment of Feet: The Transitional Script of the Long Twelfth Century. In *Turning over a new leaf: change and development in the Medieval manuscript*, E. Kwakkel, R. McKitterick, and R. Thomson, eds., Studies in Medieval and Renaissance book culture, Pp. 76–112. Leiden: Leiden University Press.

Lerer, S.
2003. Medieval English Literature and the Idea of the Anthology. *PMLA*, 118(5):1251–1267.

Nichols, S. G.
2015. What is a Manuscript Culture? Technologies of the Manuscript Matrix. In *The Medieval Manuscript Book*, M. Johnston and M. Van Dussen, eds., Pp. 34–59. Cambridge: Cambridge University Press.

Ormrod, W.
2003. The Use of English: Language, Law, and Political Culture in Fourteenth-Century England. *Speculum*, 78(3):750–787.

Pass, G. A.
2002. *Descriptive cataloging of ancient, medieval, Renaissance, and early modern manuscripts*. Chicago: Association of College and Research Libraries.

Petrucci, A.
1984. *La descrizione del manoscritto : Storia, problemi, modelli*. Rome: La Nuova Italia Scientifica.

Postlewate, L.
2007. Preaching the Sins of the Ladies: Nicole Bozon's "Char d'Orgueil". In *Cultural Performances in Medieval France: Essays in Honor of Nancy Freeman Regalado*, E. Doss-Quinby, R. L. Krueger, and E. Burns, eds. Woodbridge: Boydell & Brewer.

Sargent, M. G.
2008. What do the numbers mean? A Textual Critic's Observations on some Patterns of Middle English Manuscript Transmission. In *Design and distribution of late medieval manuscipts in England*, M. Connolly and L. R. Mooney, eds., Pp. 205–244. York: York Medieval Press.

Spence, J.
2013. *Reimagining History in Anglo-Norman Prose Chronicles*. Woodbridge: Boydell & Brewer.

Stein, R. M.
2007. Multilingualism. In *Middle English: Oxford Twenty-First Century Approaches to Literature*, P. Strohm, ed., Pp. 23–37. Oxford: Oxford University Press.

Visconti, A.
2016. Building a static website with Jekyll and GitHub Pages. *Programming Historian*.

Waters, C.
2015. *Translating Clergie: Status, Education, and Salvation in Thirteenth-Century Vernacular Texts*. Philadelphia: University of Pennsylvania Press.

Watson, N.
2009. Lollardy: The Anglo-Norman Heresy? In *Language and Culture in Medieval Britain: The French of England, c.1100-c.1500*, J. Wogan-Browne, ed. Woodbridge: Boydell & Brewer.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. d. S. Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. t. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. v. Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. v. d. Lei, E. v. Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons
2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Analysis of Fidel Castro Speeches Enhanced by Data Mining

Sergio Peignier*[1] and Patricia Zapata[2]

[1]Univ Lyon 1, INSA Lyon, INRA, BF2I, UMR0203, F-69621, Villeurbanne, France
[2]Carrera de Lingüística e Idiomas, Universidad Mayor de San Andrés, La Paz, Bolivia

Fidel Castro was a major Cuban politician of the twentieth century who influenced different left-wing regimes and political movements in Latin America and around the world. His ability to seduce the masses and captivate his audience relied to a large extent on his rhetorical abilities. Therefore, studying Castro's political speeches is a crucial step towards understanding his political success. Some previous studies have addressed this issue, mostly using small discursive samples and only a few speeches. However, using a small and possibly non-representative sample is likely to lead to biased results. To overcome this problem, this work was carried out on a large corpus of 1,018 speeches and 7,548,480 words, combining state-of-the-art data mining tools and the linguistic discourse analysis methodology proposed by Patrick Charaudeau. Combining both techniques, we provide here a more representative characterization of Castro's main discursive strategies.

Keywords: Discourse Analysis, Word-Embedding, Subspace Clustering, Hierarchical Clustering

1 Introduction

Fidel Castro came to power in 1959 as the leader of the Cuban Revolution, and governed Cuba until 2008 (e.g. Gott (2007)). During this period, he had a strong presence in the political arena, influencing different left-wing regimes and political movements in Latin America and around the world (e.g. Padmos et al. (2017)). His ability to seduce the masses relied, to an important degree, on his rhetorical skills. Indeed, according to Charaudeau (2009b), the ability of a politician to captivate people's attention depends to a large extent on his rhetorical capabilities. Thus, speech skills are crucial for a politician who wants to persuade an audience to join his cause, making such abilities

* Sergio.Peignier.Zapata@gmail.com

the finest political weapons. Therefore, studying Castro's political discourse is a crucial step towards understanding his political success.

According to Gee (2014), the discourse analysis field is divided into 'micro-communities', that develop their own methodologies based on particular paradigms. Even if different approaches tackle discourse analysis from different angles, they often share common terminologies and tools; that blur the frontiers between them and complexifies their categorization. Indeed, there is no single classification, and many taxonomies of discourse analysis methodologies have been created (e.g. Barry (2002), Maingueneau (2016)). Nevertheless, some general and important discourse analysis families have been identified by several authors. For instance, sociolinguistics (e.g. Hudson (1996)) studies the variants of language usage in a linguistic community; conversational analysis (e.g. Richards and Schmidt (1983)) focuses on talk-in-interactions; critical discourse analysis (e.g. Fairclough (2013)) studies how language establishes and reinforces power relationships within social groups; pragmatic analysis (e.g. Widdowson (1995)) studies the intentions of the audience and the speaker, as well as the context of the speech; lexicometric analysis (e.g. Pêcheux (1995), Wiedemann (2013)) is a computer-assisted family that can study a large corpus by characterizing its *structure* (e.g. vocabulary, grammar). To some extent, this technique has also been used to study discursive strategies by quantifying and interpreting the co-occurrences of some *pivot words* (e.g. words chosen ad hoc, most common words). Unlike the lexicometric family, most other methods generally extract, manually and systematically, all the underlying discursive strategies used by the politician in a corpus.

Since systematic studies are *time-consuming* and *complex* tasks, most previous systematic investigations have been conducted on small discursive samples only, containing few speeches. However, *small* and possibly *non-representative* samples may lead to biased interpretations, as shown extensively in statistics (e.g. Ellis (2010)). The family that is less impacted by this problem is the computer-assisted lexicometric one. So, at first glance, it might make sense to use *only* lexicometric methodologies. However, this sole paradigm cannot replace all the other ones, since each paradigm has been developed to address different scientific questions. Indeed, each family has its own interests and merits, and all are complementary tools for the discourse analyst.

In this work, we extend the recognized, non-lexicometric discourse analysis methodology created by Charaudeau (1995) by combining it with state-of-the-art data mining tools to study a corpus of 1,018 speeches and 7,548,480 word occurrences. For this purpose, our method uses word-embedding (e.g. Turney and Pantel (2010)) and subspace clustering (e.g. Kriegel et al. (2009)) to partition the vocabulary of the corpus into clusters of words sharing the same discursive context. Next, our method organizes intra-cluster words in dendrograms, applying hierarchical clustering. In this article, we refer to such structures as *dendrogram prototypical discourses*, or simply *DP-discourses*. The key intuition in the design of our method is to model the entire corpus, as a set of DP-discourses, and then study the most representative ones, using Charaudeau's discourse analysis methodology. In this context, our main research question could be stated as follows: Does the analysis of a large corpus, using a hybrid approach that combines a traditional discourse analysis methodology and data mining tools, provide new insights and a more representative characterization of Castro's discursive strategies?

In this paper, we used the latter approach to characterize the main discursive strategies used by Castro. Here we reveal that Castro presents himself as an authority, an expert, committed to his duties and identified with his audience. Moreover, he

presents himself as a hero that protects people against several enemies, which are depicted as the source of all problems. In this context, he represents his audience either as a victim or as a hero fighting against the enemies. In addition, he frequently alludes to the economic and social progress of his country under his administration. All these elements aim at evoking strong feelings in his audience, such as heroism, pride, hope and fear. Finally, Castro uses elaborate and detailed descriptions to increase the veracity of his speeches.

The contribution of this paper is two-fold: On the one hand, we propose a data mining framework that models a large corpus as a set of representative DP-discourses, which can be studied easily by means of non-lexicometric methodologies. On the other hand, we provide a representative landscape of Fidel Castro's discursive strategies. Indeed, this work is the first to conduct a systematic study of more than 1,000 speeches and 7,500,000 word occurrences, using the non-lexicometric methodology proposed by Patrick Charaudeau.

The rest of the paper is organized as follows: Section 2 presents related works; Section 3 introduces Charaudeau's discourse analysis methodology; Section 4 describes our data mining framework; Section 5 describes the corpus collection and composition; Section 6 presents the discourse analysis of Castro's speeches themselves; Section 7 compares our main findings to the main conclusions of previous works; Section 8 concludes this paper by offering a summary, and perspectives for future research.

2 Related works

Given Castro's political importance, several previous works have studied his rhetorical abilities from different perspectives, using varied methodologies. The presentation of these works is divided into two parts: Section 2.1 presents earlier works based on non-lexicometric methods, while Section 2.2 focuses on lexicometric approaches.

2.1 Non-lexicometric approaches

The pioneering work of Joyner (1964) used a discourse analysis methodology based on Aristotelian principles of rhetoric to study three well-known speeches. Another pioneering work, Fagen (1965), identified frequent topics and discursive strategies and studied their interrelations in order to understand the mechanisms enhancing Castro's charisma. More recently, Nieto et al. (2002) have developed a *conversational analysis* method to study the emotions conveyed by Fidel Castro and Hugo Chávez during their first conversation that was broadcast on radio and television. Belisario (2010) combined techniques from *cognitive linguistics* and *critical discourse analysis* to study discursive strategies based on metaphors in three well-known Castro's speeches. Recently, Reyes (2011) combined *sociolinguistics* and *critical discourse analysis* to study the discursive roles assumed by a politician (e.g. narrator, interlocutor), using as sources all speeches delivered by Castro between the 15th of April and the 14th of August of 2005 (124,321 word counts in total). Considering corpus size, this work has been the most representative systematic study to date.

2.2 Lexicometric approaches

The lexicometric methodology has been used by Serge De Sousa to analyze the chronological evolution of Castro's speeches. In one of his first works, De Sousa (2009a) studied 42 speeches delivered by Castro each 26th of July, for the national day of Cuba,

between 1959 and 2004. De Sousa (2009a) clustered these speeches in five historic periods, using two dimensionality reduction techniques: correspondence analysis, developed by Benzécri et al. (1973)) and the so-called 'analyse arborée' created by Luong and Mellet (2003). In order to study the temporal evolution of discursive topics, De Sousa (2009a) extracted the 42 most frequent words within each period, and compared the evolution of these lists of terms. De Sousa (2009a) presented chronological interpretations on his observations, taking into account historical events that characterized each period. Lately, De Sousa (2012) has extended this work, and obtained similar results, by considering all available speeches.

In a different work, De Sousa (2009b) used the entire corpus of Castro's speeches to study the evolution of the concept 'pueblo' [people] in the discourses. First, the author quantified the frequency of this term over time, showing that Castro used it more often between 1959 and 1964, during the early period of his rule. The author also tracked the evolution of the *semantic network* of 'pueblo', by extracting terms that had a significantly higher and lower co-occurrence with this term. De Sousa (2009b) showed that Castro conveyed an idealized representation of the people: conscious, confident, strong, proud and revolutionary.

Until now, no previous work has aimed at extracting systematically the different discursive strategies from a large corpus of Castro's speeches. While early systematic approaches relied on a small corpus, ranging from three to a few dozens of speeches, lexicometric approaches have considered many speeches but only studied the evolution of the frequency and the semantic network of a few terms.

3 Discourse analysis methodology

The discourse analysis presented in this article is based on the so-called semio-pragmatic methodology, a non-lexicometric approach developed by Charaudeau (1995). The semio-pragmatic approach is a well-recognized methodology that has introduced some important founding principles into the field (Weizman, 2008), and Charaudeau is recognized as one of the most representative authors of the French School of discourse analysis according to Weizman (2008). This methodology is based on the Aristotelian classification of the art of rhetoric, which consists of three families of discursive strategies, called Ethos, Pathos and Logos. In this section, we present the major strategies of these three families, describing their underlying objectives, i.e. their expected impact on the audience. Nevertheless it should be noted that the outcome of a given discursive strategy, i.e. the audience reaction, may differ from the expected outcome, which also depends on external circumstances. Indeed, a discursive strategy can be said to have only a potential influence on a specific audience under particular circumstances. The discourse analysis methodology presented hereafter does not aim at characterizing the reaction of the audience; it only aims at identifying underlying discursive strategies.

3.1 Ethos

Ethos strategies allow the speaker to build his discursive identity, they strengthen his credibility and they enable the identification between the speaker and his audience. Hereafter are described the main mechanisms from the Ethos family that were identified by Charaudeau (2009a).

Discursive identity: A politician alternates between three discursive identities: The *"Me" identity*, when he speaks only on his behalf, using the first person singular; the

"Me-Us" identity, when he also speaks on the audience behalf, using the first person plural; and the *"Me-Spokesman" identity*, when he speaks on behalf of a doctrine or ideal.

Embody a credible character: A politician reinforces his legitimacy, by personifying credible characters, such as: 1) a *leader*, who imposes his decisions and emphasizes his institutional *authority of power*; 2) an *expert*, who exhibits his analytical competences and knowledge of a given topic; 3) a neutral *witness*, who removes from his speech any indications of personal judgment; 4) a person *committed* to his ideals, who vehemently defends his ideas as unquestionable truths; 5) a *charismatic friend or relative*, who has a strong identificatory relationship with his audience.

3.2 Pathos

Pathos strategies aim at persuading the audience by bringing out feelings and passions. Within this category, Charaudeau (2008, 2011) identified three major mechanisms, which are presented hereafter.

Recruitment process: This strategy aims at leading the audience to accept the speaker's project willingly. To do so, the speaker refers to classic positive values, such as social welfare (e.g. freedom, justice, security); national or regional belonging (e.g. nationalism, regionalism); religious, ethnic or ideological belonging; development, economic growth and technological progress; and moral values (e.g. honesty, commitment).

Rhetoric of effects: The goal of this mechanism is to stir feelings and passions in the audience, in order to predispose people to share the speaker's point of view. Indeed, it has been extensively shown in the literature that emotions can have an important influence on decision-making (e.g. Janis and Mann (1977), Schwarz (2000)). In the context of discourse analysis, the study of the rhetoric of effects mainly focuses on basic emotions that were reported in the field of cognitive sciences (e.g. Ekman (1992), Lövheim (2012), Plutchik (2001)): Usually the speaker aims at provoking feelings such as threat, fear, compassion, hope and pride. To do so, politicians may use the *dramatization* strategy, which consists in telling dramatic life stories that involve several characters that the audience can identify with or reject.

Triadic scenario: Politicians' speeches are often organized as a *triadic scenario*, which contains the following three elements: 1) A current or latent *disastrous social situation* is described by the speaker to induce the audience to a state of anguish and lead the public to speculate about the origin of the problems. 2) An *enemy* is pointed out by the politician as the cause of all problems. Enemies are either clearly identifiable (e.g. political party) or vague entities (e.g. ethnic groups), and they are commonly embodied by the speaker's political adversaries. 3) A *hero* is proclaimed by the speaker as society's savior and protector against enemies. Heroes are either abstract entities (e.g. social classes) or real persons (e.g. the speaker himself). This mechanism proposes a seductive imaginary where the audience is both the hero and the main beneficiary.

3.3 Logos

The Logos strategies are used to convince the audience through logical reasoning and argumentation. According to Charaudeau (2005), Logos is used less often than Ethos and Pathos in the context of political speeches. Indeed, for politicians it is less important to explain concepts in a logical way than obtaining the audience's support by using the most efficient strategy. Moreover, in the political context, the goal of

Logos is limited to increasing the truthfulness of the speech. The main Logos strategies identified by Charaudeau (2005) are described hereafter.

Singularization and essentialization: The *singularization* strategy aims at reducing the number of ideas exposed in the speech, keeping the attention of the audience focused on a few concepts. The *essentialization* strategy represents complex concepts by using only a few words. Once a concept has been essentialized, the audience does no longer need to reflect on it to make sense of it, which reduces the audience's mental effort. Both strategies are often combined to form a so-called *formula* strategy. According to Charaudeau (2005), a formula produces a strong feeling of evidence and attraction in the audience. Similar strategies called *slogans* concentrate entire ideas in catchy sentences, reminiscent of proverbs that seem to convey an absolute truth.

Self-evidence and causal arguments: Politicians often use *simple causal reasoning* to persuade the audience. They often build causal arguments upon values and beliefs that are deeply rooted in the minds of the majority of the audience. In order to enhance the causal arguments, politicians often present such values as being *self-evident assumptions*, i.e. known beforehand and accepted by everyone.

Analogies and detailed descriptions: To increase the veracity of their speeches, politicians often use *analogies with the past*, making reference to important historical characters or events. Finally, politicians include detailed descriptions and narrations to enhance the veracity of their speeches.

4 Dendrogram prototypical discourses

4.1 Founding principles

According to Harris (1954) and Rubenstein and Goodenough (1965), words in natural languages are structured within linguistic environments (e.g. sentences, paragraphs), and in this context, words having similar meanings tend to share similar contexts. This assumption, known as the *distributional hypothesis*, suggests that a corpus is often constituted by several discursive contexts, each one being a set of extended linguistic environments, conveying similar/related concepts and topics. Although this theory emerged in linguistics as early as 1954, it has recently received an increasing attention in many other fields such as in cognitive sciences (e.g. McDonald and Ramscar (2001)) and natural language processing (e.g. Mikolov et al. (2013a)). This hypothesis is the founding principle of our approach.

Our method aims at modeling a large corpus as a set of so-called DP-discourses and then studying them as prototypical speeches. To do so, the key step consists in building clusters of words sharing similar discursive contexts. This was achieved using word-embedding and subspace clustering, but other data-mining techniques could be used. Then, intra-cluster words were represented as *dendrogram prototypical discourses (DP-discourses)*, using a hierarchical clustering algorithm. Finally, DP-discourses have been studied using Charaudeau's methodology; they could possibly be analysed using other discourse analysis approaches.

4.2 Vector Space Modeling

This step aims at representing the corpus' different words as real-valued numeric vectors, building a Vector Space Model. For this purpose we used Word2Vec, a well-known word embedding algorithm based on neural networks, developed by Mikolov et al. (2013a). Word2Vec Vector Space Models are able to capture the contextual and the

semantic relationship between words, such that words appearing in the same context tend to have similar vector representations. Word2Vec has two major architecture: skip-gram and CBOW. According to Mikolov et al. (2013a), skip-gram outputs better representations for small datasets, while both architectures provide similar results for large datasets; regarding runtimes, CBOW tends to be faster than skip-gram. Since in this work we are dealing with a large dataset, we decided to use the CBOW architecture, with a negative sampling technique, as detailed hereafter.

Formal definition Let a textual corpus be a list $(w^{(1)}, w^{(2)}, \ldots, w(n))$, and let V_W denote its vocabulary, such that $\forall i \in \{1, \ldots, n\}, \quad w^{(i)} \in V_W$. The frequency of a word $w \in V_W$ is simply the number of times it appears in the corpus. The context $c^{(i)}$ of a word $w^{(i)}$ is defined as the list of neighboring words, within in a window of size l: $c^{(i)} = (w^{(i-l)}, \ldots, w^{(i-1)}, w^{(i+1)}, \ldots, w^{(i+l)})$. The set of contexts in the corpus is called V_C. Each word $w \in V_W$, and each context $c \in V_C$ are represented respectively by vectors $x \in \mathbb{R}^D$ and $z \in \mathbb{R}^D$, where D is the dimensionality of the Vector Space Model. A word-context pair $\langle w, c \rangle$ exists in the corpus, if c is the context of w, at least once. The probability that the pair $\langle w, c \rangle$ exists in the corpus is denoted $P(exists|w, c)$, and the probability of the complementary event is simply $P(\overline{exists}|w, c) = 1 - P(exists|w, c)$. In Word2Vec, this probability distribution is approximated using the corresponding vector representations x and z, as follows: $P(exists|w, c) \approx p(x, z) = 1/(1 + e^{-x^\top z})$. Word2Vec relies on a negative sampling technique: its neural network is trained to learn the vector representations that allow to discriminate a world w from a set of k randomly drawn words denoted $\tilde{w}$, only using its context c. This is achieved by maximizing $log(P(exists|w, c)) + k \times \mathbb{E}(log(P(\overline{exists}|\tilde{w}, c)))$. Indeed, this expression aims at maximizing $P(exists|w, c)$ for existing pairs, while minimizing $P(exists|\tilde{w}, c)$ for unexisting ones (i.e. maximizing $P(\overline{exists}|\tilde{w}, c)$). In terms of word and context representations, the corresponding objective function formalization is: $\mathcal{L}(x, z) = log(1/(1 + e^{-x^\top z})) + k \times \mathbb{E}(log(1 - 1/(1 + e^{\tilde{x}^\top z})))$. To maximize $\mathcal{L}(x, z)$, the algorithm updates the word and context representations x and z for each example, using stochastic gradient descent.

Parameter setting The Word2Vec implementation available in the Gensim Python library (Rehurek and Sojka, 2011) was used with the following parameter setting: The Vector Space Model dimensionality was set to $D = 300$, the context window size to $l = 5$, the number of negative samples to $k = 5$, and the number of iterations across the entire corpus to $NbIter = 5$.

4.3 Subspace clustering

Once the Vector Space Model was built, we clustered its word vector representations. The aim of this step was to find groups of words sharing the same discursive context, to analyze them separately. This task could have been achieved using any traditional clustering technique (Jain et al., 1999); however Aggarwal et al. (2001) have shown that traditional data mining algorithms struggle in high dimensional spaces, such as the 300 dimensional Vector Space Model generated by Word2Vec. To overcome this problem, an alternative is to use subspace clustering. This data mining task is recognized as being more general than clustering, since it does not only search groups of similar objects but also detects the subspaces where similarities appear. In this work, we used the recent subspace clustering algorithm called SubCMedians, designed by

Peignier et al. (2018). This technique, based on a K-medians paradigm, groups data instances around centers, and updates the coordinates and the subspaces of the centers, to minimize the distance to their closest data objects, using stochastic hill climbing. The clustering procedure can be stated more formally as follows:

Formal definition Let X denote the set of vector representations, such that $x \in \mathbb{R}^D$ represents word $w \in V_W$, and D denotes the Vector Space Model dimensionality. Let $\mathcal{M}$ denote the set of centers built by SubCMedians, such that $m_i \in \mathcal{M}$ is defined in its own subspace $\mathcal{D}_i$. Let $dist(x, m_i)$ be the distance between x and m_i; in SubCMedians, $dist(x, m_i)$ corresponds to the Segmental Manhattan distance (Aggarwal et al., 1999), an extension of the Manhattan distance, that allows to deal with vectors defined in different subspaces. Each vector $x \in X$ is assigned to its closest center $m_i \in \mathcal{M}$, and the corresponding distance between them is termed the Absolute Error $AE(s, \mathcal{M}) = min_{m_i \in \mathcal{M}} dist(x, m_i)$. The objective of SubCMedians is to find the centers $\mathcal{M}$ that minimize the Sum of Absolute Errors $SAE(X, \mathcal{M}) = \sum_{x \in X} AE(x, \mathcal{M})$. Once a suitable set of centers $\mathcal{M}$ has been produced, each vector $x \in X$ is assigned to its closest center, together with its corresponding word w; such that $\mathcal{C}_i = \{\langle w, x \rangle, \ldots\}$ denotes the cluster of word-representation pairs associated to center $m_i \in \mathcal{M}$. This assignment step defines directly a clustering model of words and vector representations.

Parameter setting Peignier et al. (2018) provided a simple and effective default parameter setting procedure for SubCMedians: the user simply provides a suggested number of clusters $NbExpClust$, and the algorithm automatically adapts the number of clusters and the sizes of subspaces. In this work, we set this parameter to $NbExpClust = 2$, which turned out to be sufficient to build a satisfactory subspace clustering model, as shown in Section 6.1.

4.4 Dendrogram-based intra-cluster words representation

The pairs of words-representations belonging to each subspace cluster were organized in dendrograms, using the traditional bottom-up hierarchical clustering algorithm developed by Sokal (1958). The aim of this step was to provide an interpretable structure of the intra-cluster words, to study them using Charaudeau's methodology. Other visualization techniques could also be used for this purpose[1].

Formal definition Let $\mathcal{C} = \{\langle w, x \rangle, \ldots\}$ be a cluster of word-representation pairs. Initially, each pair $\langle w, x \rangle \in \mathcal{C}$ is considered as an isolated group $\{\langle w, x \rangle\}$. Then, at each iteration, the algorithm merges the two closest groups, considering distances in the Vector Space Model. Iteratively the number of groups decreases, until every pair belongs to the same cluster. Then the algorithm stops, and the hierarchical arrangement of groups produced by the algorithm is represented as a dendrogram. Since the construction of the hierarchical structures is based on the distances between vector representations, the corresponding dendrograms represent the contextual and semantic relationship between intra-cluster words.

Parameter setting The hierarchical clustering algorithm has two major meta-parameters. The first one is the distance used to compare individual vector rep-

[1] For instance, t-SNE plots (Maaten and Hinton, 2008) seem an interesting alternative to dendrograms, and should be tested in future works.

resentations. In this work, two vector representations ${}^{u}x$ and ${}^{v}x$, were compared using their Manhattan distance $||{}^{u}x, {}^{v}x||_1$. The second meta-parameter corresponds to the so-called linkage method, which is used to assess the similarity between two groups u and v, before merging them. Here we used the Complete linkage method, which considers the maximum distance between elements of each group, as a similarity measure: $distance(u, v) = max(\{||{}^{u}x, {}^{v}x||_1 : \langle {}^{u}w, {}^{u}x\rangle \in u, \langle {}^{v}w, {}^{v}x\rangle \in v\})$.

In this work, we tested nine meta-parameter configurations, combining three classic similarity measures: the Manhattan distance, the Cosine similarity, and the Euclidean distance, and three well-known linkage methods: Average, Complete, and Single linkage. In practice, for each one of the nine possible configurations, we used the hierarchical clustering package from SciPy Python library (Jones et al., 2019) to build the dendrograms and partition their main branches. While analyzing these structures, a common practice consists in studying first each branch independently and then combining the different interpretations to get an overall understanding. In this context, it is preferable to deal with dendrograms such that: 1) the words within each branch are densely packed together, forming groups with low intra-cluster dispersion; 2) the branches are well separated from each other, forming a partition with a high inter-cluster dispersion. This characteristic of a clustering structure is captured by a well-known clustering quality measure: the Variance Ratio Criterion (Caliński and Harabasz, 1974). This measure is simply the average ratio between the inter-cluster and the intra-cluster dispersion, and higher scores are obtained by more interpretable dendrograms. In order to choose the most suitable configuration, we have computed the average Variance Ratio Criterion of the dendrograms, obtained using each one of the nine configurations. According to the results, depicted in Table 1, among these nine configurations, the combination of Manhattan distance and Complete Linkage method, obtained the higher Variance Ratio Criterion, and hence this setting was chosen as the one providing the most interpretable dendrograms.

	Manhattan	Euclidean	Cosine
Complete Linkage	**10.52**	9.70	9.10
Average Linkage	5.09	5.02	6.98
Single Linkage	4.23	4.13	3.37

Table 1: Average Variance Ratio Criterion for nine combinations of similarity measures (columns) and aggregation techniques (rows).

5 Dataset

Before focusing on the discourse analysis, it is important to describe the corpus itself. In this section we discuss the textual origin of the speeches, the data collection procedure, as well as basic cleaning and pre-processing steps that were applied to the corpus.

5.1 Textual origin

The characterization of the speeches' textual origin and composition mechanism raised two important questions: 1) Did Castro write and/or prepare his speeches himself or did a specialized team prepared them for him? 2) Were the speeches improvised or were they prepared beforehand?

To the best of our knowledge, only one interview carried out by Ramonet (2010), has addressed these questions. According to this interview, Castro himself affirmed that

some of his speeches were carefully written, while others were more or less improvised on the spot. In this context, Castro also highlighted the differences between the two kinds of speeches and he affirmed that written speeches may diminish the ability of the speaker to modulate his tone and tend to be less emphatic than improvised ones. Consequently, a reasonable expectation is that our dataset contains both kinds of speeches.

In the same interview, Castro declared that he had never been satisfied by speeches prepared by his collaborators and always ended up preparing his own speeches. Nevertheless, even if we imagine the speeches to have been prepared by a dedicated team of collaborators, it is reasonable to expect that the speaker would have read and validated the speeches and their underlying discursive strategies.

This work does not aim at classifying and analyzing the speeches according to their putative textual origin, which would be an interesting research subject on its own. Instead we have approached the analysis of Castro's speeches in a broad sense, regardless of their degree of preparation and textual origin.

5.2 Data collection

The corpus that has been analyzed in this work has been downloaded from a dedicated official website of the Cuban government.[2] This website hosts a large collection of documents produced by Fidel Castro, including speeches, interviews, essays and letters. Moreover, this website also hosts the translations of these documents into different languages. In this work, we decided to focus specifically on speeches in Spanish, filtering out interviews, letters, essays, and translated texts. This step aimed at avoiding possible biases by preventing the inclusion of strategies belonging to other kinds of enunciation, delivered through other channels of communication. In practice, data collection was ensured by Python web-scraping custom scripts, relying on the urllib2 (Van Rossum and Drake, 1995) and the BeautifulSoup (Richardson, 2019) Python libraries, while data filtering and cleaning were facilitated by Python custom scripts using the re (Van Rossum and Drake, 1995) and nltk (Bird et al., 2009) Python libraries. Finally, the corpus underwent a comprehensive manual verification.

5.3 Pre-processing

Once the corpus had been downloaded and cleaned, we applied stop-words filtering, a pre-processing step that excluded from the study extremely common words conveying too little semantic information. Even if stop-words filtering is a very common pre-processing step in natural language processing, it can be delicate to use this pre-processing in the context of word embedding. On the one hand, Mikolov et al. (2013b) have shown that a related procedure that aims at massively subsampling very frequent words can reduce the runtimes and improve the word vector representations of less frequent words. On the other hand, Agarwal and Yu (2009) have shown that removing stop words that are actually carrying semantic information (i.e. words linked to some specific contexts or negations) may lead to a significant quality drop of the word representations. In order to reduce runtimes and improve the representations of less frequent words while avoiding issues related to the exclusion of actually meaningful stop words, we carefully chose a curated stop words list with only 34 demonstrative adjectives, indefinite and definite articles. Stop words filtering was facilitated by Python custom scripts using the nltk (Bird et al., 2009) library.

[2] http://www.cuba.cu/gobierno/discursos/

In total, the corpus contains 1,018 discourses in Spanish, 7,548,480 word occurrences, 4,161,729 not-stop words occurrences and a vocabulary of 6,453 distinct not-stop words.

6 Results

For the sake of reproducibility and completeness, the Word2Vec vector representations, the clustering memberships, and all the DP-discourses are available on a dedicated web-page[3] and the software is available on a GitLab repository.[4]

6.1 Quantitative cluster assessment

Clustering structure Using the aforementioned methodology, the SubCMedians algorithm automatically extracted 26 clusters, thus adapting its number of clusters to the dataset without being restrained by the weak number of expected clusters parameter setting (here $NbExpeClust = 2$). For each cluster, we computed its absolute and relative vocabulary size, as well as the absolute and relative number of word occurrences. As shown in Figure 1, the 26 clusters have different sizes, and the four largest clusters gather close to 60% of the number of words and almost 50% of the corpus vocabulary. Hence, considering only these four clusters seemed sufficient to perform a representative discourse analysis study.

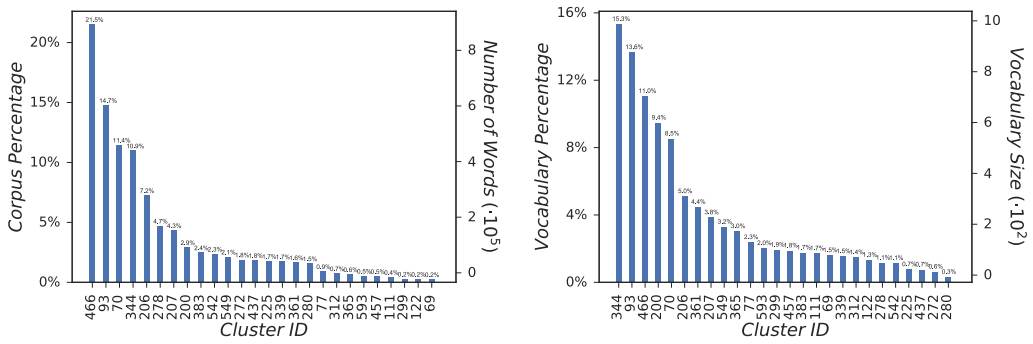

Figure 1: Absolute and relative number of word occurrences (right) and vocabulary size (left) per cluster.

Motivation for clustering assessment This work relies on the assumption that clusters represent discursive contexts, which allows their corresponding DP-discourses to be studied as prototypical speeches. According to the definition of discursive contexts presented in Section 4.1, consecutive words are likely to belong to the same discursive context. So, if clusters capture discursive contexts, consecutive words should also tend to belong to the same cluster, i.e. observing the same cluster membership for consecutive words should be mutually dependent events. Before studying the underlying discursive strategies embedded in the corresponding DP-discourses, we assessed whether our clusters exhibit this statistical property. To do so, we relied on the Weighted Point-wise Mutual Information measure, between the cluster memberships of consecutive words across all speeches, as detailed hereafter.

3 https://analisisdiscursosocialistalatinoamerica.github.io/castro.

4 https://gitlab.com/speignier/dendrogram_prototypical_discourses

Weighted Point-wise Mutual Information The Point-wise Mutual Information (PMI), as defined by Church and Hanks (1990), is a measure that quantifies the mutual dependence between two outcomes y and z, of two random variables Y and Z. The PMI is the logarithm of the ratio between the joint probability $p(Y = y, Z = z)$ and the distribution assuming independence $p(Y = y) \times p(Z = z)$. The variant called Weighted Point-wise Mutual Information (WPMI) simply weights the PMI by the joint probability: $WPMI(y, z) = p(Y = y, Z = z) \times \log\left(\frac{p(Y=y,Z=z)}{p(Y=y)p(Z=z)}\right)$. The WPMI between independent events is equal to zero this measure is positive for mutually dependent events, and it is negative for events that mutually exclude each other.

Intra and inter-cluster WPMI Let $C^{(t)}$ and $C^{(t+1)}$ be two discrete random variables, which model the cluster membership of two consecutive words from the corpus. Each random variable has K possible outcomes, denoted $\{c_1, c_2, \ldots, c_K\}$, and each outcome corresponds to one of the existing clusters ids (here $K = 26$). The $WPMI$ between the cluster memberships c_i and c_j of consecutive words is defined as follows:

$$WPMI(c_i, c_j) = p(C^{(t)} = c_i, C^{(t+1)} = c_j) \times \log\left(\frac{p(C^{(t)} = c_i, C^{(t+1)} = c_j)}{p(C^{(t)} = c_i) \times p(C^{(t+1)} = c_j)}\right)$$

Moreover, let $IntraClustWPMI$ and $InterClustWPMI$ denote respectively the sets of intra-cluster and inter-cluster WPMI values. More precisely, $IntraClustWPMI$ and $InterClustWPMI$ are simply the sets of WPMI values of consecutive words having the same ($IntraClustWPMI = \{WPMI(c_i, c_i), \ldots\}$) or different cluster memberships ($InterClustWPMI = \{WPMI(c_i, c_{j\neq i}), \ldots\}$). Let the sums of elements in $IntraClustPMI$ and $InterClustPMI$ be denoted as $IntraClustMI$ and $InterClustMI$ respectively. These values correspond to the intra-cluster and the inter-cluster Mutual Information measures.

Probabilities estimation Let $\#(c_i, c_j)$ be the frequency of consecutive non-empty words belonging to cluster c_i and c_j. In practice, the joint probabilities $p(C^{(t)} = c_i, C^{(t+1)} = c_j)$ can be estimated by $\frac{\#(c_i,c_j)}{\sum_i \sum_j \#(c_i,c_j)}$, and the marginal probabilities $p(C^{(t)} = c_i)$ and $p(C^{(t+1)} = c_i)$ can be estimated by $\frac{\sum_j \#(c_i,c_j)}{\sum_i \sum_j \#(c_i,c_j)}$ and $\frac{\sum_i \#(c_i,c_j)}{\sum_i \sum_j \#(c_i,c_j)}$ respectively.

Intra and inter-cluster WPMI comparison The $WPMI$ values were estimated using the previous method, and then the results were organized in a matrix. The rows and the columns of the matrix represent the possible outcomes of the random variables $C^{(t)}$ and $C^{(t+1)}$, respectively, and they are labeled accordingly, so $WPMI(c_i, c_j)$ is located in the row with label c_i, and column with label c_j. As depicted in Figure 2, the highest $WPMI$ values from this matrix correspond to the intra-cluster $WPMI$ for clusters 70, 93, 344 and 466, suggesting that intra-cluster $WPMI$ measures are higher than inter-cluster ones. Moreover, the intra-cluster Mutual Information $IntraClustMI = 0.102$ is revealed to be higher than the inter-cluster $InterClustMI = -0.031$. In order to determine whether $IntraClustWPMI$ and $InterClustWPMI$ follow the same distributions, we applied the non-parametric Mann–Whitney U test. This test resulted in a Mann–Whitney U statistic equal to 14,211 and a very low p-value equal to 1.8×10^{-09}. Consequently, intra-cluster WPMI are significantly higher than inter-cluster ones,

which supports the hypothesis that clusters represent discursive contexts, allowing us to proceed to the discourse analysis of the corresponding DP-discourses.

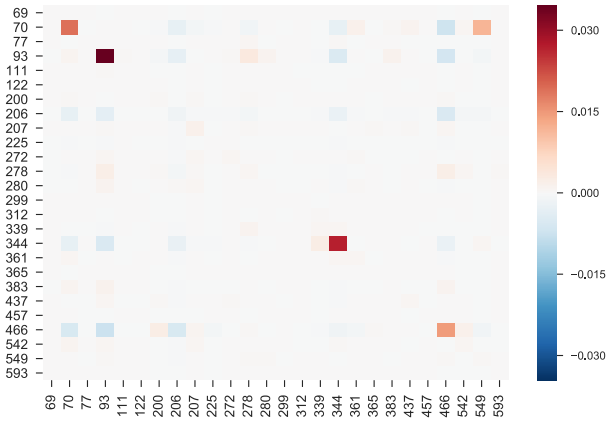

Figure 2: Matrix representing the $WPMI$ measures between the cluster memberships of consecutive words. The value $WPMI(c_i, c_j)$ is located along row labeled c_i and along column with label c_j. A strong mutual dependence between events is depicted in red, while mutually exclusive events are depicted in blue.

6.2 Discourse analysis

Regarding the previous results, we focused on the four biggest clusters, with highest $WPMI$ values, i.e., clusters 466, 344, 93 and 70. As explained in Section 4.4, the corresponding intra-cluster words were organized in DP-discourses, and analyzed using Charaudeau's methodology.

Castro's marxist revolutionary project (DP-discourse 466) This DP-discourse has a vocabulary of 713 words and 895,982 occurrences. The percentage of vocabulary size and the number of occurrences expressed in this DP-discourse are equal to 11.05% and 21.53% respectively. Analyzing this DP-discourse, we can infer that Fidel Castro's speeches often depict a revolutionary Marxist project for the future. To captivate his public, Castro characterizes his plan as being strongly tied to moral values, nationalism, social welfare and people. In this context, Castro summons his audience to join the project and also to defend it. To do so, he mainly uses recruitment processes and so-called 'rhetoric of effect' strategies. A DP-discourse representing the 100 most frequent terms from cluster 466 is illustrated in Figure 3.

Plan for the future: This DP-discourse is characterized by the presence of common nouns, and verbs in future and conditional, that evoke a forthcoming project and its near fulfillment (e.g. 'futuro' [future], 'adelante' [forward], 'será' [will be]). Different terms show that Castro depicts his plan as a difficult task (e.g. 'dificultades' [difficulties]), but nonetheless achievable (e.g. 'esfuerzos' [efforts]). Moreover, there are verbs in subjunctive and conditional, suggesting that this project is presented as a major aspiration of Castro and his public. Different verbs of obligation (e.g. 'seamos' [we must be]), in the first and third person, suggest that Castro presents the plan as an obligation. Therefore, Castro instructs Cuban people to join his project, and in doing so, he exhibits an Ethos of authority and commitment to this task. Moreover, since the notion of a future plan tends to convey an underlying idea of progress, Castro captivates his audience using a recruitment process strategy. Finally, by evoking the

construction of a better future, Castro seeks to provoke in his audience strong feelings, such as hope and pride. This approach corresponds to the rhetoric of effects.

Marxist revolution: Castro's project is actually the consolidation of a Marxist popular revolution. Indeed, this DP-discourse is characterized by a large and rich lexicon reflecting a revolutionary Marxist imaginary (e.g. 'revolucionarios' [revolutionaries], 'socialismo' [socialism]), which shows the importance that Castro assigns to this topic. This strategy corresponds to a process of recruitment based on ideological belonging. Close to this lexical group, there are several nouns, evoking abstract social welfare topics, nationalism and mainly moral values (e.g. 'justicia' [justice], 'moral' [moral], 'patria' [homeland]). Hence, Castro includes, in the description of his plan, moral precepts that are strongly rooted in Cuban and Latin American society in general. This corresponds to a process of recruitment based on moral values. In this context Castro speaks in the name of socialist and Marxist ideals themselves, employing the *Me-Spokesman* discursive identity. Interestingly, some terms transcribing the favorable reaction of the crowd, are connected to these terms (e.g. 'aplausos' [applauses]), and illustrate the strength of such discursive strategies.

Defense of the project: This DP-discourse contains many verbs in infinitive with an imperative value (e.g. 'pelear' [to fight], 'defender' [to defend]), which call upon the audience to join and protect the plan, providing Castro with an Ethos of authority and power. In addition, we find some belligerent nouns (e.g. 'frente' [front], 'firme' [firm]) as well as combinations of words with a strong semantic impact, depicting disastrous scenarios (e.g. 'miseria' [misery], 'pobreza' [poverty], 'sangre' [blood]). These elements correspond to a discursive strategy of rhetoric of effects, since they seek to make the audience feel outraged and fearful, and to trigger a defensive or aggressive reaction in the crowd. According to Charaudeau, these types of strategies are common in political discourse, since an audience immersed in these feelings, would accept a message more easily. Using these elements, Castro summons the audience using the traditional triadic scenario: Marxist revolution is leading Cuban society towards a better future; however, there is an enemy that threatens the people and their future; so Castro calls on the audience to join and defend the project. This approach also corresponds to a process of recruitment, based on the defense of social, political, moral and national ideals.

Identification: The presence of the first person plural in verbs and possessive adjectives (e.g. 'nosotros' [we], 'nuestro' [our]) shows that Fidel Castro uses the *Me-Us* discursive identity. This strategy has a two-fold goal: It allows Castro to create a higher degree of identification with his audience, and it makes people perceive themselves as major actors in Castro's plan. Since Castro presents himself as a committed ruler, identified with the people, and connected to the beneficiaries of his policies, he exhibits an Ethos of authority, identification and charisma.

Veracity: In this DP-discourse we have identified different terms showing that Castro aimed at increasing the veracity of his discourse. First, there are several words denoting negation (e.g. 'nunca' [never], 'ninguna' [none]) that are connected to terms denoting the concept of absoluteness. Since these terms are also strongly connected to the term 'duda' [doubt] we deduced that Castro asked his audience to believe in him beyond any possible doubt. Hence, in order to convey the absolute veracity of his speeches, Castro exhibits an Ethos of power and authority. Moreover, in this context, there are also different terms denoting the first person singular (e.g. 'digo' [I say], 'creo' [I believe]). This suggests that Castro exhibits the *Me* discursive identity, expressing a personal commitment, and presenting himself as the guarantor of the veracity of his speech.

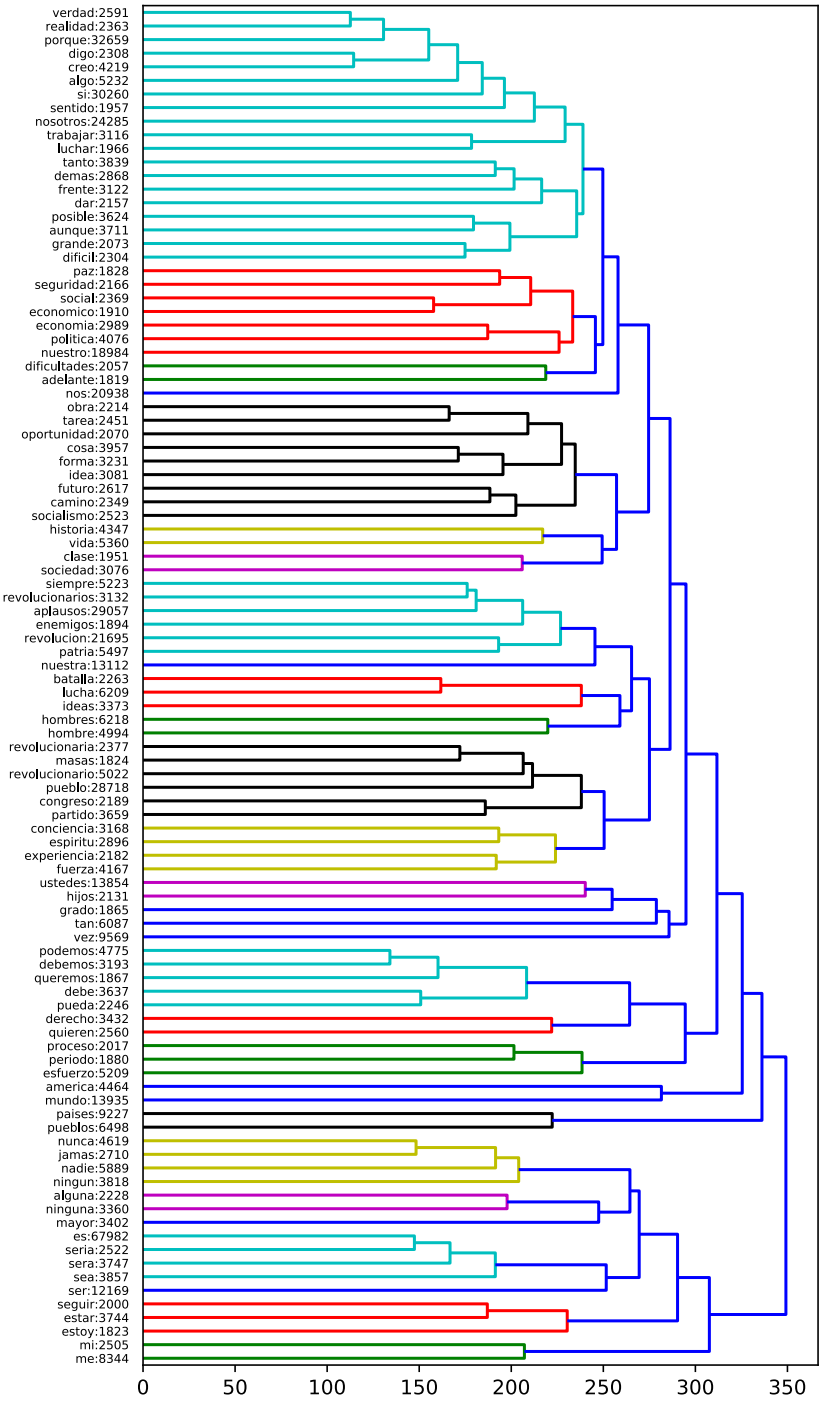

Figure 3: DP-discourse 466 containing the 100 most frequent terms.

Finally, there are several logical operators (e.g. 'porque' [because], 'aunque' [although], 'si' [if], 'tampoco' [nor]). Since these terms participate in argumentative processes by definition, we deduced that Castro aimed to increase the credibility and truthfulness of his speeches by using arguments belonging to the Logos family (possibly relying on simplified augmentation and self-evident assumptions).

Cuba's economic development (DP-discourse 93) This DP-discourse has a vocabulary of 881 words and 614,923 occurrences. The percentage of vocabulary size and the number of occurrences expressed in this DP-discourse are equal to 13.65% and 14.78% respectively. Analyzing this DP-discourse, we can infer that Castro's speeches often evoke Cuba's economic development. This concept is conveyed by several nouns from the lexical field of industrial and agricultural production, as well as verbs of necessity, obligation and action. In this context, we can also deduce that Castro mainly uses a rhetoric of effects and aims at increasing the legitimacy and the veracity of his message through detailed descriptions. A DP-discourse representing the 100 most frequent terms from cluster 93 is illustrated in Figure 4.

Needs, obligations, actions: This DP-discourse is characterized by the presence of several nouns and verbs in the present and future tense denoting needs and possibility (e.g. 'necesitan' [they need], 'pueden' [they can], 'tienen' [they have to]). Since these terms are linked to words referring to industrial and agricultural projects (e.g. 'industria' [industry], 'ganadería' [livestock], 'caña' [sugar cane]), we infer that Castro shows that these plans respond to the country's fundamental needs and open new opportunities for Cuba. This discursive strategy corresponds to a recruitment process, based on ideals of progress. Moreover, since this strategy also aims at creating in the audience feelings of hope and national pride, it also corresponds to a rhetoric of effects. We find several verbs from the lexical field of realization, in present and future tense (e.g. 'tendremos' [we will have], 'alcanza' [it reaches]). Fidel Castro is likely to have used these verbs to show that the projects were being executed and would be completed soon. Moreover, there are verbs of realization in subjunctive, which show that these plans were an important aspiration for Castro. There are also several terms from the lexical field of time that were probably used by Castro to refer to the execution schedule of the projects (e.g. 'diarias' [daily], 'mensuales' [monthly]). Castro provides these details in order to increase the veracity of his message and to show himself as an expert who masters all the details related to the execution of the projects. Moreover, using these mechanisms, Fidel Castro increases his own legitimacy, embodying an Ethos of commitment. Finally, the different verbs appearing in this DP-discourse are mainly conjugated in the first person plural and in the third person plural and singular (e.g. 'disponemos' [we dispose], 'crecen' [they grow]). The presence of the first person plural indicates that Fidel Castro uses the *Me-Us* discursive identity, showing his closeness and identification with the Cuban people.

Wealth and economic prosperity: This DP-discourse is characterized by the presence of many terms referring to the development of agricultural and industrial means of production (e.g. 'tractores' [tractors], 'fábrica' [factory]). Moreover, there are terms from the lexical field of wealth and production in general (e.g. 'bienes' [goods], 'riquezas' [resources], 'producto' [product]). We also find verbs expressing actions related to economy and production. And we mainly identified words corresponding to different economic resources and products, such as raw materials (e.g. 'caña' [sugar cane], 'petroleo' [oil]). These words are strongly linked to several terms conveying the concepts of quantity, quality and value (e.g. 'dólares' [dollars], 'toneladas' [tones], 'millones' [millions]). This suggests that Castro describes these projects in detail in order to increase the veracity and credibility of his speeches and also to show himself an expert on the subject. Moreover, this rich vocabulary creates an impression of wealth and economic prosperity. This strategy aims at generating in his audience a feeling of hope, well-being, security and pride, and thus corresponds to the rhetoric-of-effects strategy. On the other hand, since this strategy is based on the imaginary of progress

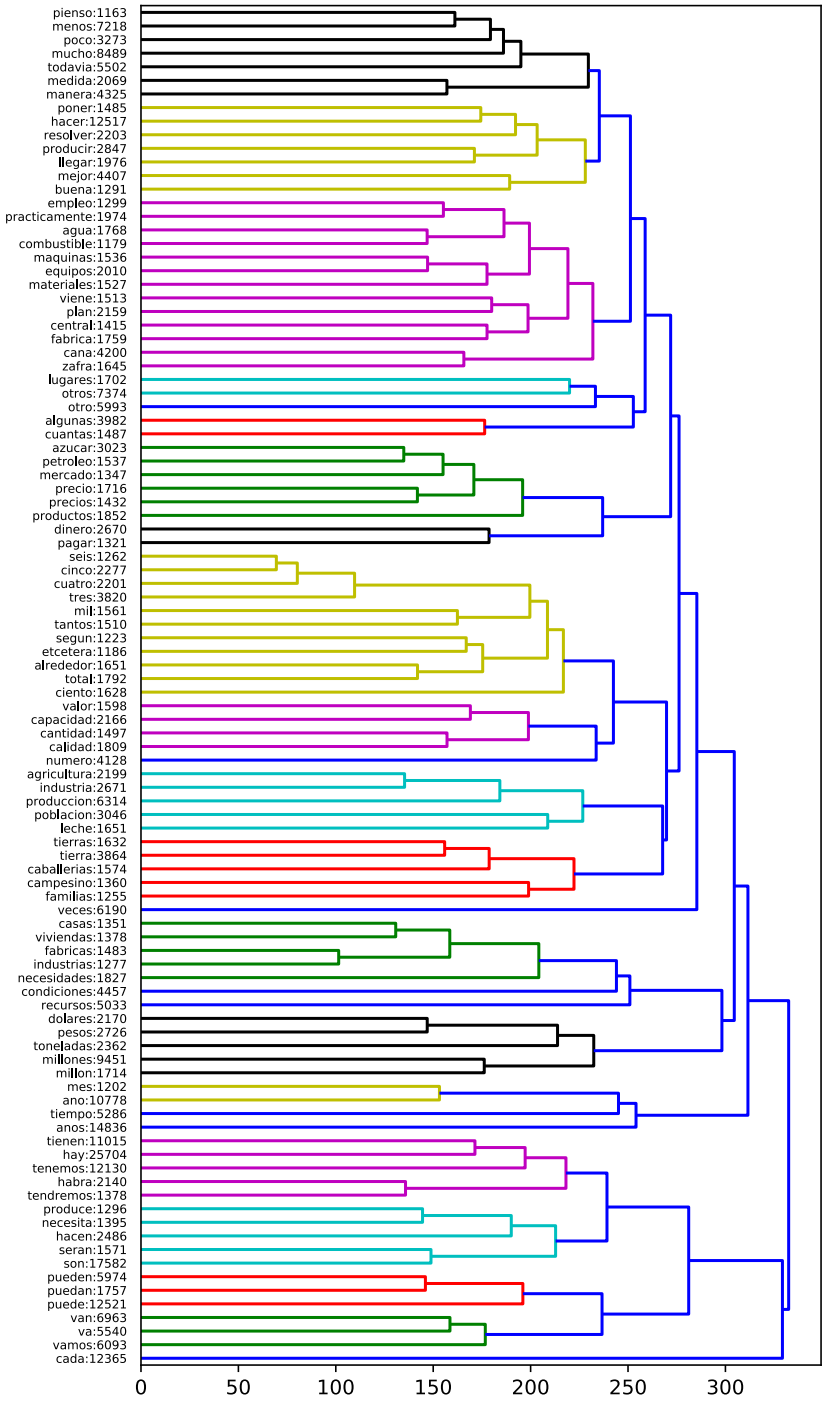

Figure 4: DP-discourse 93 containing the 100 most frequent terms.

and economic growth, it also corresponds to a recruitment process.

Social infrastructure This DP-discourse also contains several nouns that refer to the development of social infrastructure and services, such as public housing (e.g. 'casas' [houses], 'comida' [food], 'tienda' [store]). We also find nouns that evoke the beneficiaries of these social projects (e.g. 'familias' [families], 'pobres' [poor], 'campesino' [peasant]). This strategy corresponds to a recruitment process, based on the social

welfare imaginary. Moreover, since this mechanism aims at generating a feeling of hope in the audience, it also corresponds to a rhetoric of effects.

Social welfare projects (DP-discourse 70) This DP-discourse has a vocabulary of 550 words and 476,957 occurrences. The percentage of vocabulary size and the number of occurrences expressed in this DP-discourse are equal to 8.52% and 11.46% respectively. In this DP-discourse, Fidel Castro presents the work that his government and himself are carrying out. These ongoing projects address social welfare issues and are mainly related to education, health and employment. A DP-discourse representing the 100 most frequent terms from cluster 70 is illustrated in Figure 5.

Ongoing projects: This DP-discourse contains lexicon referring to the development of projects (e.g. 'construyendo' [building], 'proyectos' [projects]). There are also verbs in the present tense conveying the idea of ongoing actions (e.g. 'creando' [creating], 'convirtiendo' [transforming], 'resolviendo' [solving]). These verbs are likely to increase the veracity of Castro's message by referring to tangible, ongoing steps. The aforementioned terms are linked to adjectives and adverbs that describe these steps using a range of positive connotations such as importance, variety, novelty and quantity (e.g. 'importantes' [important], 'enorme' [huge], 'mejores' [best], 'diversos' [diverse]). These elements aim at captivating the audience by creating feelings of hope and pride; they thus correspond to a rhetoric of effects. Furthermore, Castro evokes the places where such projects are being executed (e.g. 'ciudad' [city], 'región' [region], 'Camaguey', 'Matanzas', 'Cienfuegos'). These geographical details increase the veracity and the credibility of the speeches and, in turn, indicate that the entire country benefits from the projects. This corresponds to a recruitment process by national belonging.

Society as an actor: Among the terms referring to ongoing projects, there are several infinitive verbs (e.g. 'realizar' [to make], 'participar' [to participate]), and words referring to people (e.g. 'obreros' [worker], 'trabajadores' [workers], 'sindicatos' [unions]). Since verbs in the infinitive form may have a value of imperative, we can infer that Castro calls on the audience to participate in these projects, as individuals or as members of social organizations. Therefore, Castro shows himself as the organizer, and the head of the projects, exhibiting an Ethos of authority and commitment. In addition, Castro uses possessive adjectives in the first person plural (e.g. 'nuestros' [ours]). This shows that Castro uses the *Me-Us* discursive identity in order to include society in the development of the projects. This also indicates that Castro develops an identification process between him and the Cuban people.

Education, employment and health: Specific vocabulary suggests that the projects described by Castro embrace three major social welfare objectives: education (e.g. 'estudiantes' [students]), employment (e.g. 'empleos' [employment]) and health (e.g. 'salud' [health]). These elements are major social welfare topics and they directly imply a better quality of life for the population, which is shown as the direct beneficiary of the projects. The objective of this strategy is to generate feelings of hope and well-being in the audience; this corresponds to a rhetoric-of-effects strategy. In addition, the direct reference to ideals of social welfare induce a recruitment process, since society is likely to share such ideals. Finally, the large amount of specific lexicon in this DP-discourse contributes to the veracity of the speeches and allows Castro to present himself as an expert.

Cold War and triadic scenario (DP-discourse 344) This DP-discourse has a vocabulary of 989 words and 457,130 occurrences. The percentage of vocabulary size and

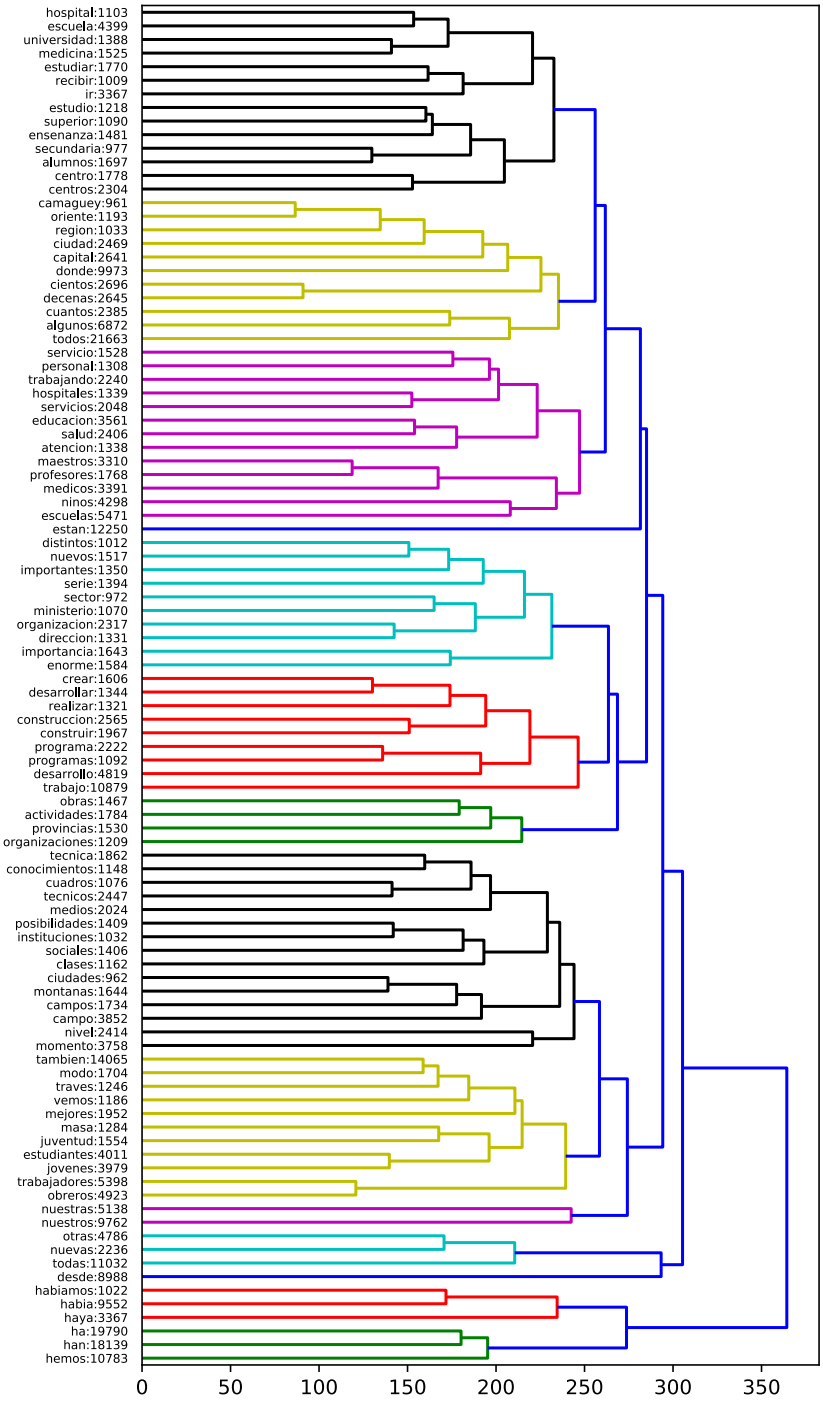

Figure 5: DP-discourse 70 containing the 100 most frequent terms.

the number of occurrences expressed in this DP-discourse are equal to 15.33% and 10.98% respectively. Here, Castro reports different historical conflicts that occur in Cuba and around the world, mainly in the context of the Cold War. These events exhibit a particularly disastrous situation and correspond to the first element of the triadic scenario. In this context, Castro characterizes the main enemies of the Cuban people: the United States and local elites. This description corresponds to the second element of the triadic scenario. Finally, the third element corresponds to the socialist Bloc, communist parties, people and the army of Cuba. A DP-discourse representing the 100 most frequent terms from cluster 344 is illustrated in Figure 6.

Historical conflicts and wars: This DP-discourse evokes different historical conflicts that took place in the context of the Cold War in Cuba and around the world: the dictatorial government of Fulgencio Batista overthrown by the Cuban revolution (e.g. 'Batista', 'dictadura', 'Moncada'); guerrilla movements (e.g. 'guerrilla'); the attempted invasion of Playa 'Girón' (Bay of Pigs Invasion) by the United States; the economic 'embargo' promoted by the United States against Cuba; the civil war in 'Angola'; the 'Vietnam' War; Latin American dictatorships ('Trujillo' [Dominican Republic], 'Somoza' [Nicaragua], 'dictador' [dictator]); the wars of decolonization (e.g. 'Argelia', 'Guinea'). These elements are connected to time markers (e.g. 'julio' [July], 'abril' [April]), and verbs in the past tense (e.g. 'hubo' [there was], 'hicieron' [they did]), showing that Castro includes analogies with the past in his speeches. This allows him to increase the legitimacy and the veracity of his discourse and depict himself as an expert on history and international politics. Furthermore, the description of the historical conjuncture is strongly impregnated by military lexicon (e.g. 'infantería' [infantry], 'artillería' [artillery]), and moral and national values ('soberanía' [sovereignty], 'democracia' [democracy], 'libertades' [freedoms]). References to war and ideals create feelings of fear, heroism, and pride in the audience, which corresponds to the rhetoric of effects. This strategy also corresponds to a recruitment process based on nationalism and moral values. Finally, the details conveyed by the military vocabulary aim at increasing the veracity of the discourse.

Cold War and triadic scenario: Castro's description of historical events is defined by the Cold War and its focus on conflicts between the socialist and the western Blocs. These Blocs constitute the binary conception of the world depicted by Castro and they are organized following the triadic scenario, where the western and the socialist Blocs correspond to the *Enemy* and the *Hero* respectively. This binary description is a simplification of a complex scenario. This presentation allows the audience to concentrate on a few ideas and thus corresponds to the singularization argument. This mechanism is often complemented by the argument of essentialization, which condenses complex ideas into a few simple terms. Hence, we deduce that some terms (e.g. 'imperio' [empire], 'burgués' [bourgeois]) may correspond to this mechanism. The combination of singularization and essentialization, correspond to the use of argumentative formulas that generate a strong feeling of evidence and attraction.

Enemy: This DP-discourse is characterized by vocabulary evoking the existence of an external enemy, which is depicted as the opponent to liberation movements around the world. The enemy is clearly associated with the United States, using different nouns (e.g. United States, 'imperio' [empire], 'Yankis' [Yankees]). Castro also describes an internal enemy, represented by the economically dominant class (e.g. 'burgués' [bourgeois]), and by the political elites (e.g. 'oligarquía' [oligarchy], 'politiqueros' [demagogues]), which receive orders from the external enemy (e.g. 'esbirros' [henchmen], 'títeres' [puppets]). The speaker uses a Marxist conception of society

to link the economic elites to the concept of labour exploitation (e.g. 'explotadores' [exploiting]). On the other hand, Castro associates the local elites, and especially the external enemy, with crimes and destruction (e.g. 'genocida' [genocidal], 'destrucción' [destruction]); these terms create an extremely negative image of the enemy. These elements contribute to depict the enemy as numerous and dangerous and create the feeling in the audience that there is a hidden threat. In this context, Castro implicitly depicts a latently disastrous situation. The goal of this rhetoric of effects is to generate fear and anxiety in the audience, in order to predispose people to accept his message.

Hero: The hero described by Fidel Castro has several facets, which he most probably adapted according to the audience and to the political context. First, Castro assigns a major role to the Cuban revolutionary army and rebel groups: different nouns present this institution as the vanguard in the fight against the enemy ('guerrilleros' [guerrilla groups], 'milicianos' [militiamen], 'cubanos' [Cubans]). Moreover, there is a large lexicon evoking civil society and socialist political movements related to Castro's government (e.g. 'socialista' [socialist], 'comunista' [communist], 'gente' [people]). Therefore, we deduce that Fidel Castro creates an amalgam between communist organizations, the people and the army of Cuba, producing a greater sense of identification between and among these actors. Furthermore, the description of these actors is associated to values and ideals, using nouns and adjective with strong semantic impact (e.g. 'heroismo' [heroism], 'sacrificio' [sacrifice]). This strategy aims at creating feelings and passions in the audience and thus corresponds to a rhetoric of effects. In addition, by using values and ideals as banners in the fight against the enemy, Fidel Castro also applies a recruitment process, based on moral, social, national and ideological belonging. Finally, Castro evokes different important historical characters, such as José Martí, the Cuban writer, politician and hero of the Cuban independence war and the well-known guerrilla commander Ernesto Che Guevara. Castro evokes José Martí, to create nationalist feelings in his audience and also to increase his legitimacy by presenting himself as the successor of this historical figure, while Che Guevara is invoked as an heroic example of a fighter against the enemy. Castro thus creates a revolutionary pantheon of heroes.

Summary In order to have an overview of Castro's discursive strategies, we decided to count the number of times each strategy was found in the DP-discourses. To do so, we conducted an exhaustive manual expert-driven analysis of the discursive strategies present in the branches of the four most important DP-discourses. For each discursive strategy that was detected in a branch, twenty random sentences from the corpus, containing the involved words, were extracted and systematically checked in order to confirm our interpretations. According to this study, Pathos, Ethos and Logos make up 51%, 34% and 15% of these strategies respectively, which is coherent with general characteristics of political speeches as described by Charaudeau (2005). The strategies' frequencies are illustrated in Figure 7. According to this radar-chart, the most frequently used Pathos strategies are rhetoric of effects, recruitment process based on social ideals, nationalism and progress, and the triadic scenario. The most frequently used Ethos strategies are the Me-Us discursive identity, the Ethos of authority, commitment, expertise and the identification with the audience. Finally, the veracity strategy seems to be the most frequently used of Logos strategies.

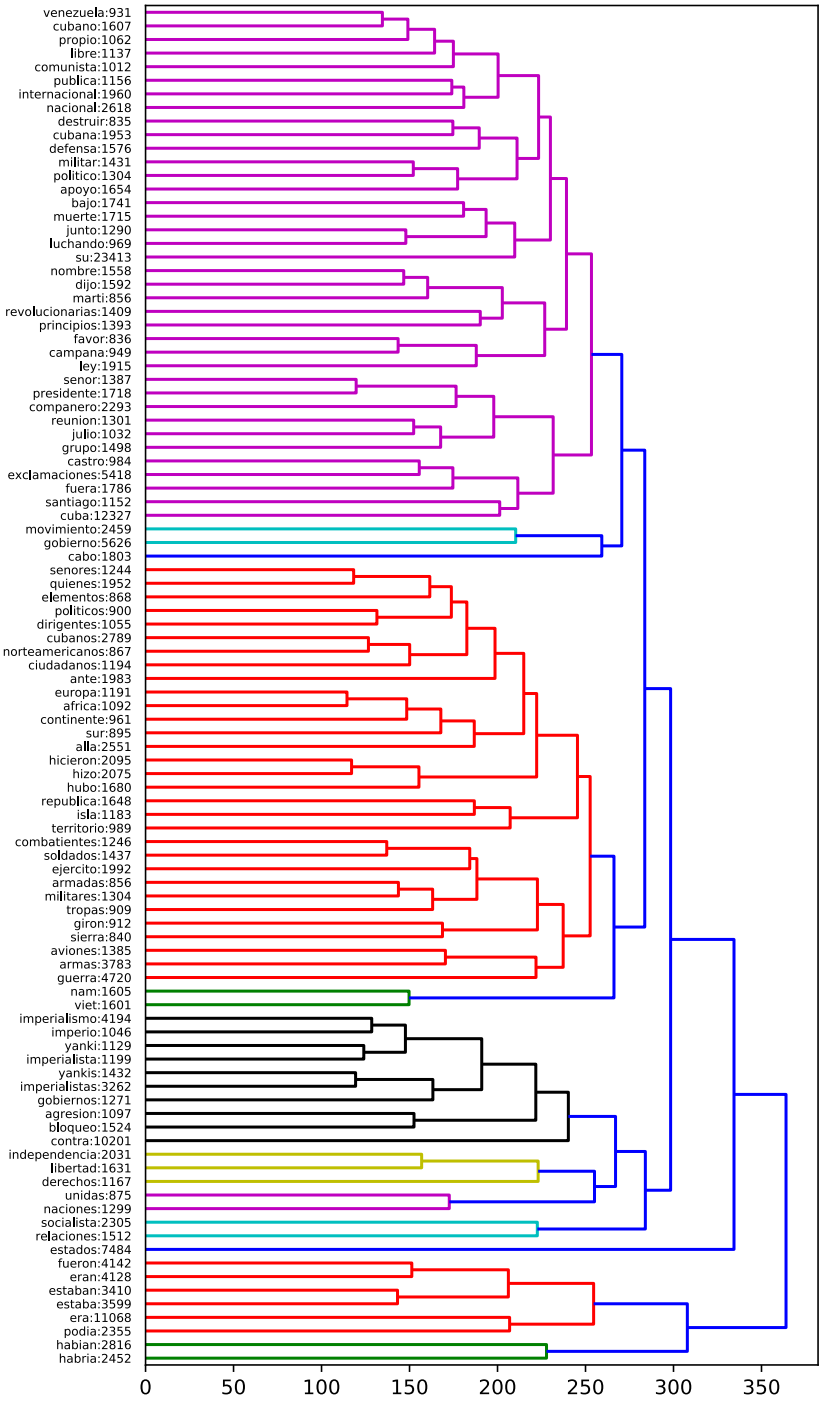

Figure 6: DP-discourse 344 containing the 100 most frequent terms.

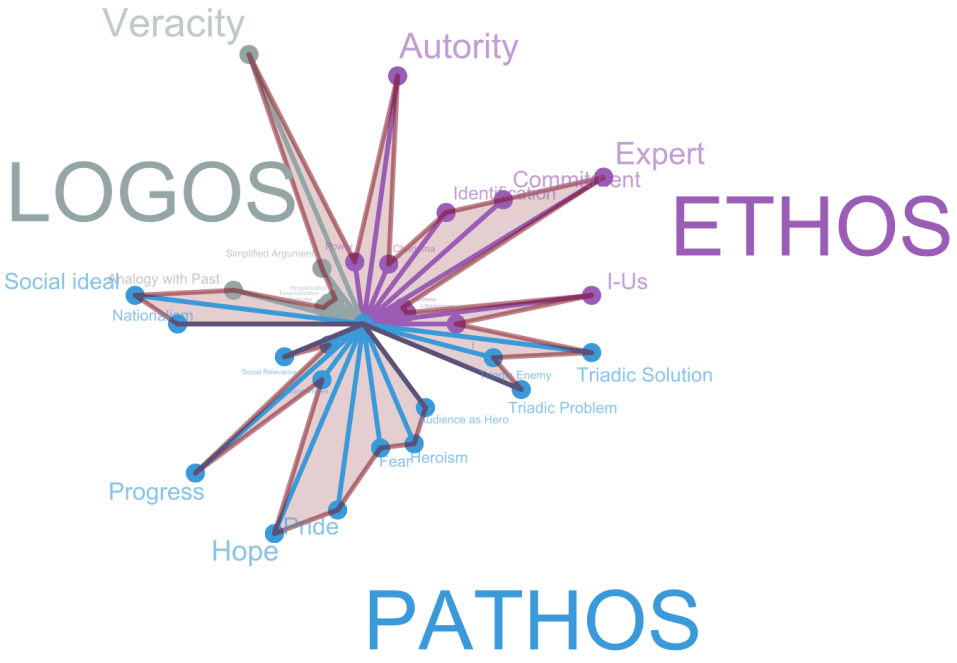

Figure 7: Radar-chart summarizing the discursive strategies of Fidel Castro. Each radius represents one strategy, and its length is proportional to its frequency across the full analysis. The Pathos, Ethos and Logos strategies are respectively colored in blue, violet and green.

Reference	Methodology	Discursive strategies	Number of strategies
Joyner (1964)	Aristotelian rhetoric	People-hero, people-victim, enemy, moral feelings: anger, fear, confidence, hope committed, expert, identification causal, analogy, induction, deduction	12
Fagen (1965)	Non-lexicometric	Castro-hero, people-victim, enemy authority, identification, charisma	6
Nieto et al. (2002)	Conversational analysis	Charisma, identification	2
Belisario (2010)	Cognitive linguistics Critical disc. analysis	People-hero, people-victim, enemy authority, identification	5
Reyes (2011)	Sociolinguistics Critical disc. analysis	Witness, expert, charisma, identification veracity	5
De Sousa (2009a) De Sousa (2012) De Sousa (2009b)	Lexicometric	People-hero, people-victim, enemy moral, ideology identification	6
DP-discourses	Hybrid: Data mining - Semio-pragmatic	People-hero, Castro-hero, people-victim enemy, nationalism, ideology progress, welfare, moral feelings: anger, fear, confidence, hope committed, witness, expert, authority charisma, identification causal, analogy, veracity	**19**

Table 2: Discursive strategies reported previously in the literature, and in this work using DP-discourses (last entry). For the sake of clarity, strategies have been coloured according to their category, following the Aristotelian classification of rhetoric. Ethos, Pathos and Logos are reported in violet, blue and green, respectively.

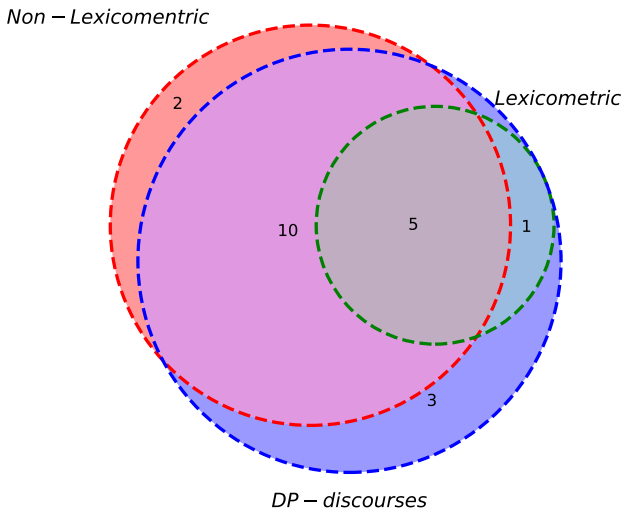

Figure 8: Venn diagram representing the number of discursive strategies found by lexicometric approaches (green circle), non-lexicometric approaches (red circle) and those retrieved using DP-discourses (blue circle).

7 Discussion

In this work, we hypothesized that the use of a data mining driven methodology could provide a more representative characterization of Castro's discursive strategies. In order to assess this hypothesis, and to illustrate the contribution of the paper to the existing literature, we compared our conclusions with those of previous works. The discursive strategies reported so far in the literature are summarized in Table 2, and the number of discursive strategies found by general families of methods (i.e. lexicometric, non-lexicometric and based on DP-discourses) are represented using a Venn diagram in Figure 8.

As shown in Table 2 most of the previous works, focused on particular aspects of Castro's discursive strategies: Nieto et al. (2002) and Reyes (2011) mostly studied the discursive identity of Castro, while Belisario (2010) and Fagen (1965) analyzed Castro's triadic scenario and his discursive identity. De Sousa (2009a,b, 2012) focused on the evolution of the major topics developed in the speeches, with respect to the historical scenario. Finally, Joyner (1964) was the only previous work that aimed at presenting a general picture of Castro's discursive strategies, using a non-lexicometric methodology derived from Aristotelian rhetoric.

As depicted in Table 2 and Figure 8, this work reports the highest number of discursive strategies, i.e. 19 in total, including 16 out of 18 strategies that were described by at least one of the eight previous works.

The two missing strategies, namely inductive and deductive reasoning, have only been reported by Joyner (1964). These strategies belong to the Logos family, and they are usually carried by complex articulations of words, which are likely to be lost by Vector Space Modeling algorithms. Indeed, Vector Space Models capture the semantic relationship between words, by assigning similar vector representations to words appearing in the same context; as a result, fairly rare and complex associations of words are likely to be lost at the expense of more frequent ones. In this case, specialized approaches, such as a systematic expert-driven analysis, should be applied to retrieve

these strategies.

On the other hand, three important discursive strategies that were found in this work have been neglected in the literature, possibly due to small corpus selection biases. Indeed, previous studies have not focused on the importance of recruitment processes based on ideals of progress, social welfare and nationalism. Nevertheless, as presented in Section 6.2, these discursive strategies have been widely used by Castro, with the objective to lead his audience to accept his project willingly.

Consequently our approach was able to draw a representative landscape of Castro's discursive strategies that agrees with the major conclusions of previous works, while also offering new insights into this research question.

8 Conclusion

Summary This paper presents a two-fold contribution to the digital humanities community: On the one hand, we propose a new discourse analysis data mining framework to study large data copus. This new approach combines state-of-the-art data mining tools with the well-known semio-pragmatic linguistic discourse analysis methodology On the other hand, we have provided a broad and representative characterization of the main discursive strategies used by Castro. This study was conducted on a large corpus of more than 1,018 speeches and 7,500,000 words, combining state-of-the-art data mining tools and the semio-pragmatic discourse analysis methodology. According to this study, Castro presents himself as an authority, an expert committed to his duties and identified with his audience. His speeches are organized around a Cold War triadic scenario, his government and socialist movements worldwide are presented as heroes that protect people against the source of all problems, i.e. the enemy which is represented by the USA and the bourgeoisie. In this context, he shows the audience as a potential hero and beneficiary, and he alludes to the progress made by his country. These elements evoke strong feelings such as heroism, pride, hope and fear. Finally, Castro tends to include many details to increase the veracity of his speeches. A comparison between these findings and those of previous works reveals that our method confirmed most of the previously reported discursive strategies and provided new insight into the importance of recruitment processes based on nationalism, welfare and progress. These discursive strategies aim at captivating the audience by incorporating references to classic positive values such as social welfare policies, ideals of technological progress, economic growth and nationalist ideals. Thus, our study provides a broad and representative characterization of the main discursive strategies used by Finally Castro.

Wider relevance Given that our methodology resulted in a representative characterization of Castro's rhetoric strategies, it seems promising to apply it to other case studies, and to use it to analyze political rhetoric in a broader context. For instance, considering the resurgence of populism in Europe and America (Berezin, 2009, De la Torre, 2010, Greven, 2016), it could be particularly interesting to characterize the rhetorical strategies of populist leaders. Indeed, according to Jansen (2011), populist movements are mainly based on two components: popular mobilization and populist rhetoric. Characterizing, and analyzing, the discursive strategies of various populist leaders is therefore an important step towards understanding this phenomenon. In this context, a hybrid discourse analysis methodology, as presented in this paper, could provide a representative and broad overview of the populist rhetoric and make it possible to

unravel common populist discursive strategies. Furthermore, this kind of framework could facilitate the elaboration of fast, large-scale and more objective analyses of political discourses at critical moments (i.e. elections, political crisis), which could have a direct impact on society. Interestingly, this research topic has recently motivated the creation of global research projects such as the Populismus Observatory, [5] and the Team Populism. [6] Both projects aim at developing research communities to promote the study of populist rhetoric by sharing data, information and tools. In this context, the framework presented in this paper could be integrated into such projects as a complementary analysis tool.

Perspectives and Future Work In the future, we plan into incorporate the temporal dimension to the discourse analysis. The corpus of Fidel Castro's speeches is particularly well adapted to these kinds of temporal considerations, since the date of each speech is also reported. Simply by counting the number of words belonging to each DP-discourse at different periods of time and by following the frequency changes, we could study the evolution of Castro's discursive strategy over time and with respect to the speeches' historical context. Another promising research path consists in using our data mining framework to characterize the discursive strategies of other politicians, and possibly using other non-lexicometric discourse analysis methodologies. Future perspectives also include a thorough assessment of this method and a comparison with alternative approaches that could rely on different pre-processing steps (e.g. Brants (2000)), different word embedding techniques (e.g. Le and Mikolov (2014)), alternative subspace clustering methods (e.g. Kriegel et al. (2009)), topic modelling algorithms (e.g. Blei et al. (2003), Jain et al. (1999)) and complementary visualization techniques (e.g. Maaten and Hinton (2008)).

References

Agarwal, S. and H. Yu
2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Aggarwal, C. C., A. Hinneburg, and D. A. Keim
2001. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, Pp. 420–434. Springer.

Aggarwal, C. C., J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park
1999. Fast algorithms for projected clustering. In *ACM SIGMoD Record*, volume 28, Pp. 61–72. ACM.

Barry, A.
2002. Les bases théoriques en analyse du discours. *Documents de la Chaire MCD*, 159.

Belisario, A. G. V.
2010. Sistemas metafóricos en discursos de Fidel Castro: "decir la verdad en el primer deber de todo revolucionario". *Letras*, (81):139–162.

Benzécri, J.-P. et al.
1973. *L'analyse des données*, volume 2. Dunod Paris.

5 http://observatory.populismus.gr/
6 https://populism.byu.edu

Berezin, M.
2009. *Illiberal politics in neoliberal times: culture, security and populism in the new Europe*. Cambridge University Press.

Bird, S., E. Klein, and E. Loper
2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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

Brants, T.
2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, Pp. 224–231. Association for Computational Linguistics.

Caliński, T. and J. Harabasz
1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Charaudeau, P.
1995. Une analyse sémiolinguistique du discours. *Langages*, Pp. 96–111.

Charaudeau, P.
2005. Quand l'argumentation n'est que visée persuasive. l'exemple du discours politique. *Argumentation et communication dans les médias. Québec: Éditions Nota Bene*, Pp. 23–43.

Charaudeau, P.
2008. Pathos et discours politique. *Émotions et discours. L'usage des passions dans la langue. Rennes: Presses universitaires de Rennes*, Pp. 49–58.

Charaudeau, P.
2009a. Identité sociale et identité discursive. un jeu de miroir fondateur de l'activité langagière. *Identités sociales et discursives du sujet parlant*, Pp. 15–18.

Charaudeau, P.
2009b. Le discours de manipulation entre persuasion et influence sociale. In *Acte du colloque de Lyon*.

Charaudeau, P.
2011. Réflexions pour l'analyse du discours populiste. *Mots. Les langages du politique*, (97):101–116.

Church, K. W. and P. Hanks
1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

De la Torre, C.
2010. *Populist Seduction in Latin America*. Ohio University Press.

De Sousa, S.
2009a. Le discours de Fidel Castro. Éssai de lexicométrie politique. *Lexicometrica, Explorations textométriques*, 2:68–94.

De Sousa, S.
2009b. Le peuple dans le discours Fidel Castro. *Communication au Colloque Représentation du Peuple*, Pp. 1–14.

De Sousa, S.
2012. À l'épreuve des temps... temps lexical et temps politique dans le discours de Fidel Castro (1959-2008). *A. Dister, D. Longré et G. Purnelle (Éds), JADT*, Pp. 337–349.

Ekman, P.
1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ellis, P.
2010. The essential guide to effect sizes: An introduction to statistical power. *Meta-Analysis and the Interpretation of Research Results.: Cambridge University Press*.

Fagen, R. R.
1965. Charismatic authority and the leadership of Fidel Castro. *Western Political Quarterly*, 18(2-1):275–284.

Fairclough, N.
2013. *Critical discourse analysis: The critical study of language*. Routledge.

Gee, J. P.
2014. *An introduction to discourse analysis: Theory and method*. Routledge.

Gott, R.
2007. *Cuba*. Ediciones Akal.

Greven, T.
2016. The rise of right-wing populism in europe and the united states. *A Comparative Perspective. Friedrich Ebert Foundation, Washington DC Office*.

Harris, Z. S.
1954. Distributional structure. *Word*, 10(2-3):146–162.

Hudson, R. A.
1996. *Sociolinguistics*. Cambridge university press.

Jain, A. K., M. N. Murty, and P. J. Flynn
1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Janis, I. L. and L. Mann
1977. *Decision making: A psychological analysis of conflict, choice, and commitment*. Free press.

Jansen, R. S.
2011. Populist mobilization: A new theoretical approach to populism. *Sociological theory*, 29(2):75–96.

Jones, E., T. Oliphant, P. Peterson, et al.
2019. SciPy: Open source scientific tools for Python.

Joyner, G. M.
1964. *Persuasive elements in the speeches of Fidel Castro*. PhD thesis, Texas Tech University.

Kriegel, H.-P., P. Kröger, and A. Zimek
2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1.

Le, Q. and T. Mikolov
2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, Pp. 1188–1196.

Lövheim, H.
2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses*, 78(2):341–348.

Luong, X. and S. Mellet
2003. Mesures de distance grammaticale entre les textes. *Corpus*, (2).

Maaten, L. v. d. and G. Hinton
2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Maingueneau, D.
2016. *Les termes clés de l'analyse du discours*. Le seuil.

McDonald, S. and M. Ramscar
2001. Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Mikolov, T., K. Chen, G. Corrado, and J. Dean
2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean
2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, Pp. 3111–3119.

Nieto, M. J. et al.
2002. La afectividad en la comunicación política. *Opción: Revista de Ciencias Humanas y Sociales*, (39):36–53.

Padmos, R. et al.
2017. Fidel and raúl castro's ideological influence on foreign policy in reaction to the us trade embargo. B.S. thesis.

Pêcheux, M.
1995. *Automatic discourse analysis*, volume 5. Rodopi.

Peignier, S., C. Rigotti, A. Rossi, and G. Beslon
2018. Weight-based search to find clusters around medians in subspaces. In *Proceedings of the ACM Symposium on Applied Computing*. ACM.

Plutchik, R.
2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Ramonet, I.
2010. *Fidel Castro: biografía a dos voces*. Debate.

Rehurek, R. and P. Sojka
2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Reyes, A.
2011. *Voice in political discourse: Castro, Chavez, Bush and their strategic use of language*. A&C Black.

Richards, J. C. and R. W. Schmidt
1983. Conversational analysis. *Language and communication*, Pp. 117–154.

Richardson, L.
2019. Beautiful soup.

Rubenstein, H. and J. B. Goodenough
1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Schwarz, N.
2000. Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

Sokal, R. R.
1958. A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, 28:1409–1438.

Turney, P. D. and P. Pantel
2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Van Rossum, G. and F. L. Drake
1995. *Python library reference*. Centrum voor Wiskunde en Informatica.

Weizman, E.
2008. *Positioning in media dialogue: Negotiating roles in the news interview*, volume 3. John Benjamins Publishing.

Widdowson, H. G.
1995. Discourse analysis: a critical view. *Language and literature*, 4(3):157–172.

Wiedemann, G.
2013. Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, Pp. 332–357.

Character Centrality in Present-Day Dutch Literary Fiction

Roel Smeets[1], Eric Sanders[2], and Antal van den Bosch[3]

[1]Radboud University, Department of Literary and Cultural Studies
[2]Radboud University, Centre for Language and Speech Technology
[3]KNAW Meertens Institute

In the critique of literary representation, the depiction of literary characters has been studied from ideological perspectives. Hierarchies can be exposed by determining the centrality of a character relative to other characters. As an addition to such close reading methods, the present paper proposes an approach to character centrality that combines network analysis with narratology. This explorative study is based on a dataset of demographic metadata on 2,137 characters from a corpus of 170 contemporary Dutch novels. We extract social networks of characters from each novel, and rank all characters according to five centrality metrics. Then, we perform a multiple linear regression to test which of the demographic variables predicts a character's position in the rankings. Our results suggest that immigrant and female characters score higher on two of the five centrality metrics. As a narratological evaluation, we contextualise this observed pattern in relation to a close reading of Özcan Akyol's *Eus* (2012), a novel from the corpus that thematises both descent and gender. We demonstrate that our data-driven and empirically informed approach to character centrality lays bare surprising patterns of representation which only gain relevance in light of close readings of specific cases.

Keywords: Dutch literature, character representation, social network analysis

1 Quantifying the Critique of Literary Representation

Literary studies have a rich tradition of critically analysing hierarchies in literary texts from an ideological perspective. In the wake of the poststructuralist turn (J.Culler (1983)), the critique of literary and cultural representation has evolved into an engaging field with roots in various ideological strands, such as Marxism, postcolonialism and feminism. These so-called 'hermeneutics of suspicion' (Ricoeur (1979), Felski (2009)) have focused on a variety of topics, including the hierarchical representation of story

characters, the inhabitants of fictional story worlds. In Dutch literary studies, this is illustrated by a range of studies that critique the representation of characters of a certain gender, descent or class (e.g. Pattynama (1994), Meijer (1996a), Meijer (1996b), Pattynama (1998), Minnaard (2010), Meijer (2011)). For the sake of the present research, it suffices to say that this field of study is concerned with hierarchies between characters and their identities, although this is of course not their only focus. These ideological approaches to character representation are commonly (and/or implicitly) concerned with how important, influential, dominant or central a character in a narrative is as opposed to other characters. 'Centrality' will be used in this article as an umbrella term to refer to abstract notions such as importance, dominance, influence and power. When a character is central, it means that he or she is important, dominant, influential or powerful in a specific way.

A number of quantitative studies on character representation in Dutch literature have been conducted only recently (van der Deijl et al. (2016), Van der Deijl and Smeets (2018), Koolen (2018): 162-244). These studies make use of new applications of social network analysis and other quantitative methods to reconstruct and analyse story worlds in a data-driven way, following earlier research on non-Dutch texts (e.g. Alberich et al. (2002), Stiller et al. (2003), Elson et al. (2010), Lee and Yeung (2012), Karsdorp et al. (2012), Agarwal et al. (2013), Jayannavar et al. (2015), Karsdorp et al. (2015b), Lee and Wong (2016)). For the study of hierarchies between characters, these methods provide the means for a formalisation and quantification of the concept of 'character centrality'. This is potentially interesting for the study of character representation as practised in the critique of literary representation. Except for these recent examples, studies on character representation mainly use close reading methods. This can lead to powerful interpretations for one or a few cases, but such qualitative readings do not result in general insights into the centrality of characters at a larger scale. However, the importance of a character in a narrative might very well be expressed by the numerical frequency with which he/she features in the narrative. While being fully aware that data-driven approaches are not ideologically neutral either, the present study aims to bridge this gap by considering the centrality of characters in both narratological and statistical terms.

In this contribution, we try to answer the following question: To what extent can a data-driven and empirically informed approach to character centrality contribute to the ideological critique of literary representation? First, we provide a succinct overview of how character centrality has been understood in the narratological tradition of analysing literary texts. Second, we confront these narratological considerations with a computational approach to identifying networks of characters. Third, we describe our data and the method we used to rank all 2,137 characters in a corpus of 170 contemporary Dutch novels on the basis of five statistical metrics. Fourth, we analyse and interpret the results of a multiple regression analysis that tested which demographic feature (gender, descent) was the best predictor for a character's place in the rankings. Based on simple descriptive statistics of the gender and descent distributions among characters in the corpus, we hypothesise that male and non-immigrant characters will score higher than female and immigrant characters. Fifth, in order to evaluate the output of the statistical model narratologically, we confront these findings with a close reading of one novel from the corpus. We conclude with the argument that our approach to character centrality lays bare surprising patterns of representation, although they only start to make sense when contextualised through a qualitative reading of a specific case.

2 Centrality in Narratology

Narratology is one of the traditional methodological toolkits for the study of narratives. This toolkit offers various instruments to analyse the centrality of characters in literature, of which we will mention two of the most straightforward. A character's position in the story world is already predetermined by some basic structural features of a literary text. The mode in which a novel is narrated is commonly a first indicator of how important a character is in the storyline. Some narrative layers in a text are embedded in others, which is particularly relevant for the position of narrating characters in first-person novels. As narrating characters belong to the highest narrative layer, they consequently 'produce text that is not perceived by the characters' that belong to more embedded narrative layers (Van Boven and Dorleijn (2013): 33).[1] As such, 'the narrating instance is located on a higher textual level' and is 'above the world of the characters' who do not have a narrating role (ibid.). In this sense, it is logical to ascribe a more central role to a narrating character in a first-person novel than to the other characters, as the narrator is the one who is in the best position to control the flow of information.

Focalisation is a narratological concept that is also applicable to the centrality of characters. It was coined by the French structuralist Gerard Genette to distinguish between who narrates and who perceives in a text (Genette (1972)). Others have suggested revisions of the concept (e.g. Bal (1977), Nelles (1990), Jahn (1996)); the revision that has become most popular is that of the Dutch scholar Mieke Bal. She defined focalisation as 'the relation between the vision and that which is "seen", perceived' (Bal (2009): 145-146), which made it possible to discern hierarchical relations between characters who occupy active focalising roles and characters who are mainly in a passive position in which they are being focalised by other characters. The extent to which a character features in active focalising roles is thus another indicator of his/her place in the character hierarchy.

3 Centrality in Network Theory

Network theory has been occupied with the question of how to measure the centrality of nodes in a network. Relational, networked structures are interesting in this respect as they can yield insights into the centrality of certain actors as opposed to others. The centrality of a node can be measured in a number of ways to consider different aspects of the network structure. In 1978, the American sociologist Linton Freeman observed that there is 'certainly no unanimity on exactly what centrality is or on its conceptual foundations, and there is very little agreement on the proper procedure for its measurement' (Freeman (1978): 217). He conceptualised three basic centrality measures – degree, betweenness and closeness – which are still being used today, albeit frequently in revised form, and which are thought to 'cover the intuitive range of the concept of centrality' (idem: 237). It is worth mentioning that Freeman's intent was not 'to "lock in" to any sort of ultimate centrality measure' (idem: 217), as centrality is a rather abstract concept and therefore hard to pinpoint statistically. Existing measures as those used by Freeman at best help to clarify what might be understood as central, but they do not necessarily give any definitive answers on which actors are most important in a network.

Before Freeman's innovative proposition, centrality was mainly viewed in terms of

[1] Unless otherwise indicated, all translations are the authors' own.

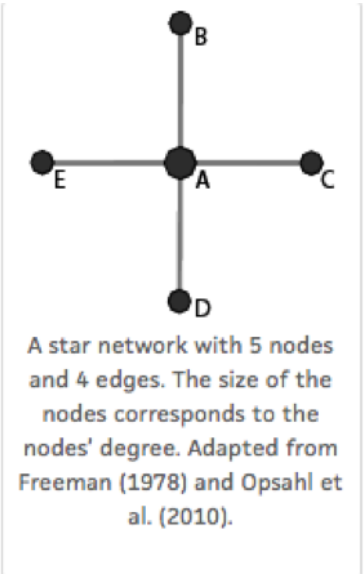

Figure 1: Adapted from Opsahl and Others (2010)).

degree, which is the most straightforward measure of centrality. In Figure 1, node A has an advantage over B, C, D and E because it has more relations to others in the network: A has a degree of 4, whereas B, C, D and E each have a degree of 1. The main limitation of *degree centrality*, however, is that it does not take into account the overall structure of the network. A node can be related to many other nodes but located in the periphery of the network, which results in a situation where the node is far removed from the opposite side of the network.

As an alternative to degree centrality, *closeness centrality* is defined as the sum of distances to all others nodes in the network. An advantage of closeness is that it takes into account the relative access that a node has to other nodes in the network. In Figure 1, node A has a higher closeness than B, C, D and E, as it is directly connected with its neighbours, whereas B, C, D, and E need to cross through A to reach a node other than A. The disadvantage of closeness centrality, however, is that it cannot properly be applied to networks that are not fully connected. By definition, nodes in two disconnected components of a network are unable to reach one another, and therefore closeness cannot be computed for the overall structure of a network with disconnected components.[2]

Freeman was the first to propose *betweenness centrality*, which computes the extent to which a node lies on the shortest path between two other nodes. In Figure 1, node A has a high betweenness centrality because it connects all four nodes with each other. As it is applicable to networks with disconnected components, betweenness has an advantage over closeness. However, as a metric, it is limited because nodes are often not located on the shortest path between two others nodes. Because of that, B, C, D and E, in Figure 1, all have a betweenness centrality of 0.

Network theory makes a distinction between unweighted and weighted graphs. In a weighted graph, the edges represent the intensity with which two nodes are connected. As the basic centrality measures of degree, closeness and betweenness are devised for application to unweighted, binary networks, alternative metrics have been

[2] In case a network has a lot of disconnected components, a convenient approach would be to only compute closeness centrality for the largest component of the network.

proposed. Degree centrality has been redefined for weighted graphs by not focusing on the number of relations but on the sum of the weights of those relations (Barrat et al. (2004)). Dijkstra's algorithm (Dijkstra (1959)), named after the Dutch computer scientist Edsger W. Dijkstra, has been used to redefine closeness and betweenness centrality by looking at the shortest paths in terms of distances (Newman (2001), Brandes (2001)). As these new proposed metrics target primarily the weights and are less reliant on the number of relations, a second redefinition was needed to take into account both weight and number of relations (Opsahl and Others (2010)).

Every network thus demands a specific approach; there is no general method that applies to every network. The first question should be which elements constitute the network, the second how those elements are related. Then, it should be decided if the network is binary and unweighted, or if the elements are gradually related to one another. The appropriate centrality measures should be derived from the specific nature of the network (weighted/unweighted, unipartite/bipartite[3]) and the question through which is it approached, as not every centrality measure is relevant in all possible instances.

In the following we will explore to what extent it is useful to make a synthesis between the narratological and the network analytic approach to centrality. We do so by considering literary texts as social networks made up of characters which can be ranked according to the centrality measures described above. The structure of these character networks will be adjusted to the mode of narration and focalisation of the novels. In the context of character representation, narratology thus informs a quantitative and statistical conceptualisation of centrality. Conversely, the study of social networks of fictional characters is informed by some basic narratological insights.

4 Data & Method

In order to test how a data-driven approach to character centrality might contribute to the critique of literary representation, we devised a computational model that takes into account both network theoretical and narratological considerations on mode of narration and focalisation. There is an emerging branch of studies that apply network analysis to fictional populations of characters (e.g. Alberich et al. (2002), Stiller et al. (2003), Elson et al. (2010), Lee and Yeung (2012), Agarwal et al. (2013), Jayannavar et al. (2015), Karsdorp et al. (2015b), Lee and Wong (2016), Moretti (2013):211-240,Rydberg-Cox (2011), MacCarron and Kenna (2012)), though all use different methods for their purposes. One of the main challenges for this type of analysis is the conceptual issue of how to define and automatically identify characters in texts (Dekker et al. (2019)). This can be done manually (e.g. Moretti (2013):211-240) or automatically (e.g. Elson et al. (2010)). Vala et al. (2015) have shown that automatic detection is a difficult task due to the poor performance of existing pronominal and coreference resolution techniques.[4] Because of this poor performance, we do not aim for full coreference resolution, but instead use a semi-automatic method that departs from a predefined

[3] Unipartite networks exclusively consist of elements from the same category, e.g. people connected to people. Bipartite networks consist of elements from different categories, e.g. people connected to organisations. The number of elements in multipartite networks can be extended endlessly in theory, but it is usually restricted to three different categories (tripartite).

[4] An alternative approach to character detection is automatically classifying animacy in in texts (Karsdorp et al. (2015a)).

set of characters. Inspired by narratological theories on what constitutes a character (Herman and Vervaeck (2005): 60-61; Van Boven and Dorleijn (2013): 335), we define characters as *people or creatures which to a greater or lesser extent are presented as human, existing of not more than a few linguistic features including one or more names.* For each novel, a list of names is created with Named Entity Recognition (NER); characters whose name frequency is above a normalised threshold value (based on the number of words of the text) will be regarded as characters.[5]

The other challenge is how to define and automatically identify relational ties between those characters. One of the most used definitions of character relation frames connections between characters in terms of conversations or dialogues (Stiller et al. (2003), Elson et al. (2010), Lee and Yeung (2012), Moretti (2013): 211-240, Jayannavar et al. (2015), Lee and Wong (2016)). The quantifiable unit of the conversation is, however, not the best indication for character interactions, as there are plenty of characters that do not enter into a conversation but are related to one another in some other way. For instance, two characters with family ties might never speak to each other, but such a relation should definitely be regarded as a character relation. Another way to define relational ties is in terms of co-occurrence in the same window of N words, sentences, paragraphs or chapters (Alberich et al. (2002), Grayson et al. (2016)). Defining character relations in terms of adjacency in the text will be able to capture more instances of character interaction than when it is defined in conversational terms. This is the most bottom-up definition of character relations, as characters do not have to communicate in a literal sense (as is the case in conversation networks) to be considered as having some form of interaction.

Based on these considerations, we operationalise the strength of character relations through *co-occurrences of character name variants in a window of N tokens*. We experimented with different window units and sizes for different types of novels to find the 'sweet spot' where not too many and not too few character interactions are detected (cf. Grayson et al. (2016)). However, such a sweet spot is different for every novel. In order to be able to compare the novels, we decided to use the same window unit and window size for every novel. As sentences are the smallest linguistic structures which are semantically meaningful in themselves (cf. Mann and Thompson (1988)), we used sentences as the window unit, which we tokenized using Ucto.[6] The window size was set to two sentences, as semantic relations are known to extend over two sentences through connectives (cf. Bluhdorn (2010)).[7] We devised a customised co-occurrence approach for each narrative mode, which we describe in detail below.

We have used a sample corpus of 170 contemporary Dutch novels, consisting of all submissions to the 2013 Libris Literatuur Prijs, one of the most prestigious literary prizes in the Dutch language area. This prize is awarded to novels published in the year before, in this case in 2012. In that year, 1,397 Dutch novels were published; our sample of 170 novels thus makes up 12.2 percent of the total number of novels

[5] There are several NER-tools available, but not all are suitable for the same task. NER-tools have to be trained for specific languages, and their accuracy depends on the nature of the training data (e.g. a tool trained on newspaper articles performs badly on literary fiction). For the current research the Namescape-tagger is used, which is trained on Dutch literary fiction and which is demonstrated to be the most accurate for the present purposes, although it is still not perfect as a F1-score of 0.72 was reported (Smeets (2017)).

[6] https://github.com/proycon/python-ucto, last accessed: 3-7-2018.

[7] As the plain texts of the novels in our corpus are unstructured, we could not rule out the possibility that characters co-occur in two sentence windows that transcend the boundaries of a paragraph or chapter. We are aware that this creates noise, as in those cases it could be argued that there is no meaningful interaction between characters.

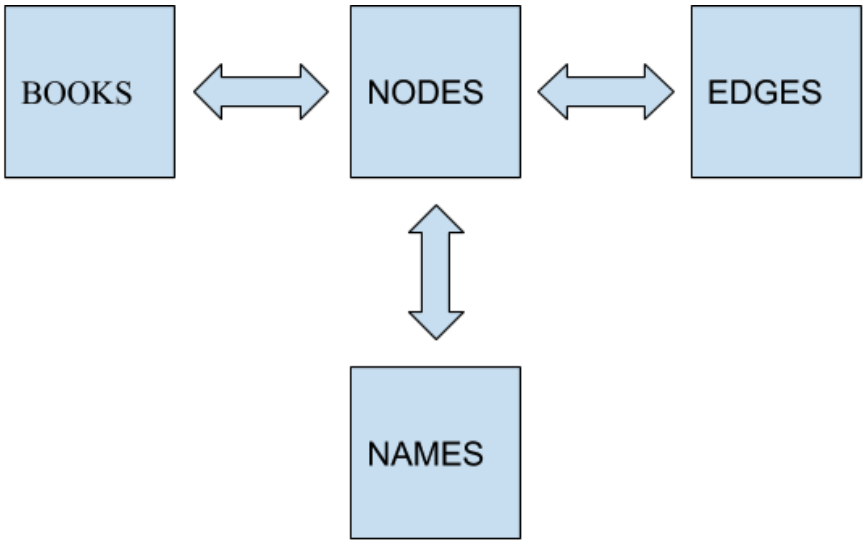

Figure 2: Visualisation of database linkage

published in that year.[8] With a few exceptions, these novels belong to the genre of literary fiction.[9]

In earlier research (van der Deijl et al. (2016)), the following demographic information for 1,176 characters in the corpus was manually gathered, if known: gender, age, country of descent, city of descent, country of residence, city of residence, profession. In more recent research (Volker and Smeets (2019)), the number of characters was increased to a total of 2,137 characters and semi-automatically enriched with 4,459 relational ties between each character. These relational ties were grouped in one of the following five categories: family, friend, lover, colleague, enemy.[10] These character metadata were stored in four interrelated database tables (see Figure 2).

Lists of all variants of a character's name were automatically generated with named-entity recognition and were stored in a table called NAMES. BOOKS contains all relevant metadata of the novels, such as title, the name, gender and age of the author, the publisher and the filename of the digital version of the novel. NODES contains all relevant metadata of the characters, such as name, gender, country of descent, city of descent, country of residency, city of residency, education and profession. EDGES contains all relevant metadata on the character relations, such as the specific nature of the relation (friend, family, enemy, lover, colleague). All tables are linked to one another through a unique book id. NAMES, NODES and EDGES are also connected through a character id. The character networks are computed through an Object-Oriented model written in the Python programming language, consisting of three main classes: Book,

[8] This number is based on all Dutch language novels published in 2012 with NUR-code 301 (literary fiction), that is, 1,780 in total. These include 383 duplicates or reissues, which were subtracted from the total number. Thus, the total number of 'original' Dutch literary fiction published in 2012 is 1,397.

[9] For a list of all 170 novels, see: http://www.librisliteratuurprijs.nl/2013-groslijst.

[10] We used top-down relational labels assigned to characters by two expert annotators. These annotators based their annotations on rather narrow definitions: e.g. the label 'enemy' was only assigned when the relation is clearly hostile, the label 'friend' when the relation is clearly friendly. Differences in annotations between the annotators were resolved through discussion. The annotators also accounted for changing relations between characters. In those cases double labels were assigned, such as Colleague_Enemy. Double labels were also assigned when the nature of the relation changed over time, such as friends becoming enemies.

Character and Network.[11]

Each book in the corpus has a unique id from 1 to 170. Every character in the corpus has a unique character id that corresponds to a book id stored in database BOOKS. For instance, *De lichtekooi van Loven* by Ineke van der Aa is represented by the book id 1. In database NODES, character 'Louise' is represented by this same book id followed by character id 1 and her name (1_1_Louise). In database NAMES, this same unique identifier is followed by every name variant of the character. The name variants for this character are 'Louise', 'Louisje' and 'Louiseke', which is represented in NAMES as 1_1_Louise_Louise, 1_1_Louise_Louisje and 1_1_Louise_Louiseke. Each novel's text was then searched for each of these name variants, after which these variants were replaced by the unique character identifier.[12] As such, the locations of each character in the text were automatically identified.

The corpus was divided into three sub-corpora based on their mode of narration: third-person, first-person, multi-perspective. Third-person novels are narrated by an anonymous narrator who follows one main character. First-person novels are narrated by an I-narrator. Multi-perspective novels are narrated by multiple narrators, either in third or first person. For every subcorpus a slightly different co-occurrence approach was used based on the specific mode of narration. For all novels, irrespective of their mode of narration, relations between characters were pre-established when they were annotated with one of the relational labels stored in EDGES (friend, family, lover, enemy, colleague). In all cases, the procedure below was used to establish the weight of the relations.[13]

Third-person novels (63 novels) For every character in the novel, a sliding window approach was used in which co-occurrences of two characters were mapped in a window of two sentences. Whenever two characters occur in the range of the same two sentences, a relation between those characters was established. The more often such co-occurrence takes place, the stronger their relation becomes.

First-person novels (73 novels) *a)* As the first-person narrator has by definition high centrality in narratological terms, the relations of the first-person narrator with all other characters were simply defined by counting every occurrence in the novel of characters other than the first-person narrator. As every character is embedded in the narration of the first-person narrator, it can be argued that every character occurrence represents a relational tie with the first-person narrator. The more often a character occurs in the novel, the stronger its relation with the first-person narrator.

b) For every character other than the first-person narrator, a sliding window approach was used in which co-occurrences of two characters were mapped in a window of two sentences. Whenever two characters occur in the range of the same two sentences, a relation between those characters was established.

---

[11] All software and data are accessible through the following open access GitHub repository: https://github.com/roelsmeets/character-networks. This repository does not contain the corpora because of copyright issues.

[12] A similar approach is used by Grayson et al. (2016): 4, who replace character aliases with a character's name.

[13] In some cases, two characters have a relational label such as "family" assigned to them while the weight of their relation is 0. This is possible as characters do not have to be adjacent in the text to have a family tie, just as people in real-world networks can be family without being in each other's physical presence or without talking about each other.

The more often such a co-occurrence takes place, the stronger their relation becomes.
Note that this approach will in most cases rightfully lead to relatively strong relations between the first-person narrator and all other characters, whereas this is not the case for the relations between and among all other characters.

Multi-perspective novels (34 novels) For each of these novels, student assistants annotated where a character perspective begins and ends in the text. These annotations also contain information on the narrative mode and focalisation: a first-person or third-person narration was annotated as such, and for third-person narration the main focaliser was annotated. On the basis of those annotations, each novel was divided into separate sections. For sections narrated in first or third person, the first- or third-person method was applied. After that, the co-occurrence counts between characters were aggregated for all the separate sections.

All these relations are symmetrical, and thus undirected. This means that the character relations are not regarded in terms of directionality, which is a logical consequence of the co-occurrence approach, as adjacency is a priori a symmetrical issue. Furthermore, the resulting network, with characters as nodes and character relations as edges, will both be undirected and weighted. Not every relation between any two characters will have the same status, as the strength of a relation is increased when two characters co-occur more often in the novel.

With Python's software package networkx,[14] the resulting networks for each individual novel were used to rank the characters on the basis of five centrality metrics. Among those metrics are the above described degree, betweenness and closeness centrality, as well as eigenvector and Katz centrality, two metrics on which Google's PageRank algorithm is based. PageRank is used by Google's search engine to rank web pages by relevance. PageRank, eigenvector and Katz are all based on the same, seemingly circular assumption that a node in a network becomes more important when it is connected to other important nodes (Page et al. (1998)). The computation of all these metrics was based on the *weighted* edges. Then, a regression analysis was carried out to see which of the demographic variables (gender, descent) is the best predictor for a character's place in the rankings.

5 Results regression analysis

Because of the exploratory nature of the present research and the absence of prior research on this topic, we did not have any formal hypothesis about which demographic factors would possibly determine a character's place in the rankings. However, we preferred to not just enter all possible variables into the regression equation as this would have possibly obscured the results of the analysis. Therefore, we formulated a non-formal hypothesis based on traditional, non-statistical research in the critique of literary representation. Several studies suggest that female characters and/or characters of mainly non-Western descent are often represented in a stereotypical manner and are therefore likely to be staged in less central, more marginal positions in literary texts (e.g. Pattynama (1994), Meijer (1996a), Meijer (1996b), Pattynama (1998), Minnaard (2010), Meijer (2011)). Gender and descent might therefore be possible

[14] https://networkx.github.io/, last accessed 7-5-2018.

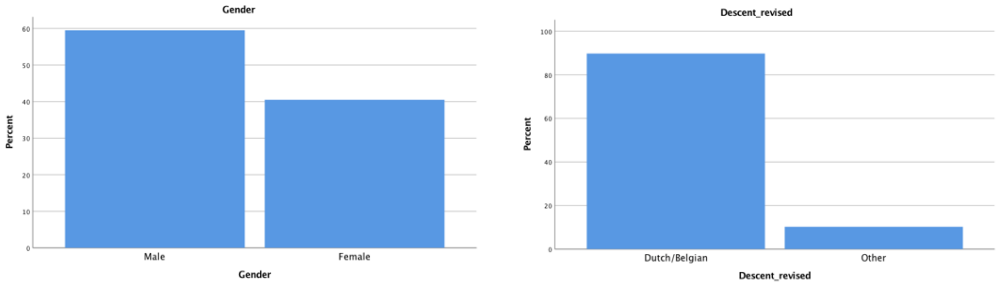

Figure 3: Gender and descent distributions among characters in the corpus ($N = 2137$). The numbers are based on types, not on tokens.

predictors of a character's position in the rankings. Based on descriptive statistics of our data, we suspected that male and non-immigrant characters would end up as more central,[15] as these types of characters are simply more present in the dataset (see fig. 3).[16] More precisely, we hypothesise that both male characters and non-immigrant characters will score higher on the centrality metrics than female characters and characters with a migrant background.

For each of the five centrality metrics (degree, betweenness, closeness, eigenvector, Katz), a multiple linear regression was conducted to predict characters' centrality scores based on their gender and descent. Gender is coded as 0 for male and 1 for female. Descent was coded as 0 for non-immigrant and 1 for immigrant. As our aim was to generalise across all novels, we did not include the division into sub-corpora (third-person, first-person, multi-perspective) in the statistical model.

No significant results were found for betweenness, closeness and eigenvector centrality. Gender and ethnicity are thus no predictors for characters' scores on betweenness, closeness and eigenvector centrality.

However, significant results were found for degree and Katz centrality. First, for degree centrality, a significant regression equation was found ($F(2, 2128) = 6.424$, $p < 0.01$), with an R^2 of 0.006. Characters' predicted degree centrality is equal to a B value of 0.428 + 0.024 (GENDER) + 0.059 (DESCENT) (see fig. 4). This means that, on degree centrality (on a scale from 0 to 1), female characters scored 0.024 higher than male characters, and immigrant characters scored 0.059 higher than non-immigrant characters.

Secondly, for Katz centrality a significant regression equation was found ($F(2, 2128) = 6.124$, $p < 0.01$), with an R^2 of 0.006. Characters' predicted Katz centrality is equal to a B value of 0.272 + 0.009 (DESCENT) + 0.007 (GENDER) (see fig. 5). This means that,

[15] Referring to characters who either have or do not have a migrant background, the terms 'immigrant' and 'non-immigrant' are used in a loose sense. In this article, immigrant characters can also refer to characters who are born in the Netherlands or Belgium but whose parents migrated to the Netherlands or Belgium. In this broad definition, immigrant characters are considered to have some sort of bond with a socio-cultural tradition that is not the same as their current country of residence. We chose the Netherlands and Belgium as a point of departure as the books in the corpus are either written by Dutch or Flemish authors who operate in a shared literary field of Dutch literature.

[16] A chi-square goodness of fit test was calculated comparing the occurrence of male and female characters with the hypothesized occurrence of a 50-50 gender distribution. Significant deviation from the hypothesized values was found ($\chi 2$ (1) = 82,030, p < .001). Also, a chi-square goodness of fit test was calculated comparing the occurrence of characters with a Dutch/Belgian and Other country descent with the hypothesized occurrence of a an equal distribution among those categories. Significant deviation from the hypothesized values was found ($\chi 2$ (1) = 1350,773, p < .001). This means that the 40-60 gender divide and the 89,8-10,2 divide in descent are not due to chance, but is a statistically significant difference.

Table 1: Linear model of predictors of degree centrality, with unstandardized coefficients (columns 2 and 3) and standardized coefficients (column 4)

Model		Unstandardized B	Unstandardized Std. Error	Standardized Beta	Sig.
1	(Constant)	0.438	0.006		0.000
	Descent_revised	0.058	0.019	0.065	0.003
2	(Constant)	0.428	0.008		0.000
	Descent_revised	0.059	0.019	0.066	0.002
	Gender	0.024	0.012	0.043	0.048

Table 2: Linear model of predictors of Katz centrality, with unstandardized coefficients (columns 2 and 3) and standardized coefficients (column 4)

Model		Unstandardized B	Unstandardized Std. Error	Standardized Beta	Sig.
1	(Constant)	0.273	0.002		0.000
	Gender	0.007	0.003	0.061	0.005
2	(Constant)	0.272	0.002		0.000
	Gender	0.007	0.003	0.062	0.004
	Descent_revised	0.009	0.004	0.045	0.038

on Katz centrality (on a scale from 0 to 1), immigrant characters scored 0.009 higher than non-immigrant characters, and female characters scored 0.007 higher than male characters.

These findings suggest that our initial hypothesis, based on traditional critiques of literary representation, should be rejected. Contrary to what we expected, female characters and immigrant characters scored higher, at least on two of the five centrality metrics used in the analysis. Furthermore, it should be noted that a higher frequency distribution of a character type does not necessarily lead to a more central position in a character network, as the results of the regression analysis has shown. Although male and non-immigrant characters are more present in the corpus, they do not end up as more central in network analytic terms. The question remains as to how these results can be interpreted in a close reading context.

6 Narratological evaluation

The quantitative representational patterns suggested by the outcome of the multiple linear regression require a narratological evaluation, as it is unclear what their significance is for the critique of literary representation. In concrete terms, the outcome for degree centrality is that female and immigrant characters have significantly more relations than male and non-immigrant characters. More specifically, women and characters with a migrant background often co-occur with a wider range of fellow characters in the novels. The higher scores of female and immigrant characters on Katz centrality indicate that they often co-occur with characters who also have relatively

high Katz centrality. In sum, female and immigrant characters have both *more* relations in general and more relations with *important* characters.

In order to make sense of this pattern, we subsequently conducted a small narratological exploration of character centrality in one novel from the corpus and confronted it with the results of the statistical analysis. As a case study, we used a novel that thematises both gender and ethnicity, as these issues of representation were taken as points of departure for the regression analysis. For the sake of brevity, we only used the two concepts of narrative mode and focalisation as points of departure. Note that there is a wide variety of other narratological concepts and perspectives that might potentially lead to alternative insights.

Eus (2012) by Özcan Akyol is a semi-autobiographical, first-person novel, in which the reader follows the life of the first-person narrator Eus, the son of Turkish immigrants living in Deventer, a small city in the Netherlands. Eus gets involved in criminal activities and ends up in jail, where he starts a writing career. This plotline foregrounds the theme of upward social mobility: a character with a migrant background who initially has a hard time finding his way in Dutch society eventually finds his creative ambition and becomes a successful author.

This theme is underscored by Eus's foregrounding of the economic and social hierarchies that exist between Dutch people and people with a migrant background. At the beginning of the novel, Eus states that he and his friends '... didn't dare to go to the better neighbourhoods', although they 'knew that they existed' (Akyol (2012): 24). An implicit opposition is thus postulated between 'better neighbourhoods' and Eus's own rough neighbourhood. Later on, Eus is more explicit when he characterises the 'indigenous youth, rich kids' as 'white scum' (idem: 120).

Another less prevalent, but latently present theme in the novel is the way men with a migrant background engage with (Dutch) women. Throughout the novel, women are treated with little respect by Eus and his friends. Female characters are either objects of sexual desire or considered a man's possession. They are repeatedly referred to as 'whores' (idem: 36, 58, 85, 145) and 'sluts', or variants of the term (idem: 43, 57, 62, 86, 145, 157, 176, 253). The male characters seem mostly interested in whether or not a woman is 'fuckable' (163). In general, women are constantly objectified. The following quote is a good example:

> Sometimes I stared out of the window for hours, in search of the hottest girls in school, about whom I then started fantasising. How beautiful they were! Nice tits! Nice ass! (idem: 50)

On the basis of such thematic cues, one could argue that two hierarchical oppositions take shape in the narrative:

1. immigrants $\longleftrightarrow$ non-immigrants
2. male $\longleftrightarrow$ female

Although there are some indications in the quotes given above, it is not evident what the specific hierarchy of these relations might be. The two basic concepts of narrative mode and focalisation might help to clarify this. First of all, the novel is narrated by Eus, which means that he controls the flow of information in the narrative. It is a logical consequence of the I-narration that Eus decides which events to report and which to leave out. When he, e.g., reports that 'I was born and raised in Koekstad, a small town by the IJssel, exactly on the border of two eastern provinces' (idem: 13),

he chooses to use an alias ('Koekstad') for a town the reader might know as Deventer. As an I-narrator, Eus is able to manipulate the narrative at will.

Furthermore, he is also the main focaliser: the narrative events are filtered through his perceptions. This means that the description of events is not neutral but colored by the vision and judgement of Eus. The following two sentences are illustrative:

> The next two years I went to the Hegius school, where I was surrounded by the beautiful, posh girls who pursued the highest level of education. Rumour had it that these girls had an above-average interest in foreign boys because they never saw those types of boys (idem: 37)

A whole group of girls is lumped together by Eus and characterised as 'beautiful' and 'posh'. Eus thus foregrounds their physical appearance and highlights their poshness, thereby suggesting that these girls are spoiled rich kids. In the second sentence he repeats a rumour relating to their supposed sexual interest in boys with a migrant background. These two sentences show an extremely biased representation of a specific type of characters (in this case: female, highly educated).

These basic narratological observations are of major importance for the interpretation of character hierarchies in the novel. As first-person narrator and main focaliser Eus is both a character with a migrant background and male, so the non-Dutch and male perspectives are a priori more dominant than the Dutch and female perspectives. Taking into account that Eus's friends and fellow criminals (Kosta, Ata, Meltem, Mahir) are also predominantly of a migrant background and male, one could argue that the centre of gravity lies with non-Dutch and male characters.

But although this interpretation of character centrality is based on plain narratological insights, it relies heavily on qualitative evidence (i.e. the few quotes we used to illustrate our point). A more data-driven and quantitative approach might potentially shed a different light on the question of character centrality in this novel. How does our network analytic approach relate to this narratological approach? Table 3 shows the characters in the novel ranked by their scores on degree centrality.

First of all, this novel conforms to the general pattern as observed in the regression equation only with regards to descent. 12 of the 21 identified characters have a migrant background, and they are higher in the rankings than the Dutch characters, which is in line with the general pattern according to which characters with a migrant background have significantly higher degree centrality.

However, with regards to gender, the novel deviates from the pattern. 14 of the 21 identified characters are male, and they occupy higher positions in the rankings on degree centrality, indicating that the male characters in *Eus* have *more* relations than female characters.

For Katz centrality, a similar pattern emerges. Table 4 lists the characters in the novel ranked by their scores on Katz centrality. Here, too, both characters with a migrant background and male characters occupy higher positions in the rankings than non-immigrant and female characters. The first types of characters are thus connected to more important characters than the latter.

Interestingly, these rankings are in line with the findings of our narratological approach to character centrality in the novel. Our narratological argument that the perspectives of male characters with a migrant background are dominant in terms of narrative mode and focalisation is backed up by our quantitative argument that these types of characters have higher scores on degree and Katz centrality. In this specific case, our narratological argument that the male and immigrant perspectives are dominant does not conflict with the character rankings for this particular novel.

Table 3: Characters in Eus (2012) ranked by degree centrality score

	Name	**Gender**	**Descent**	**Degree**
1	Kosta	male	immigrant	0.65
2	Kareltje	male	non-immigrant	0.55
3	Eus	male	immigrant	0.50
4	Turis	male	immigrant	0.40
5	Meltem	male	immigrant	0.40
6	Ata	male	immigrant	0.40
7	Mahir	male	immigrant	0.35
8	Selma	female	immigrant	0.30
9	Metin	male	immigrant	0.30
10	Haakneus	female	immigrant	0.30
11	Levine	female	non-immigrant	0.30
12	Theo	male	non-immigrant	0.15
13	Nathan	male	non-immigrant	0.15
14	Eef	female	non-immigrant	0.15
15	Inez	female	non-immigrant	0.1
16	Ömer	male	immigrant	0.1
17	Angelo	male	non-immigrant	0.1
18	Vinny	male	non-immigrant	0.1
19	Osman	male	immigrant	0.05
20	Daphne	female	non-immigrant	0.05
21	moeder Eus	female	immigrant	0.00

Table 4: Characters in Eus (2012) ranked by Katz centrality score

	Name	**Gender**	**Descent**	**Katz**
1	Kosta	male	immigrant	0.218218982823223
2	Mahir	male	immigrant	0.21821840342212764
3	Eus	male	immigrant	0.2182184034212662
4	Kareltje	male	non-immigrant	0.2182182875401764
5	Turis	male	immigrant	0.21821822960063142
6	Ata	male	immigrant	0.218218113719957
7	Meltem	male	immigrant	0.2182179398985992
8	Selma	female	immigrant	0.21821782401818632
9	Haakneus	female	immigrant	0.218217824018094
10	Levine	female	non-immigrant	0.21821782401787862
11	Metin	male	immigrant	0.2182178240177709
12	Theo	male	non-immigrant	0.2182177660783798
13	Nathan	male	non-immigrant	0.21821770813789643
14	Eef	female	non-immigrant	0.21821765019709
15	Inez	female	non-immigrant	0.21821759225711432
16	Angelo	male	non-immigrant	0.218217592257022
17	Vinny	male	non-immigrant	0.218217592257022
18	Ömer	male	immigrant	0.21821759225692972
19	Daphne	female	non-immigrant	0.21821753431670787
20	Osman	male	immigrant	0.21821753431661559
21	moeder Eus	female	immigrant	0.21821747637639374

In sum, our narratological evaluation highlights two important points with regards to the interpretability of our regression analysis. 1) A narratologically oriented analysis of a case can provide a qualitative contextualisation of a statistical argument. The mode of narration and focalisation in *Eus* illustrate the dominance of the male, immigrant perspective, which is supported by the characters rankings for the novel. A narratological argument may also give nuance to or conflict with a statistical argument, but that is not the case for this specific novel. 2) A specific novel (in this case, *Eus*) can very well deviate from an observed general statistical pattern. Characters with a migrant background score higher than Dutch characters in the novel in terms of network centrality, which is in line with our regression model. But contrary to the general pattern, the male characters score higher than the female characters. This highlights the importance of a qualitative contextualisation of the statistical analysis.

7 Conclusion

In this contribution, we have demonstrated that a data-driven and empirically informed approach to character centrality informs the ideological critique of literary representation in at least three ways.

First, instead of focusing on a limited number of cases, general patterns can be discerned in a larger corpus that represent a specific literary-historical period, which can lead to new, surprising insights. Contrary to what is suggested by the wide range of ideologically oriented close readings of character representation in literature, our results suggest that female and immigrant characters take up a more central position in the social networks of present-day Dutch literary fiction than non-immigrant and male characters, statistically speaking. This remarkable outcome requires an explanation, particularly in light of the highly imbalanced frequency distribution of immigrant and non-immigrant characters in the corpus. Almost 90% of the characters in the corpus are non-immigrants, but our regression model suggests that immigrants are more central in the networks than non-immigrants. Possibly, these higher centrality scores might be explained by the probability that novels that thematise descent, and stage a higher number of immigrants, also ascribe more central roles to them. Overall, the novels have fewer immigrant characters (only around 10% of all characters have an immigrant background), but these immigrants score higher on degree and Katz centrality. Something similar holds for female characters: there are fewer female characters than male characters in the corpus (almost a 40-60 ratio), but they have relatively high centrality values. In order for immigrant or female characters to be central in network theoretical terms, a high frequency of occurrence is not a necessary prerequisite as long as they interact with a high number of other (central) characters. This is a possible explanation of the discrepancy between the descriptive statistics and the outcome of the regression equation.

Second, combining a narratological close reading with network analysis enables a formalisation of abstract terms as importance, influence or power that are typically used in a strict metaphorical sense. Our case study on *Eus* illustrates that ideologically oriented interpretations regarding gender or ethnicity can be either backed up or nuanced by network statistics which make such interpretations less susceptible to unarticulated and implicit presuppositions.

Third, this study demonstrates that quantitative statistical patterns only make sense when confronted and contextualised with close readings of specific cases. Statistical trends might indicate general patterns of literary representation, but they can only

serve as an analytic backdrop for the individual analysis of particular novels. The novel *Eus*, for instance, lives up to the pattern only with regards to descent but not with regards to gender. The extent to which a single novel either conforms to or deviates from the general pattern can then be used to determine the particularity of a certain aspect of representation.

Our main contribution to the field of Digital Literary Studies lies in bringing together the methodological toolkits of narratology and social network analysis, which often seems to be lacking in data-driven approaches to networks in literary texts. As opposed to other network extraction methods, we propose a method that departs from domain knowledge and combines this with a bottom-up operationalisation of character interactions. As such, this article is situated within the mixed-method framework of the text-oriented Digital Humanities, and provides an argument for more strongly connecting qualitative and quantitative strands of research in the field.

8 Acknowledgements

The authors would like to thank annotators Maartje Weenink and Lisa Rooijackers for their precise annotations and Lucas van der Deijl and the anonymous reviewers for commenting on earlier versions of this article.

References

Agarwal et al.
2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. *IJCNLP*.

Akyol, O.
2012. *Eus*. Prometheus.

Alberich et al.
2002. Marvel universe looks almost like a real social network. *arXiv*.

Bal, M.
1977. *Narratologie: Essais sur la signification narrative dans quatre romans modernes*. Klincksieck.

Bal, M.
2009. *Narratology. Introduction to the Theory of Narrative*. Toronto/Buffalo/London.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, and A. Vespignani
2004. The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, 101(11):3747–3752.

Bluhdorn, H.
2010. *40 Jahre Partikelforschung*, chapter A semantic typology of sentence connectives. Stauffenburg Linguistik.

Brandes, U.
2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177.

Dekker, N., T. Kuhn, and M. Van Erp
2019. Evaluating
named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*.

Dijkstra, E. W.
1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

Elson et al.
2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Pp. 138–147.

Felski, R.
2009. After suspicion. *Profession*, Pp. 28–35.

Freeman, L. C.
1978. Centrality in social networks: Conceptual clarification. *Social Networks*.

Genette, G.
1972. *Narrative Discourse. An Essay in Method*. Oxford.

Grayson, S., K. Wade, G. Meaney, and D. Greene
2016. The sense and sensibility of different sliding windows in constructing co-occurrence networks from literature. *International Workshop on Computational History and Data-Driven Humanities*.

Herman, L. and B. Vervaeck
2005. *Vertelduivels. Handboek verhaalanalyse*. Vantilt.

Jahn, M.
1996. Windows of focalization: Deconstructing and reconstructing a narratological concept. *Style*.

Jayannavar et al.
2015. Validating literary theories using automatic social network extraction. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*.

J.Culler
1983. *On Deconstruction. Theory and Criticism after Structuralism*. Routledge.

Karsdorp et al.
2012. Casting a spell: Identification and ranking of actors in folktales. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*.

Karsdorp et al.
2015a. Animacy detection in stories. In *6th Workshop on Computational Models of Narrative*.

Karsdorp et al.
2015b. The love equation: Computational modeling of romantic relationships in french classical drama. In *Proceedings of the Sixth International Workshop on Computational Models of Narrative*.

Koolen, C.
2018. *Reading Beyond the Female. The Relationship Between Perception of Author Gender and Literary Quality*. ILLC Dissertation Series.

Lee and Wong
2016. Hierarchy of characters in the chinese buddhist canon. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*.

Lee and Yeung
2012. Extracting networks of people and places from literary texts. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*.

MacCarron, P. and R. Kenna
2012. Universal properties of mythological networks. *EPL*.

Mann, W. and S. Thompson
1988. Rhetorical structure theory: toward a functional theory of text organization. *Text: Interdisciplinary Journal for the Study of Discourse*.

Meijer, M.
1996a. *Het omstreden slachtoffer: geweld van vrouwen en mannen*, chapter De verschrikkelijke sneeuwman: projectie, geweld en nieuwe mannelijkheid in het werk van Jan Wolkers. Ambo.

Meijer, M.
1996b. *In tekst gevat. Inleiding tot de kritiek van de representatie*. Amsterdam University Press.

Meijer, M.
2011. *Diversiteit*, chapter Zwartheid in de witte verbeelding. Printart Press Kft.

Minnaard, L.
2010. The spectacle of an intercultural love affair. exoticm in van deyssel's blank en geel. *Journal of Dutch Literature*.

Moretti, F.
2013. *Distant reading*. Verso.

Nelles, W.
1990. Getting focalization into focus. *Poetics Today*.

Newman, M. E.
2001. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132.

Opsahl and Others
2010. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245 – 251.

Page et al.
1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pattynama, P.
1994. Oorden en woorden. over rassenvermenging, interetniciteit, en een indisch meisje. *Tijdschrift voor Vrouwenstudies*.

Pattynama, P.
1998. *Domesticating the Empire. Race, Gender, and Family Life in French and Dutch Colonialism*, chapter Secrets and Danger: Interracial Sexuality in Louis Couperus's The Hidden Force and Dutch Colonial Culture around 1900. University Press of Virginia.

Ricoeur, P.
1979. *Freud and Philosophy: An Essay on Interpretation*. Yale University Press.

Rydberg-Cox, J.
2011. Social networks and the language of greek tragedy. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*.

Smeets, R.
2017. Finding characters: An evaluation of named entity recognition tools for dutch literary fiction. In *European Alliance for Digital Humanities (EADH) day*.

Stiller et al.
2003. The small world of shakespeare's plays. *Hum Nat*.

Vala et al.
2015. 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. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Van Boven, E. and G. Dorleijn
2013. *Literair mechaniek*. Coutinho.

van der Deijl, L., S. Pieterse, M. Prinse, and R. Smeets
2016. Mapping the demographic landscape of characters in recent dutch prose: A quantitative approach to literary representation. *Journal of Dutch Literature*, 7(1):20–42.

Van der Deijl, L. and R. Smeets
2018. Tussen close en distant. personage-hiërarchieën in peter buwalda's bonita avenue. *Tijdschrift voor Nederlandse Taal -en Letterkunde*, 134(2):123–145.

Volker, B. and R. Smeets
2019. Mirrors or alternative worlds? comparing ego networks of characters in contemporary dutch literature with the population in the netherlands. *Poetics (In press)*.