Abstracts DHBenelux 2017 conference

Tuesday 4 July 2017

Session A

1. Coin Production in the Low Countries, fourteenth century to the present

Rombert Stapel[1], Jaco Zuijderduijn[2], Jan Lucassen[1], Kerim Meijer[1]
International Institute for Social History, Amsterdam, Netherlands
Lund University, Lund, Sweden

This project collects, combines and makes available data on mint production in the Low Countries (Netherlands, Belgium, Luxembourg) and has developed a web application to query and visualize the data, which is also linked to a digital map of (changing) historical boundaries in the Low Countries from 1100 to the present (available in Linked Open Data). It provides scholars with a user-friendly approach to large datasets, and allows them access to such variables as regional production figures and coin denominations.

Introduction

Monetization is a key concept in economics and in economic history. Throughout history currencies were a crucial element of economic exchange: first in the form of metal coins, which made up the lion's share of currencies, and were widely used in everyday transactions. Only much later paper money also emerged: before the First World War very few normal people would have ever seen paper money. Finally, nowadays non-material book money has become much more important than currencies, and with the onset of mobile banking and virtual currencies such as Bitcoin, it seems future generations will see much less currencies than people in the past.

Historical societies depended as much on media of exchange as we do today: coins and paper money helped a great deal in realizing everyday transactions, as did various forms of credit. Coin production figures are of crucial importance for understanding development in the long run.[1] The study of coinage, their quantity, denominations, use (e.g. in wage payments) and monetary policy in general provides important insight in economic and social history and this project provides historians a firm quantitative basis for their research.

In this paper, we will present the project and its goals, give an overview of the process of data collection and the web application we built to query and visualize the data (including geospatial visualizations), and provide some of the results for historical research that stem from our dataset.

Project

Coin Production in the Low Countries, fourteenth century to the present provides an overview of coin production figures covering many centuries. Of course we deal with omissions: not all mint accounts go back to the fourteenth century, and not all administration has survived. The website allows for an overview of the mint house data we have at our disposal at the moment, and

[1] Jan Lucassen and Jaco Zuijderduijn, 'Coins, currencies, and credit instruments. Media of exchange in economic and social history', *Tijdschrift voor sociale en economische geschiedenis* 11 (2014) 1-13.

visualizes the missing data. *Coin production in the Low Countries, fourteenth century to the present* also does not pretend to be the final dataset: like any other dataset it reflects the data that has been collected and made available up until now. Although we are confident we cover the vast majority of the coins minted in the Low Countries, some new or overlooked sources may emerge in the future; we are likely to make additions in time. The dataset represents the data we presently have, and is a tool to be used by scholars looking for variables related to coin production.

Our goal for this project was to take the aforementioned datasets, check the validity of the collected data, select and/or (re)calculate the relevant variables for our project, combine the different datasets, and present our selected variables in a web application which allows the user to query and visualize the data.

Manual web application

Figure 1. Number of coins minted in Flanders between 1334 and 1700, organised per alloy (status: November 2016).

In the web application[2], the user can query the data and create (and export) their own subsets. Different queries and selections can be made at the top left. This includes the possibility to display the *Value in denier groot,* a common coin used as money of account, in hourly wages. The query starts by clicking 'Run'. There are three tabs: 'Table', 'Chart', and 'Map'. The variables in the table and chart can be adjusted freely on the right. The map needs some further

[2] https://datasets.socialhistory.org/dataverse/coinproduction/search/.

introduction. At the moment, the map is used to give a rough indication how complete our dataset is for particular mint houses and authorities in time. We have turned to the works by Hugo Vanhoudt and H. Enno van Gelder, supplemented with data from our own datasets, to determine the years of activities of mint houses and authorities.[3]

The colour of a region that minted coins (e.g. Duchy of Brabant) will be dependent on the number of years (in a particular query) we know that region was minting coins and for which of those years we have actual production figures in our dataset. This also applies to the mint houses, where we have used pie charts. For this purpose, we have created a GIS map of all major authorities in the Low Countries in time.[4] This means that borders will change with time and mint houses will pop up and disappear.[5] A slider on the top left corner of the map allows the user to change the years. On the right of the application, different options regarding the map can be selected, choosing whether the colours and pie charts should change instantaneous with the slider or not.

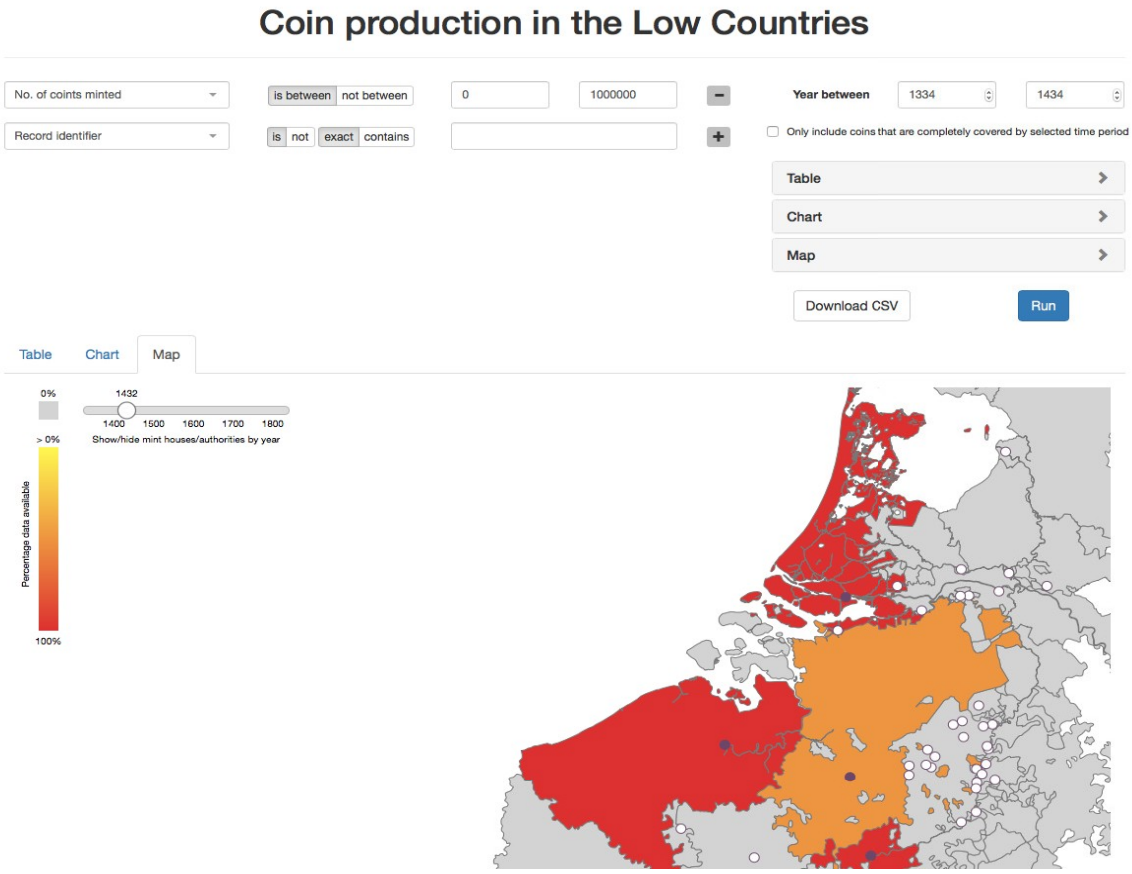


Figure 2. Map of the Low Countries (1432) with percentage of data availability in that year (status: November 2016).

[3] H. Vanhoudt, *Atlas der munten van België van de Kelten tot heden* (Heverlee 2007, 2nd edition); H.E. van Gelder, *De Nederlandse munten* (Utrecht 2002, 8th edition).

[4] For some important disclaimers regarding these GIS maps, see the introduction at http://hdl.handle.net/10622/HPIC74/.

[5] This process was visualized in a movie of the period 1100-2016, where each frame is a year: http://hdl.handle.net/10622/5KGG1T.

2. Mapping the Place: "De Krook Quarter"

Piraye Hacıgüzeller, Sally Chambers, Christophe Verbruggen and Hans Blomme
Ghent Centre for Digital Humanities, Ghent University

The presentation will elaborate on a new project Ghent Centre for Digital Humanities (Ghent CDH) is starting to carry out, "Mapping the Place: 'De Krook Quarter'", which involves "deep mapping" of a historical district in Ghent. In the presentation, the context, framework, workflow and impact of the project will be described and discussed.

The objective of the "Mapping the Place" project is to harness the well-demonstrated power of cartography as a participatory tool (Perkins 2007). Specifically, the project aims to contribute to the participatory governance of cultural heritage in Europe through "deep mapping" a district in Ghent (Belgium) that embodies place-based heritage such as Vooruit (a people's palace established in 1913 that has been turned into a vibrant international contemporary arts centre), the Minard Theatre, De Krook (the newly built city library and digital innovation centre) and adjoining former Wintercircus, and the surrounding streets (Kuiperskaai) that used to connect a Latin Quarter and red light district. In collaboration with the heritage institutions responsible for management of these places, Ghent CDH will employ a variety of participatory mapping tools and methodologies in order to involve a range of communities in a deep mapping project.

Deep maps are "thick spatial descriptions" of places breaking away from Cartesian paradigm in cartography. The latter, known also as "Western scientific mapping" (Pickles 2004; see Turnbull 1996), limit both content and methods of mapping as it traditionally aims to map only empirically observable phenomena that is considered to constitute reality exclusively. Deep maps, on the other hand, inspired by the concept of "thick description" coined by anthropologist Clifford Geertz (Bodenhamer et al. 2105), are based on a much more flexible and fruitful definition of what can constitute a map and what constitutes places aiming to bring together a large and rich array of spatial qualities. Deep mapping is even more promising today as digital cartography opens up many possibilities to collect and crowdsource new types of geospatial information and visualise, integrate and analyse it in novel ways with the help of technologies such as geographical information systems, virtual and augmented reality and, real time mapping.

The participatory deep map of Ghent, displayed in De Krook and Vooruit, will be an innovative, open ended, multi-vocal and largely digital cartographic process that will bring together geographical information, sensual experiences, memories, oral histories, creative narratives, emotions, knowledges, imaginations, practices and events. The map is planned to be produced through the following five types of activity: a) playful community mapping exercises (Pinder 1996; 2005) will be organised for diverse groups in order to carry out a certain cartographic task (e.g. mapping an area) and their knowledge and experiences of the places will be revealed in the process through their interaction (e.g. Grasseni 2004) b) a digital online crowdsourcing platform for heritage places will be created where people can enter cartographic information (see Perkins 2013); c) geospatial data on people's emotions (http://biomapping.net/), movement, sound and smell will be collected in real-time and converted into data sculptures or paintings by artists (see, e.g., www.refikanadol.com/); d) multi-layered geographic information systems and three-dimensional virtual reality displays will be installed in De Krook affording a diverse groups of visitors to annotate their experiences and knowledge about heritage places focused in the deep mapping project e) (non-) digital map-based or -aided games (e.g. geocaching) will be designed, developed and/or employed in order to facilitate conversation about heritage places in question between diverse group of people as well as informing and engaging them with these places. The layers of the participatory deep map will be distributed across many locals in De Krook comprising a geographical information systems component, virtual

reality room, game room, exhibition room, digital sculpture and painting rooms, screens for real time mapping, and computers with access to the digital crowdsourcing platform.

References

Bodenhamer, D.J., Corrigan, J. & Harris, T.M. eds., 2015. *Deep maps and spatial narratives*, Indiana: Indiana University Press.

Grasseni, C., 2004. Skilled landscapes : mapping practices of locality. *Environment and Planning D: Society and Space*, 22, pp.699–717.

Perkins, C., 2007. Community mapping. *The Cartographic Journal*, 44(2), pp.127–137.

Perkins, C., 2013. Plotting practices and politics: (Im)mutable narratives in OpenStreetMap. *Transactions of the Institute of British Geographers*, 39(2), pp.304–317.

Pickles, J., 2004. *A history of spaces: Cartographic reason, mapping and the geo-coded world*, London & New York: Routledge.

Pinder, D., 1996. Subverting cartography: the situationists and maps of the city. *Environment and Planning A*, 28, pp.405–427.

Pinder, D., 2005. Arts of urban exploration. *Cultural Geographies*, 12(4), pp.383–411.

Turnbull, D., 1996. Cartography and science in early modern Europe: mapping the construction of knowledge spaces. *Imago Mundi*, 48, pp.5–24.

3. Cinemas on the Move: A geospatial analysis of the role of traveling cinemas in the Dutch cinema landscape

Jolanda Visser, Julia Noordegraaf and Ivan Kisjes
University of Amsterdam

The emergence of the cinema as a new cultural industry at the dawn of the twentieth century has had a significant impact on the social, cultural and economic infrastructures of modernizing societies. Cinema's technological and cultural innovation, combined with economic competition, significantly reconfigured the role and place of entertainment culture in public life. Besides being an economic factor of importance, it also has literally "taken place" in urban and rural infrastructures, transforming the organization and experience of modern public space.

The ways in which cinema has taken place in Dutch public space has been the subject of a number of studies. Some focus on the history of specific cinema theatres and the urban context in which they function (Visser 2012; Noordegraaf et al. 2016). Others have investigated national and local cinema networks and focused on the organization and economics of the industry (Dibbets 1980 & 2006; Oort 2016). Yet other studies focused on the ways in which movies reached their audiences and how this correlates with specific religious and ideological orientations (Boter and Clara Pafort-Overduin: 2009), or studied the popularity of certain genres or stars (Van Beusekom 2013). In addition, a comprehensive database has been created that facilitates data-driven research on national Dutch film culture.[6]

At the same time, though, the study of the role of cinema in modern public life has focused primarily on urban contexts. When plotting the locations of cinemas from the Cinema Context on a map, it

[6] www.cinemacontext.nl

appears that the majority of cinemas is located in urbanized areas. In fact, there were cinema screenings in less urbanized areas as well; those were frequented by traveling cinemas. At present, the role and impact of these traveling cinemas in Dutch cinema culture remains entirely unknown. In this paper, we present the results of the very first study of the impact of traveling cinemas on Dutch film culture. Using a combination of network and geospatial analysis software, the paper contributes: 1. new insights into the way cinema as a leisure industry contributed to the shaping of modern Dutch identity; and 2. a reflection on the affordances and limitations of GIS and network analysis tools for (cinema) historical research.

Central Question

Our research aims to establish the role and place of traveling cinemas in the Dutch, post-WWII cinema landscape. What was the relation between the permanent and traveling cinemas, in terms of geographical distribution, market share, and distribution and exhibition practices? In order to answer this question, we approach the Dutch cinema landscape as a network with socio-economic (distribution, consumption) and cultural (programming) dimensions. In order to analyse this network, we combine a geospatial analysis of the network of permanent and traveling cinemas and owners/exhibitors in The Netherlands in 1949 with an in-depth case study of one particular section of this market. This combination allows us to combine a macrosocial analysis of the role of traveling cinemas in the national cinema market with an analysis of the contextual features that explain causality in one specific case (Ragin 1987).

Method

For the research, we adopted a two-tiered approach. First, we extended the data on the location of permanent cinemas and their owners in the Cinema Context database with newly assembled data on the places frequented by traveling cinemas. Then, we mapped these cinemas according to their typologies, distinguishing between permanent theatres, theatres with occasional screenings and traveling cinemas in QGIS. This resulted in a geospatial analysis of the organization of the Dutch industry that, for the first time, includes data on traveling cinemas.

Second, the networks of cinema exhibitors of permanent and traveling cinemas have been analyzed by processing the data on theatres and owners/exhibitors in Gephi. The resulting graph allowed us to acknowledge the influence of cinema chains as well as individual, non-networked entrepreneurs. By projecting these data on historical maps in QGis, we could compare the geographical distribution of different types of cinemas with the network of cinema owners/exhibitors. We identified a number of clusters where permanent cinemas and mobile cinemas were related and used this analysis to select one case for further, in-depth analysis of film flows within a cinema chain with a traveling department. The selected case study tracks the film flows of the cinema chain of Joh. Miedema and his competitors in the Northern province of Friesland in 1949.

Results

Some of the data sets used already existed (Cinema Context database), some had to be digitized partly (census data) and some had to be created (film programming, traveling cinema locations and screenings). In the first phase of the project the data of the cinemas and the networks of cinemas were combined. The first results showed the geographical networks of Dutch permanent cinemas in relation to the network of owners/exhibitors. In general, as also shown by Dibbets (1980), one can conclude that half of the cinemas belonged to a cinema chain, leaving the other half as isolates.

After adding the mobile cinema networks, we identified a clear geographical distribution for exhibitors of a cinema chain with a traveling department, among others in the provinces Friesland and Drenthe. The selected case study focused on the network of Joh. Miedema in

Friesland, which comprised 10 permanent cinemas surrounded with places he claimed for his mobile department. It appears he used these mobile screening locations for constructing a buffer zone around the permanent cinemas in his chain, to ward of competition from other owners in the region. Reconstructing film programming practices within that network and comparing that to those of his competitors in the province of Friesland in 1949 provides new insights in the economics of a cinema chain with a traveling department, the socio-economic and cultural context of these various sites visited, and patterns of taste. Based on the first results of this research, the benefits and pitfalls of the combined use of Gephi and QGIS will also be evaluated.

References

Beusekom, Ansje van. "Distributing, programming and recycling Asta Nielsen films in the Netherlands, 1911-1920." In *Importing Asta Nielsen: The international film star in the making 1910-1914*, edited by Martin Loiperdinger & Uli Jung, 259-272. New Barnet, Herts UK: John Libbey/KINtop, 2013.

Boter, Jaap, and Clara Pafort-Overduin. "Compartementalisation and Its Influence on Film Distribution and Exhibition in The Netherlands, 1934-1936." In *Digital Tools in Media Studies: Analysis and Research : An Overview*, edited by Michael Ross, Manfred Grauer, and Bernhard Freisleben, 55–68. Bielefeld: Transcript Verlag, 2009.

Dibbets, Karel. "Bioscoopketens in Nederland: Economische concentratie en geografische spreiding van een bedrijfstak, 1928-1977." Doctoraalscriptie, Universiteit van Amsterdam, 1980. online: http://kd.home.xs4all.nl/home/Karel%20Dibbets%20%20Bioscoopketens%20in%20Nederland%201980.pdf

Dibbets, Karel. "Het Taboe van de Nederlandse Filmcultuur: Neutraal in Een Verzuild Land." *Tijdschrift Voor Mediageschiedenis* 9, no. 2 (2006): 46–64.

Hallam, Julia, and Les Roberts, eds. *Locating the Moving Image: New Approaches to Film and Place*, 2014.

Horak, Laura. "Using Digital Maps to Investigate Cinema History." In *The Arclight Guidebook to Media History and the Digital Humanities*, edited by Charles R Acland and Eric Hoyt, 65–102. Falmer: Reframe Books, 2016.

Noordegraaf, Julia, Opgenhaffen, Loes, & Bakker, Norbert. "Cinema Parisien 3D: 3D Visualisation as a Tool for the History of Cinemagoing". *Alphaville*, 11 (2016): 45-61.

Oort, Thunnis van. "Industrial Organization of Film Exhibitors in the Low Countries: Comparing the Netherlands and Belgium, 1945–1960." *Historical Journal of Film, Radio and Television* (March 17, 2016): 1–24. Online first: http://dx.doi.org/10.1080/01439685.2016.1157294

http://dx.doi.org/10.1080/01439685.2016.1157294

Oort, Thunnis van. "'Coming up This Weekend': Ambulant Film Exhibition in the Netherlands". (Forthcoming).

Ragin, Charles C. *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press, 1987.

Visser, Jolanda, *Samen naar The Movies – 100 jaar Bioscoop op de Haarlemmerdijk 161*, The Movies Art House Cinemas and Film Distribution Amsterdam: 2012.

Session B

1. Soft skills in hard places: the changing face of DH training in European research infrastructures

Jennifer Edmond, Trinity College Dublin
Vicky Garnett, Trinity College Dublin

Research infrastructures are becoming an increasingly distinct presence in the landscape of the digital humanities, creating unique research ecosystems that interact with, but remain distinct from, the traditional university-based ones. It is a research sector still very much in the process of defining itself, however, in particular in the arts and humanities, not only in terms of how exactly infrastructures support research but also in terms of how a word with such "hard" connotations (conjuring up images of roads and bridges) can encompass the many "soft" resources and skills, from data to know-how, that we now recognise as a part of infrastructural provision for research in Europe. This tension is already in how research infrastructure is defined, with some camps preferring to fall back on long lists of elements infrastructure may or may not comprise, such as data, services and tools, while others remain more theoretical, placing them in the role of "mediating" (Badenoch and Flickers, 2010) or "below the level of the work" (Edwards et al.., 2012). Regardless of how we conceptualise it, however, infrastructure is undeniable as a rising presence, with a growing impact on how research is conceptualised and carried out, how research results are communicated and shared, and how the potential scale of a humanities project can be conceptualised.

There is one element in this landscape of change that has steadfastly remained based within the universities, however: that is the manner in which new generations of researchers are formed, through training and education. Some of the reasons for this lie in the need for specialised procedures, staff, resources and expertise to deliver formal educational programmes, a layer of provision that research infrastructures seldom have. Indeed, it is the lack of this layer that most distinctly differentiates activities of the research infrastructure from those of the more familiar academic context. As we continue to develop our understanding of what it means to 'teach' the digital humanities (eg. Fyfe, 2011, Hirsch, ed, 2012, or Bellamy, 2012), however, we need also to reconsider the utility, responsibility and potential contributions of other actors than universities in this process, and how we integrate them into recognised learning pathways. It is not infrastructures do not offer training opportunities, just that the paradigm informing much of this training has historically been founded upon a more narrow conceptualisation of the added value of the infrastructural space for creating and sharing unique knowledge. As such, projects and platforms would traditionally create materials to assist users approaching specific tools developed or hosted by the infrastructure, serving a very narrow conceptualisation of the user and his or her needs.

There has been an increasing number of examples of the infrastructural community expanding their activities to fill spaces less easily addressed by traditional, formal, course- and institution-based training contexts, however. Hands-on training with specific collections or objects, or using transnational access to build skills, for example, are mechanisms that have been developed to great effect by infrastructures, as has the model of partnering with other organisations to deliver credit-bearing programmes. These are mechanisms that have arisen in part because of the opportunities that exist, for example, when researchers work in close proximity to specific scientific instruments, as in the fields of cultural heritage and preservation, but have also arisen as accidents of design. Many research infrastructure funding schemes include fixed elements drawn directly from the longer tradition of infrastructure development in the fields of science and technology, mechanisms that do not necessarily fit humanities modes of work or interaction.

Even as such programmes remained largely ad hoc extensions of the originating user-support model of training, they exposed the potential of research infrastructures not only as places that support research, but where unique knowledge was being created, and where this knowledge could and should be shared. The development of a theoretical understanding of the strengths of the research infrastructure, what knowledge they contribute to digital humanities, and how this knowledge could be more systematically shared has been a primary goal of the training programme of the PARTHENOS (Pooling Activities, Resources and Tools for Heritage e-Research, Optimization and Synergies, http://www.parthenos-project.eu/) cluster project, itself a collaboration between a number of research infrastructures and their affiliated projects.

As an infrastructure cluster, PARTHENOS is charged with deepening understanding of what infrastructure is and how common activities can be better aligned for maximal benefit to researchers between the communities that have built landmark research infrastructures at European level. The PARTHENOS training framework seeks first and foremost to make a distinction between research work that does and that does not engage with data and service infrastructures such as the PARTHENOS partners represent. At the next level, the framework seeks to address the digital humanities not only as a set of domains, but also as a set of roles and actors, following upon the work of the DigCurv project (http://www.digcurv.gla.ac.uk/). By reconceptualising a didactic system from the first principles of who might need digital infrastructure and what they might need to know or be able to do, PARTHENOS has been able to create bespoke training materials that draw from the uniques experiences within research infrastructures and the unique knowledge they create. The materials exist within a simple but evolving framework, addressing experience levels from the novice (for example: "What is an Infrastructure"), to the intermediate (for example: "Management Challenges in Research Infrastructures") and advanced (for example: "Introduction to Infrastructures as Collaborations") levels. Modules are designed to build bridges between potential users and the entire context of the research infrastructure and how they operate, answering fundamental questions about what resources are available and how they operate, through to much more fundamental explorations of the opportunities and challenges that exist in this environment, issues that even expert practitioners struggle to define and address.

The paper will embed a presentation of PARTHENOS's work in a theoretical discussion of the role of research infrastructures in the development of skills and careers in the digital humanities. It will give an overview of some of the practical interventions the project has made to address the thorny issues of developing training and education programmes outside of the academy, including awareness raising, foresight work, embedding in higher education, partnerships and accreditation. Working in concert with its constituent partners (the DARIAH, CLARIN and E-RIHs Research Infrastructures, as well as their partner projects, such as CENDARI, EHRI, ARIADNE, and IPERION CH), the PARTHENOS team is testing the potential for infrastructural knowledge, for its transmission as materials for self-directed use by independent learners and trainers, and for its capacity to be integrated in the programmes of universities and professional organisations alike. Through this programme of engagement PARTHENOS will not only bring an extended horizon for training to research infrastructures and their users, but to all of digital humanities.

References

Badenoch, A., and A. Fickers, Materializing Europe: Transnational Infrastructures and the Project of Europe (Palgrave McMillan, 2010)

Bellamy, Craig, 'The Sound of Many Hands Clapping: Teaching the Digital Humanities through Virtual Research Environment (VREs)', Digital Humanities Quarterly, 6 (2012)

Edwards, Paul N., Knobel, Cory P., Jackson, Steven J., and Bowker, Geoffrey C., Understanding Infrastructure: Dynamics, Tensions, and Design <http://hdl.handle.net/2027.42/49353> [accessed 16 November 2012]

Fyfe, Paul, 'Digital Pedagogy Unplugged', Digital Humanities Quarterly, 5 (2011)

Hirsch, Brett D., Digital Humanities Pedagogy: Practices, Principles and Politics (Cambridge: Open Book Publisher, 2012) <http://www.openbookpublishers.com/product/161/digital-humanities-pedagogy--practices--principles-and-politics> [accessed 7 April 2017]

2. Ranke.2 - How to Get Digital Source Criticism on the Teaching Agenda

Stefania Scagliola - C2DH – Centre for Contemporary and Digital History
University of Luxemburg

Abstract

The term ***Ranke.2*** refers to the need to reassess Leopold von Ranke's method for historical source criticism, in the light of the impact of digitization and the world wide web on the position of the archive and the craft of the historian. It is also the proposed title of a platform for lessons on digital source criticism, a project that is being developed at the Centre for Contemporary and Digital History at the University of Luxemburg.

While a number of scholars have successfully addressed various theoretical and epistemological implications of the digital turn for the historical craft, little is known about how this subject is dealt with in the realm of teaching. This paper pleads for an assessment of the concept of Digital Source Criticism from the perspective of Digital Humanities Pedagogy. It starts off with some reflections on why and how Ranke's concept has to be reconsidered. Then it discusses whether source criticism can still be regarded as a specific historical method. The third section of the paper is an account of a small-scale exploration among humanities scholars involved in teaching at the humanities faculty of the University of Luxemburg. They were asked to share their understanding of how digital source criticism should be taught. The paper concludes with a plea for a integrating small scale DH interventions into the traditional historical curriculum.

'Everything has changed and everything has stayed the same'

With the arrival of digitally-based 'fake news' and the inability of sections of the public to distinguish it from the 'real thing', the vital importance of digital source criticism should be evident. What is less evident is how it affects the craft of the historian. Historians educated in the 21st century are witnessing the consolidation of the 'digital turn' with profound consequences for the historical profession. The German scholar Leopold von Ranke was responsible for an earlier radical change in scholarly practice in the 19th century: he introduced the so-called 'archival turn'. He also introduced the concept of the 'seminar'and encouraged a new generation of aspiring scholars to visit numerous archives, scrutinize and compare documents, and trace back the identity and motives of the author and the circumstances under which a document came into existence. Ranke made a distinction between 'external' source criticism, which focuses on the creation, appearance and alleged or real authenticity of a source, and 'internal' source criticism, which evaluates the evidential value that can be attributed to a particular source. This new approach became widespread and problematized the tradition of 'universal histories', based on broad philosophical concepts and ideas about the evolution of mankind. Rigorous fact-checking came in place of myth-making. Ranke's innovation in the second half of the 19th century coincided with the period of modern state formation and the creation of national archives. It gradually became the backbone of professional history, with a strong orientation towards the archive as the guardian of authenticity and historical relevance (Risbjerg Eskildsen 2008).

We now live in globalized world with cultural and disciplinary boundaries that are blurred, with digital technology that has permeated the academic research practice, and with the opportunity to copy, alter and remix data with relative ease. It therefore comes as no surprise that concern about the origin, authenticity and value of historical sources in digital form is increasing (Jones and Hafner, 2012) How this has affected the historical profession and what changes need to be introduced has been discussed by several scholars (Fickers 2012, Sternfeld 2014, Zaagsma 2014, Föhr 2015). They plead for a critical reflection on the nature of sources in digital form and for an investment in digital skills to be enable students and practitioners to apply digital tools in a professional manner and understand their potential, bias and limits. Critical reading and thinking are no longer enough in terms of safeguards, but have to be complemented with a more technical and mathematical understanding of digital phenomena. (Scagliola 2016)

In addition to the traditional Rankian inquiry into the context in which a historical source came into existence, two additional processes of creation and possible manipulation need to be scrutinized. The first involves identifying alterations and loss of context that occur during the transformation from analog source to digital object. (Fickers 2012, Treleani 2013). Transparency should be the norm, as to who was involved in the chain of digitization, what choices were made and what tools were used. If this is absent, the scholar must have enough contextual and technical knowledge to be able to identify and reconstruct to the extent possible this gap and evaluate how this may influence the historical interpretation of the object.

The second process relates to a better understanding of the algorithm-based selection bias of search engines since these increasingly determine our reference frame and have also penetrated academic library systems (Van Dijk 2010, Vaidhyanathan 2009). It looks as if our earlier dependency on the policy of the national archive with regard to granting access to documents based on national security and other concerns, has been substituted by one on the biggest stake holders in search technology: Google. The merits and perils of algorithm-based search technologies have been the object of academic debates and have led to reflections on the epistemology of the digital environment (Wouters et al 2013, Liu 2014). However, these remain limited discussions between the 'usual suspects within the community of DH scholars'. They do not seem to matter enough to push for reforming if not revolutionizing the curriculum.

Crap-Detection or Digital Philology?

The question we face is how to go about to adjust and adapt the classical humanities curriculum to the requirements of 21st century academic research. Where do we start? Should we make a distinction between general academic digital skills and those that are calibratred for specific fields of research such as history?

When observing the learning subject 'methods of research', which is often taught in the first year of a humanities bachelor curriculum, one gains the impression that with the 'Googlelization of knowledge' and the more general digitization of information (Vaidhyanathan 2009) topics that in the past belonged to distinctive (sub-)fields of research such as critical media studies, information science, literacy studies and education studies are now more and more alike. This calls for a renegotiation of boundaries and specification of what is distinctive about history.

When we look at the realm of education, the call for training young people in assessing the trustworthiness of what they consult and of what they engage with through social media, is a recurrent feature. There are many initiatives aiming at making the use of digital media less dangerous for the novice in the field. (Scanlon 2014, Cartelli 2013, Bellanca 2010) The writer Howard Rheingold has re-introduced Hemingway's journalistic principles for 'crap-detection', and points to the importance of web resources that give advise on how to detect false information (Rheigold 2013).

However, when students enter academia with the intent to explore the world of historical narratives, philosophical concepts and general cultural heritage, will the possession of general critical media literacy be enough to avoid pitfalls? It seems that some special skills are needed. In addition to being able to distinguish fake from real, they should also be able to trace back the history of the various versions of a document. This philological inquiry in a digital environment requires understanding the back end of a digital document and sometimes requires applying forensic software to detect the trail of binary digits that each manipulation has left. Moreover, Web.2 and forthcoming Web.3 technology also require students and academics to be able to express their thoughts and insights in other ways then writing a text in the form of an essay. Therefore, digital source criticism when applied to history, involves more than a mere critical reading of digital sources and writing of articles that are published online. It entails the active application of tools to trace and detect changes, and to create digital content. It is not just one more method as part of a wider repertoire of the historian's craft, it is a new concept of conducting historical research. This has serious implications for what needs to be put in practice and consequences for its relationhip to the existing curriculum. This has an established status with engraved social practices, in which lecturers are involved who have put effort in it. Changing these practices requires patience and diplomacy.

On the Verge of Transformation

Passive if not active resistance among lecturers when trying to introduce digital methods in the humanities is not uncommon. This is often seen as being an instinctive reaction to protect established positions of power and expertise (Scanlon 2013, De Jong et all 2011). Fear for new technologies and distrust of rosy promises about what such technologies can do, also play a role. Another obstructive element can be the rigid organizational structure of traditional academic teaching, that is based on the time span of lectures of just one or two hours. This hardly leaves space for learning new skills let alone experimenting. (Henderson and Romeo 2013).

To explore the space for the subject of Digital Source Criticism at the Faculty of Humanities of the University of Luxemburg, a small-scale user study was conducted.[7] The Faculty is a salient environment for testing interest in Digital Source Criticism, as it is experiencing considerable institutional changes. As of October 2016, the new Centre for Contemporary and Digital History has been established, that will take up innovative research and teaching in close collaboration with its former basis, the Institute of History.

The first part of the user study consisted of a presentation of the envisioned format for lessons on DSC during the main meeting of the Institute of history, followed by a survey.

The plan is to create an appealing video essay around a particular data type in which the digital version is problematized and compared to its analog version. Subsequently students have to read literature and conduct research, and finally create a digital publication or object with a similar type of data with the help of digital tools. The survey to collect feedback on this format was set out to 40 colleague historians, a mix of professors, lecturers and Ph.D. students. This yielded nine benevolent responses, which all stressed the importance of the topic, but also the existing limitations to integrate it into their lessons, due to lack of expertise and time, and of space within the limits of the prescribed ICTS.

The next step was to organize focus groups with colleagues from the new center. Four meetings were held with three to four participants, a mix of junior and senior colleagues. In addition, a few face-to-face interviews were held. The background of the participants varied, most of them were

[7] The consultation of lecturers is work in progress; it should be completed in the coming months and should yield a more solid foundation for designing and realizing **Ranke.2,** the new teaching platform on Digital Source Criticism.

historians, among which media studies was overrepresented. Special weight was given to the feedback of an information scientist and of two historians specialized in digital methods, all three with ample teaching experience. Again, they were first shown the presentation on the ideal typical format of the Digital Source Criticism lesson, after which three main questions were presented:

I. In what way is digital source criticism relevant for your research?
II. What do you regard as necessary digital skills for students (basic, academic, specific for historians);
III. What would you choose to integrate in your courses, the video essay, the assignments, the hands-on component or a combination?

The feedback to the presentation and questions was in most cases recorded and later transcribed. In a few cases notes were jotted down during the interview. The most salient concerns and preferences that came out of the consultations are summarized below:

- The level of digital literacy when entering the university

The level of competences is too diverse because of lack of systematic coverage of the topic in secondary education. An entrance test should be considered to be able to cover the gaps with individual training units.

- Limited Time.

Digital Literacy and competences to deal with digital data, are best taught in collaborative projects that take up time because of the need to teach skills. Think of how much time it takes to learn to write according to academic standards. At the same time, lecturers of thematic courses consider digital source criticism as a topic that belongs to the subject 'research methods' - a subject with a limited amount of hours in the curriculum which is offered only once, most often in the first year of a bachelor. Most teaching is thematic and not about methods.

- The 'branding' of the term Digital Source Criticism is problematic

Creating a special term for this type of source criticism *suggests* it is a different and new practice. A lecturer of 'methods of research' suggested to use the generic term *Source Criticism*, that can be applied to any source, regardless of whether it is an analogue or digital form.

- There is a need for continuity in the 'framing' of the problem.

Some lecturers of media studies stated that giving too much attention to the transformation from analog to digital, risks to obscure the many transformations and manipulations that already occur between analog media (e.g. in the process of editing of newsreel). They prefer to frame the subject in a more general way, e.g. 'reflecting on transformations'.

- The majority of researchers and PhD work with non-digitized sources.

Taking into account how many lecturers and researchers work with thematic subjects and with data and literature that is not digitized, it would be disproportionate to place Digital Source Criticism, a methodological topic, as a central subject on the curriculum. The principle of 'hybrid' research cultures should be emphasized as it connects better to the dominant teaching practice.

Conclusion

To address such concerns a smart communication strategy should be considered in which 'digital source criticism' is presented as a 'hybrid concept' that encompasses both differences and continuities in dealing with source criticism. What could be considered is to substitute the principle of a series of lessons that would take up much of the time in the curriculum, with smaller teaching units with a digital component. These could be complementary in a thematic course, and more central in a methodological subject. A way to support this approach

would be to follow the pedagogical principle of the SAMR model, which stands for *Substitute, Augment, Modify, Redefine*. It was designed to gradually integrate technology into the curriculum (Puentedura 2014). The process starts with first merely substituting tasks that have to be completed manually with a technology, and then gradually adding technological components to familiarize new users to the possibilities that they offer. The outcome of this gradual process should lead to a redefinition of the original task.

This SAMR model approach is currently being considered as an instrument to realize the envisioned transition. At the same time, however, master and PhD students will be immersed in intensive DH collaborative courses with experimental components at the new centre.

The policy of combining gradual change with immersive and experimental learning could be the solution to create a common ground among different generations of historians and future generations of students of history.

References

James A. Bellanca (2010), *21st Century Skills: Rethinking How Students Learn*, Solution Tree Press. See also: http://www.p21.org/about-us/our-history

Catherine Francis Brooks (2016). 'Disciplinary convergence and interdisciplinary curricula for students in an information society'. In: Innovations in Education and Teaching International, http://www.tandfonline.com/toc/riie20/current

Antonio Cartelli (2013), (ed) *Fostering 21st Century Digital Literacy and Technical Compentency,* Information Science Reference.

Jose Van Dijck (2010), Search engines and the production of academic knowledge. *International Journal of Cultural Studies*, 13(6) .doi:10.1177/1367877910376582.

Andreas Fickers (2012) 'Towards A New Digital Historicism? Doing History in the Age of Abundance.' *VIEW Journal of European Television History and Culture*, 1(1).

Pascal Föhr, "Poster ‚Historical Source Criticism in the Digital Age'," *Historical Source Criticism*, 31. März 2015,http://hsc.hypotheses.org/328..

Michael Henderson, and Jeoff Romeo (2016), Teaching and Digital Technologies: Big Issues and Critical Questions: Cambridge University Press.

Rodney H. Jones and Christoph A.Hafner (2012), *Understanding Digital Literacies; a Practical Introduction,* Routledge.

De Jong, Ordelman, Scagliola, Audio-visual Collections and the User Needs of Scholars in the Humanities; a Case for Co-Development, *Proceedings of Supporting Digital Humanities,* 2011, Copenhagen. http://files.beeldengeluid.nl/pdf/r-en-d_audio-visual-collections-and-userneeds_dejong-ordelman-scagliola_20111117.pdf

Alan Liu (2014) "Theses on the Epistemology of the Digital: Advice For the Cambridge Centre for Digital Knowledge." http://liu.english.ucsb.edu/theses-on-the-epistemology-of-the-digital-page

Ruben Puentedura (2014), SAMR and TPCK: A Hands-On Approach to Classroom Practice http://www.hippasus.com/rrpweblog/archives/000140.html

Harold Rheingold (2013). http://rheingold.com/2013/crap-detection-mini-course/ retrieved 1-5-2017.

Kasper Risbjerg Eskildsen, 'Leopold ranke's archival turn: location and evidence in modern Historiography', *Modern Intellectual History*, 5, 3 (2008), pp. 425–453 C _ 2008 Cambridge. doi:10.1017/S1479244308001753

Eileen Scanlon, E. (2014), Scholarship in the digital age: Open educational resources, publication and public engagement. *BrEduc Technol,* 45: 12–23. doi:10.1111/bjet.12010

Matteo Treleani (2013), 'Recontextualisation; ce que les média numériques font aux documents audiovisuels', in: R éseaux,1, (no 177) http://www.cairn.info/publications-de-Treleani-Matteo--99590.htm

Stefania Scagliola (2016), Digital Source Criticism in the 21st Century: Reconsidering Ranke's Principle in the Digital Age, blog Digital History Lab, August 2016. http://www.dhlab.lu/blog-post/digital-source-criticism-inthe-21st-century-reconsidering-rankes-principles-in-the-digital-age/

Joshua Sternfeld (2014), 'Historical Understandings in the Quantum Age', *Journal of Digital Humanities,* Vol 3, nr. 2, http://journalofdigitalhumanities.org/3-2/historical-understanding-in-thequantum-age/

Siva Vaidhyanathan (2009), 'The Googlization o f Universities', in: The NEA 2009 Almanac of Higher Education, 2009 http://www.nea.org/assets/img/PubAlmanac/ALM_09_06.pdf

Paul Wouters, Anne Beaulieu , Andrea Scharnhorst and Sally Wyatt (2013) (eds), Virtual Knowledge; Experimenting in the Humanities and the Social Sciences (Eds.)

Gerben Zaagsma, 'On Digital History", *BMGN - Low Countries Historical Review* 128/4 (2013) 3-29.

3. Individual presentation: Video essays and the new possibilities for film criticism and pedagogy

Irina Trocan,
Cinema and Media PhD, National University of Film and Theatre Bucharest

The shift of film criticism to the online sphere in recent years has led to a number of mutations, including the increase in popularity of a relatively new format: the video essay. Roughly an audiovisual version of film criticism - a mode of analysis that employs the discussed object (the cinematic work) directly -, the video essay quotes the film even as it deconstructs it. It can therefore be easier to grasp without necessarily being simplified as discourse – a seven-minute clip can be as rich and thoughtful as a longform essay – and allows for the survival of intelligent film criticism in a rather dyslexic cultural environment.

The aim of this presentation is to summarize the current state of video essays and their aesthetic and didactic possibilities. In 2017, the history of video essays is simultaneously too short and too long. Since the form is roughly a decade old in popular view, in order to discern its influences, one would have to look beyond the practice itself to examine either the more timeworn tradition of essay cinema – the non-narrative films of Chris Marker, Jean-Luc Godard, Harun Farocki – or the audiovisual histories and TV broadcasts on the subject of cinema – Mark Cousins' *The Story of Film: An Odyssey* or *A Personal Journey with Martin Scorsese through the American Cinema* being popular examples. However, a decade of video-essay-making is also long enough for the form to have experienced its first moments of crisis and for attempts to theorize it to become increasingly difficult and dangerously reductive. For instance, video essays made cca. 2014 were problematic in their over-reliance on voice-over (i.e. audio commentary of the author overlapped with the images), whereas in 2017, being aimed at social media distribution, several of them adopt the irrelevant/muted audio, text-on-screen format, thus placing all the weight on the visual component; faced with the newer pattern, commenters have gone from pleading for less voice-over to asking for more of it. This constantly changing media landscape makes it urgent to develop strategies for

aesthetic evaluation and curation of video essays – otherwise, the overproduction of online content will obscure the best ones and the more provocative possibilities of the form.

Essential (though understated) production guidelines

Due to their ability to quote from film with no need of processing it into a new language, popular video essays are often made from immediately striking fragments: striking film imagery (as in Stanley Kubrick films), dialogues (Aaron Sorkin-scripted one-liners), or even blatant juxtapositions (comparing two stylistically similar films in a split-screen, with the aim of proving just how much the later film borrows from the earlier, usually canonic one). However, their range of subjects largely overlaps with that of cinephile/pop online criticism: overviews of a certain artist's filmography, a certain genre, film festival, national cinema, trend of technical evolution in film craft.

There are already a few prominent plaforms for launching video essays, which provide video-essayists with opportunities (on-the-job training, access to neccesary media) even as they sometimes limit their creative options. The first and already most controversial is the video-on-demand platform Fandor with its annexed publication, Keyframe; others are the BFI/Sight & Sound website; the Netherlands-based platform Filmkrant; MUBI (also annexed to a VOD platform), and the most academic-oriented, [in]Transition (which is more similar to a distributor than a producer, to borrow terminology from the film market).

While there are also video-essayist 'superstars' with distinctive styles, for the sake of brevity, I will only focus on the institutional guidelines which they must follow. Studying these authors' work over several years proves that, even in this seemingly lax working process, shifting editorial demands can have a significant impact on what they produce and how widely it circulates. I would further argue that the formative training of the video-essayists (whether they are filmmakers, critics, academics) is itself only partly relevant to the rigor or whimsicality of their videographic criticism. Although the format is in rapid development and expansion, and making a video essay is hypothetically accessible to anyone who owns a computer and editing software, hierarchies and mandatory style markers can easily be traced among the most well-known video essays made to date, which once again indicates that the total creative freedom of the Internet is merely a utopian dream.

Challenges to the development of video essays

The difficulties of this new form tend to be pragmatic, since the video essays depend on very precarious factors. The first is their survival and continued availability in the online sphere, which the recent Fandor scandal - involving the withdrawal of several hundred video essays - has proved tenuous. The second is the legal circumstance of their right to exist, namely the Fair Use copyright exception: this states that clips of artwork can be used by individuals without permission and copyright ownership as long as the ultimate purpose is different from the straightforward exploitation of the material. A note on Fair Use in the brochure *The Videographic Essay: Criticism in Sound and Image* ends with a disclaimer that they merely offer peer advice – they are not, nor do they claim to be, lawyers.

Video essays as study material

Among the most remarkable feats of video essays is the popularization of film as theory – or audiovisual thinking. As Volker Pantenburg points out in his comparative study of Farocki and Godard, theory has thus far been predominantly linguistic, even when it is self-reflexive and proposes a break with the dominant “amalgam of structuralism, Lacanian psychoanalysis, post-structuralism, and Marxism”. As Pantenburg puts it, “writing against the film theories of the 1970s continues to assume a clear distinction between the films on the one side and their analysis and theorization on the other.”

Similarly, in his 2012 essay *Visualization Methods for Media Studies*, Lev Manovich could be talking about video essays when using terms like “collection montage” and claiming there is a future in visualization of media artifacts when grouping them by intrinsic, yet-unarticulated features: “the most important question, which is still unresolved, is how to combine distant and close readings”. For this, video essays could be a powerful tool of scholarship and a more complex way of conveying information than written language.

Bibliography

Eric Faden, Catherine Grant, Kevin B. Lee, Jason Mittell, *The Videographic Essay: Criticism in Sound and Image*, caboose books

Pantenburg, Volker, *Farocki/Godard: Film as Theory (Film Culture in Transition)*, Amsterdam University Press 2015

Wees, William C. (1993), *Recycled Images: The Art and Politics of Found Footage Films*, Anthology Film Archives

Manovich, Lev (2001), *The Language of New Media*, MIT University Press

Manovich, Lev (2012), *Museum Without Walls, Art History Without Names: Visualization Methods for Humanities and Media Studies*, manovich.net

Witt, Michael (2013), *Jean-Luc Godard, Cinema Historian*, Indiana University Press

Session C

1. The Pyramid of Conscientious Digital Humanities Research: how to get a 'general idea of what you should be seeing'

Serge ter Braake,
University of Amsterdam

'The only way to know if your results are useful or wildly off the mark is to have a general idea of what you should be seeing.'[8]

The question how to cope with a massive number of digital humanities texts, and the tools to process them, has led to publications on 'algorithmic criticism', 'tool criticism' and 'data criticism'. What these publications have in common is the quest for a conscientious way to deal with tools and data, balanced with the humanist domain knowledge and methodologies.[9] Humanities texts can be poems that were written after a sudden burst of inspiration, well crafted texts on the history of an empire, the most inner thoughts of a diary writer or conscientiously crafted bookkeeping accounts of long gone rulers. The field of Digital Humanities tends to treat these texts quite badly. Texts are ripped out of their original contexts, chopped into pieces, linked to other texts, and used for analyses that go far beyond their original intentions.

Depending on the research question of the individual researcher, or research group, this 'text ransacking' is not necessarily a bad thing. Digital Humanities can, should, and does, ask questions that go beyond the scope of texts that could be studied intensely by one human being. There are however, plenty of dangers involved in using digital tools without really knowing what they exactly do. First of all there is the question when we know enough of what a tool does to perform conscientious digital analyses. Secondly there is the question if we keep (enough) in touch with the material we study with digital methods. Where lies the domain knowledge threshold that is necessary to deal with digital data carefully? At what point do we have a 'general idea of what we should be seeing?'

The danger of 'black box tooling' is increasingly getting attention.[10] The dangers of losing touch with the original source material requires some further explanation. For some humanities scholars, digital humanities research mainly extends the work they already are doing: same kind of data, larger approaches. When Father Robert Busa initiated the *Index Thomisticus* in the 1940's, he obviously already was familiar with the work of Thomas of Aquinas. When literary scholars want to study the language use in the works of Jane Austen we may assume they have already read quite a bit of

---

[8] Megan R. Brett, 'Topic Modeling: A Basic Introduction', *Journal of Digital Humanities*, vol 2., nr. 1, Winter 2012

[9] To cite only a few: On 'algorithmic criticism' the slightly dated but still insightful: S. Ramsay, Reading Machines: Toward an Algorithmic Criticism (Chicago 2011). On data criticism: Frederick W. Gibbs and Trevor J. Owens, The Hermeneutics of Data and Historical Writing (2012 revision)', in: Jack Dougherty and Kristen Nawrotzki eds., *Writing History in the Digital Age* (Michigan, 2013); On Tool criticism: S. ter Braake,, A.S. Fokkens, N. Ockeloen and C. van Son, 'Digital History: towards new methodologies' in: Bozic, Mendel-Gleason, Debruyne and O'Sullivan eds., *2nd IFIP Workshop on Computational History and Data-Driven Humanities* (2016).

[10] See for example the Tool Criticism Workshop in Amsterdam: http://event.cwi.nl/toolcriticism/ ; Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, Frank van Harmelen, 'Semantic Technologies for Historical Research: A Survey', *Semantic Web Journal*, Volume 6, Number 6 (2015) 539-564; Ter Braak et al, 'Digital History'.

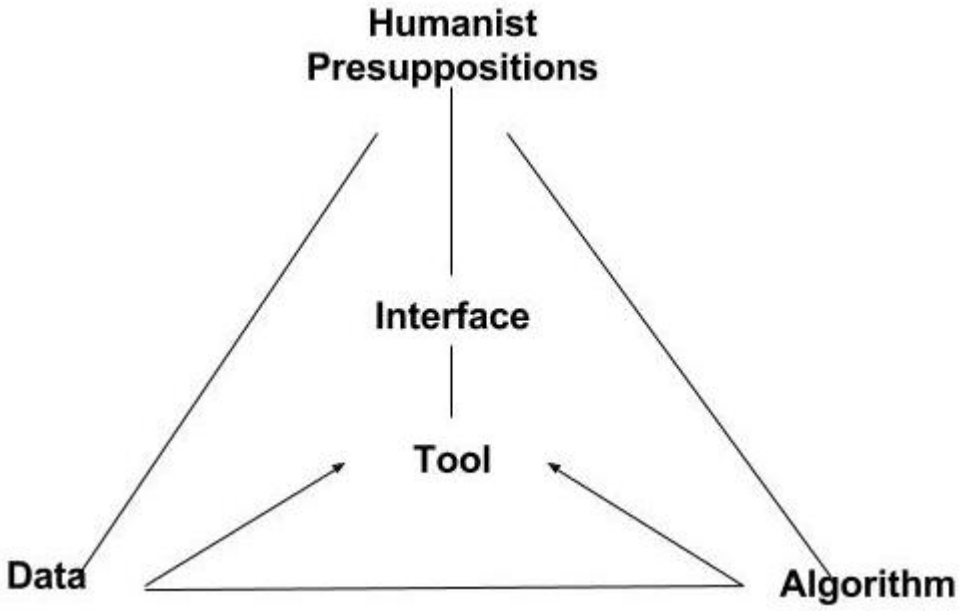
Austen. These scholars certainly already have a general idea of what they could be seeing. When historians use large newspaper archives for digital research however, including different newspapers spanning numerous decades, things become more complex. Historians are often experts on one or several historical topics, with the necessary archival sources attached to them. Few historians are experts on a wide variety of historical newspapers. This problem is enlarged by the way digital tools deal with these newspapers. Text is transformed into 'data', taken away from the page and its surroundings and is transformed together with other pieces of text into an aggregated result.[11]

The questions I want to address here are:

1. When does a researcher know enough of a tool to use it conscientiously?
2. When does a researcher know his material well enough to use digital tools for distant reading analyses?

And finally, springing forth from this:

3. At what point do we decide that the answers to 1) and 2) are not cost efficient anymore? At what point should we decide that a 'simple' tool and close reading practices are more practical for humanist research than complicated tools used on large datasets?



If we want to visualise the interplay between researcher, algorithm, tool, interface and data, then we can come to a pyramid of conscientious digital humanities research, as visualised below. On top there is the humanist researcher, with all of his or her presuppositions acquired from prior knowledge. This researcher will mostly be working with an interface, but also has to understand the tool behind the interface and the data and algorithms behind the tool. If the humanist misses either a sufficient grasp of the computer algorithms, or of the data that is used, the results that are

[11] For example the ShiCo tool, tracing concepts through time: https://github.com/NLeSC/ShiCo. See for reflections on the loss of context C. Jeurgens, 'The Scent of the Digital Archive: Dilemmas with Archive Digitisation', *BMGN - Low Countries Historical Review* 128 (4) 92013) pp. 30–54

provided by the tool through an interface may be misinterpreted, or significant errors may not be spotted.

In short, there should be a 'general idea' of what we could seeing, both by knowing the tool and the data. In this presentation, I will present a proposal, a step-by-step plan, of what could be done to reach this general understanding by taking the example of my own research on concept drift in *De Gids* and *Vaderlandsche Letteroefeningen*, two nineteenth century journals dealing with all kinds of topics of general interest. These steps include: 1) manual close reading; 2) digital close reading; 3) digital analysis; 4) criticism of the results; 5) reflection on steps 1 and 2: were they sufficient? 6) reflections on step 3: was this tool the best to use for this purpose?

When going through this cycle these questions should always be considered: at what point are the requirements for conscientious digital humanities research too high to be worth the effort? At what point is the pyramid too costly? When is it more efficient, and in fact conscientious, to settle for a 'simpler' tool? At what threshold should the digital make room again for more traditional humanities?

---

2. This is my ground truth, tell me yours: Potentials of multiple annotations for digital humanities

Berit Janssen
Meertens Institute, Amsterdam and
Institute for Logic, Language and Computation, University of Amsterdam

Many methods in digital humanities rely on computational methods, which may be trained on a set of reference annotations, also referred to as *ground truth*. However, human judgements are rarely unanimous: this led to research into how information from human judges can be best combined to increase knowledge of the "true" relationships in data (e.g., Dong, 2010). However, in many domains, for instance in music information retrieval, it may be assumed, that multiple annotator judgements may form equally valid interpretations of data such as music similarity or chord estimation (Koops, 2016; Schedl, 2014). The present contribution shows how multiple annotations can be used to reveal human strategies and knowledge by investigating how annotators may agree or disagree on different subgroups in data.

As an example, I present a data-set of annotations on phrase similarity in 360 Dutch folk songs.[12] These folk songs are categorized into 26 groups of variants, or tune families. Three annotators worked independently to give labels to phrases within tune families, or groups of variants. The labels consisted of a letter combined with a number, with which annotators could indicate similarity in three categories: "almost identical" (same letter and number), "related but varied" (same letter but different number), and "different" (different letter and number). The annotators did not agree on phrase similarity at all times, but with Fleiss' κ=0.71 (Fleiss & Cohen, 1973), the agreement was substantial.

The dataset was used to evaluate pattern matching algorithms: these algorithms compared each phrase in the dataset against the melodies within the tune family from which the query phrase was taken, and returned a match score. For evaluation purposes, the three annotations were combined through a majority vote: if two or more annotators had given any phrase in a given variant the same

[12] Available from liederenbank.nl/mtc

label as that of the query phrase, the variant was considered to contain an instance of the phrase, which a pattern matching algorithm should find (cf. Janssen, van Kranenburg & Volk, 2017).

The added value of combining multiple annotations is that next to the evaluation of pattern matching algorithms, also the annotators themselves may be compared to the majority vote. This comparison shows that individual annotators agree around 87% with the majority vote: they miss about 10% of the relevant phrase instances, and find about 10% irrelevant occurrences, as compared with the majority vote. Flexer and Grill (2016) showed how such inter-rater disagreement introduces an upper bound for various tasks in music information retrieval.

The current work presents a way to learn from inter-rater disagreement: the dataset is categorized into tune families, which form homogeneous groups of melodies with high distinctiveness between groups. An analysis of the distribution of disagreement with the majority vote over tune families reveals that individual annotators disagree with the majority vote in different ways, such that some tune families lead to few disagreements for one annotator, but many disagreements for another annotator. This differs from the errors produced by the three-best performing pattern matching algorithms: they show similar trends over the tune families, such that a tune family in which one algorithm produces many irrelevant results will also be more difficult to handle by other algorithms. This suggests that the strategies of the compared pattern matching algorithms may be similar, while the annotators bring different strategies to the table.

References

Dong, X. L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., & Zhang, W. (2014). From data fusion to knowledge fusion. *Proceedings of the VLDB Endowment*, *7*(10), 881-892.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, *33*(3), 613-619.

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. *Journal of New Music Research*, *45*(3), 239-251.

Janssen, B., van Kranenburg, P. & Volk, A. (2017, in press). Finding occurrences of melodic segments in folk songs employing symbolic similarity measures. *Journal of New Music Research*.

Koops, Hendrik Vincent, et al. "Integration And Quality Assessment Of Heterogeneous Chord Sequences Using Data Fusion." *International Society for Music Information Retrieval Conference*. 2016.

Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, *8*(2-3), 127-261.

---

3. Digital History Projects as Boundary Objects

Max Kemman
University of Luxembourg
max.kemman@uni.lu

Digital history is concerned with the incorporation of digital methods in historical research practices. Thus, digital history aims to use methods, concepts, or tools from other disciplines to the benefit of historical research, making it a form of *methodological interdisciplinarity* (Klein, 2014). This requires expertise of different facets, such as history, technology, and data management, and as a result many digital history activities are a collaboration of scholars and professionals from different backgrounds.

Such collaborations would fit Svensson's characterisation of digital humanities as a *fractioned trading zone* (Svensson, 2011, 2012). Simply stated, this means first that digital humanities functions as heterogeneous collaborations, i.e., with participants from different disciplinary backgrounds, and second that the participants act voluntarily.

In this paper, we will investigate these two aspects in the context of digital history to understand how digital history projects function as heterogeneous collaborations, and what the participants' incentives are for entering such collaborations.

We will look at digital history projects as *boundary objects*, a concept developed by Leigh Star and Griesemer to describe an object that maintains a common identity among the different participants, yet is shaped individually according to disciplinary needs (Star and Griesemer, 1989; Star, 2010). This concept could be used for example to refer to the tool under development, or the data on which the tool and historian will work. However, in this paper we will approach the project itself as boundary object; the project binds the participants together, and all participants subscribe to a common description of the project's goals, while at the same time the participants shape the project according to their own needs. As one digital history project coordinator described it in an interview:

> "[Y]ou have a research idea, and you fit that to the call you're applying to, and then you get funding … And if you then hire researchers, yes they too have their own idea of course, and their own line of research they're working on, and they try to fit that in the research project."

This leads us to investigate the *incentives* for collaboration. When writing about interdisciplinary collaboration in digital history, this is almost always done to underscore the positive or even necessary effects (e.g. Eijnatten et al., 2013; Hitchcock, 2014; Sternfeld, 2011). However, such collaboration is not trivial and requires dedication and investments from all involved, e.g. as shown by Siemens (2009; 2012). In order to investigate the activities of individual participants we will follow the work of Weedman on incentives for collaborations between earth scientists and computer scientists (1998). For several digital history projects based in the BeneLux, we have interviewed the participants and inquired about their reasons for joining the project, their individual goals with the project, and the expected effects of their participation after the project has ended. For example, in an interview one historian noted about their project:

> "[W]e're supposed to be advising the team developing the tool. And trying to then carry out research on a specific case study. And so originally it was like wow we're going to be able to use the tool, but very quickly it became clear ok actually probably we're not going to be able to use the tool."

By looking into the incentives of all the participants of a project, we will unpack the trading zones of digital history projects, to gain an understanding of how heterogeneous, interdisciplinary collaborations work, and how participants shape these collaborations. This will allow us to look into why a situation as described above by this historian occurs, and how individual shaping of the project can lead to this. Moreover, we will argue that these incentives go beyond disciplinary boundaries, which means that the trading zone in a digital history project is more complex than the (in)famous Two Cultures as described by C.P. Snow.

This research is part of PhD research on how the interdisciplinary interactions in digital history affect the practices of historians on a methodological and epistemological level (Kemman, 2016). By unpacking digital history projects, we aim to gain better insight in how digital history functions as a coordination of practices between historians and collaborators from different backgrounds, and how individual incentives shape this coordination.

References

Eijnatten, J. van, Pieters, T., and Verheul, J. (2013). Big Data for Global History: The Transformative Promise of Digital Humanities. *BMGN - Low Countries Historical Review*, 128(4):55–77.

Hitchcock, T. (2014). Big Data, Small Data and Meaning. Available from: http://historyonics.blogspot.co.uk/2014/11/big-data-small-data-and-meaning_9.html .

Kemman, M. (2016). Dimensions of Digital History Collaborations. *DHBenelux*. Belval, Luxembourg.

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

Leigh Star, S. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology & Human Values*, 35(5):601–617.

Leigh Star, S. and Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berke- ley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3):387–420.

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

Siemens, L. and INKE Research Group (2012). From Writing the Grant to Working the Grant : An Exploration of Processes and Procedures in Transition. *Scholarly and Research Communication*, 3(1).

Sternfeld, J. (2011). Archival theory and digital historiography: Selection, search, and metadata as archival processes for assessing historical contextualization. *American Archivist*, 74(2):544–575.

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

Svensson, P. (2012). Beyond the Big Tent. In Gold, M. K., editor, *Debates in the Digital Humanities*. University of Minnesota Press, online edition.

Weedman, J. (1998). The Structure of Incentive: Design and Client Roles in Application-Oriented Research. *Science, Technology & Human Values*, 23(3):315–345.

Session D

1. Modelling and Analyzing Character Networks in Recent Dutch Literature

Roel Smeets (PhD candidate)
Radboud University Nijmegen, Department of Literary and Cultural Studies

Keywords: social network analysis, character networks, Digital Literary Studies, Dutch literature

Character relations

When we interpret novels we are influenced by (hierarchical) relations between characters. These relations are not neutral, but value-laden: e.g. the way in which we connect Clarrisa with Richard is of major importance for our interpretation of the gender relations in *Mrs Dalloway* (1925). In literary studies, character relations have therefore lain at the foundation of a variety of critical studies on literature (e.g. Minnaard 2010, Song 2015). A basic premise in such criticism is that ideological biases are exposed in the (hierarchical) relations between representations of certain groups (i.e. gender, ethnicity, social class).

Close reading – the common, traditional method in literary studies – is well suited for fine-grained analyses of the nuances and subtleties of character relations, but falls short when it comes to finding patterns among character relations or testing hypotheses on character relations in larger bodies of literary texts (cf. Stronks 2013).

Social Network Analysis

In computational linguistics, in recent years a broadening range of research has been carried out on the computational analysis of social networks in (literary) texts (e.g. Elson et al. 2010, Karsdorp et al. 2012). On the basis of automated, computational models character relations of all kinds are formalized and mapped in large amounts of texts. Although in its infancy, this branch of research shows that social networks can in fact be reliably extracted automatically from narrative texts (Van de Camp 2016), and relationships can also be classified accurately by computational models trained on examples, e.g. as being romantic (Karsdorp et al. 2015)

The current research project departs from the hypothesis that a computational approach to character relations can reveal (hierarchical) patterns between characters in literary texts in a more data-driven and empirically informed way. In order to test this hypothesis, experiments are being conducted with different forms of social network analysis of characters in a corpus of 170 recent Dutch literary novels. The two major methodological challenges are:

1. to define the nodes that constitute the social network of a novel
2. to define and to weigh the relations between the nodes

The first methodological challenge is about doing a form of character detection: NLP techniques as Named Entity Recognition and Resolution, pronominal resolution and coreference resolution come to mind. However, automatic character detection in literary texts is far from a convenient classification task (Vala et al 2015).

The second methodological challenge is about finding a way to decide when and how two or more characters in a text 'interact'. When Franco Moretti in his famous book *Distant reading* (2013) made a character network of Shakespeare's *Hamlet*, he did that on the basis of occurrences of character X (the addressee) in the lines of character Y (the speaker). Novels are fundamentally different than dramatic plays in that respect: characters in novels usually don't speak to each other in a direct way, and the definition and weighing of character interaction therefore requires a different approach.

Top-down and bottom-up approach

In this talk I will argue that a practical combination of manually gathered data and computational analysis can gain insight in patterns between character relations in recent Dutch literature. Instead of using a bottom-up approach of character detection, I will start top-down using a predefined list of names of characters from each novel in my corpus. Furthermore, I will use manually gathered data from earlier research to ascribe demographic features to the characters that constitute the nodes of the network (Van der Deijl et al 2016). As such, it will be possible to relate demographic backgrounds of characters to their respective place in the character network of the novel. More data are currently being gathered manually from the research corpus: thematic relations as family, friend, lover, colleague and enemy, which will be used to depict the nature of the relations between the characters in the corpus.

I will demonstrate in this talk how manually gathered data (demographic features and thematic relations) can be used for defining both the nodes of the network and the nature of relation between the nodes. Moreover, I will show how a top-down approach based on manually gathered data can be complemented and enriched by a bottom-up, computational analysis of co-occurrences, which will we be used for weighing the relations (or: interactions) between the character nodes. The co-occurrence analysis will consist of precisely delineated textual windows (on the sentence level) in which will be searched for different tokens (variants of names, pronouns) for specific character entities in adjacency with tokens belonging to other character entities.

References

Camp, Matje van de. 2016. *A link to the past: Constructing Historical Social Networks from Unstructured Data*. PhD thesis, Tilburg University (Tilburg School for Humanities).

Deijl, Lucas van der, Pieterse, Saskia, Prinse, Marion & Smeets, Roel. 2016. 'Mapping the Demographic Landscape of Characters in Recent Dutch Prose: A Quantitative Approach to Literary Representation.'In: J*ournal of Dutch Literature* (7:1).

Elson, David, Dames, Nicholas & McKeown, Kathleen. 2010. 'Extracting Social Networks from Literary Fiction'. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010), Uppsala.

Karsdorp, Folgert, Kranenburg, Peter van, Meder, Theo & Antal Van den Bosch. 2012. 'Casting a spell: Identification and ranking of actors in folktales.' In: F. Mambrini, M. Passarotti, and C. Sporleder (eds.), *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-2), pp. 39–50.

Karsdorp, Folgert, Kestemont, Mike, Schöch, Christof, & Bosch, Antal van den. 2015. 'The Love Equation: Computational Modeling of Romantic Relationships in French Classical Drama.'In: *Proceedings of the Sixth International Workshop on Computational Models of Narrative*, pp. 98-107

Minnaard, Liesbeth. 2010. 'The Spectacle of an Intercultural Love Affair: Exoticism in Van Deyssel's Blank en geel'. In: *Journal of Dutch Literature* (1:1).

Moretti, Franco. 2013. *Distant Reading*. London: Verso.

Song, Angeline M.G. 2015. *A Postcolonial Woman's Encounter With Moses and Miriam.* New York: Palgrave Macmillan US.

Stronks, Els. 2013. 'De afstand tussen close en distant. Methoden en vraagstellingen in computationeel letterkundig onderzoek'. In: *Tijdschrift Voor Nederlandse Taal-en Letterkunde* (4).

Vala, Hardik, Jurgens, David, Piper, Andrew & Ruths, Derek. 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*, pages 769–774, Lisbon, Portugal, Association for Computational Linguistics.

2. Spinozist discourse in Dutch textual culture (1660-1720) A computational approach to the dissemination of the Radical Enlightenment

Lucas van der Deijl, University of Amsterdam

Lia van Gemert, University of Amsterdam

Erik van Zummeren, University of Amsterdam
Contact: l.a.vanderdeijl@uva.nl

Key words: Spinozism, Radical Enlightenment, topic modeling, discourse analysis, text mining

Since the linguistic turn, the term 'discourse' has been an important instrument for many humanities scholars (Bové 1995). It has become common practice to study cultural history through the language and discussions in which it was mediated. Currently, the growing availability of digitised historical material provides new ways and scales to study historical discourses, which have been recognised by digital humanities scholars at an early stage (Olsen & Harvey 1988). However, digital approaches to historical corpora face the problem that the often loosely defined term 'discourse' is not easy to formalise. In traditional literary studies, the very lack of definition is inherent to the influential post-structuralist paradigm that reinvented the term, in which meaning is considered 'indefinite' by definition. Within this tradition, discursive elements are measured through both manifest and latent semantic relations, with an equal focus on what is said and what is left out, forgotten or suppressed. Quantitative methods, to the contrary, require a more reductive understanding of what a discourse comprises (e.g. Jockers 2013; Ramsay 2011). They primarily rely on information represented in computationally measurable text elements, which challenges the traditional use of the term. Digital Humanities thus promise new opportunities for cultural history, but also require a critical translation of traditional methodology.

A dominant approach in the study of intellectual discourses focuses on concepts (e.g. Mandelbaum 1965; Lovejoy 2001; Kuukkanen 2008). Philosophers and computational linguists have created models and methods in order to account for conceptual change or drift through time computationally (Betti & Hein 2014; Kenter et al. 2015). Secondly, studies that employ digital text analysis to approach historical discourses often use 'topics' as a representation or indication of discursive patterns in large text corpora (e.g. Nelson 2010). Topic modeling is a useful technology for narrowing down a research corpus into a selection that could be of interest to the researcher. The method also allows tracing the evolvement of dominant themes over time. It is especially useful when the researcher has no strong intuitions about the corpus: the power of topic modeling is its independence from assumptions (Underwood 2012). The use of topics as a measure for 'discourse' in the traditional sense is, however, problematic. A topic is formally defined as a 'distribution [of words] over a vocabulary' and is no more than a set of words that are statistically likely to co-occur in a given text (Blei 2012). A discourse in the Foucauldian sense comprises (historical) values, shared assumptions, 'common sense', associations, automated modes of writing and thinking, which constitute and regulate power relations through language and intertextuality (e.g. Foucault 1977; Bové 1995). When following Foucaults notion of discourse, collocations – the basic linguistic element for topic modeling – could be misleading. The operationalisation of discourses through topics may be intuitive, but is theoretically far from evident.

The study of the dissemination of concepts and discourses is especially relevant in the context of the so-called Radical Enlightenment, a movement of proto-Enlightenment intellectual innovation in which Spinoza played a key role (Israel 2001; Jacob 1981; Krop 2014). As a result of the explosive theological and scientific debates that threatened the stability of the Republic throughout the seventeenth century, radical discourses that challenged orthodox-Calvinist doctrine were firmly suppressed through censorship and prosecution of authors, publishers and printers (Israel 1997). In spite (or because) of this censorship, radical discourses circulated 'underground', in clandestine publications and circuits (cf. Darnton 1982). Many cultural historians have also indicated how authors communicated radical ideas indirectly and ambiguously through literary genres such as novels and pornography (Van Bunge 2003; Elias 1974; Leemans 2002; Wortel 2006). The Foucauldian meaning of 'discourse' as a possible means for the reinforcement of power relations becomes evident during the Radical Enlightenment.

Rather than elaborating on the theoretical difference between topics, concepts and discourses on an abstract level, this paper demonstrates it through a case study. It presents computer assisted discourse analysis as an approach to a specific historical question: how did Spinozist philosophy disseminate into a 'Spinozist' discourse in early modern Dutch textual culture (1660-1720)? In this study, Spinozist philosophy was reduced to a set of characteristic concepts (cf. De Bolla 2013), which were identified through tf-idf[13] frequency analyses and then refined by hand. The concepts were represented as networks of co-occuring words in seventeenth century Dutch translations of eight works written by the philosopher, translated by Pieter Balling (? – 1664) and J.H. Glazemaker (1620-1682) (Thijssen-Schouten 1967; Steenbakkers 1999).[14] These conceptual networks were used as a measure to identify Spinozist 'discourse' in a corpus of 500 texts published between 1660 and 1720. For pragmatic reasons, the vocabularies were assumed to be stable, but this paper addresses possible advancements based on the literature on conceptual and linguistic drift (Betti & Hein 2014; Kenter et al. 2015). Also, conventional procedures applied in computational intellectual history were modified in order to reduce the problems caused by spelling variation in historical Dutch (e.g. in Herbelot et al. 2012; Tangherlini & Leonard 2013).

The results obtained through the concept-orientated 'top down' approach are contrasted with a more 'bottom up' transformation of the corpus based on topic modeling. This paper evaluates the differences between both approximations of Spinozist discourse and shows how Spinozist texts unknown to the computer were successfully identified and described. Based on these results, it formulates a working hypothesis on the dissemination of Spinozist discourse in Dutch textual culture and advances the debate on the resonance of (Radical) Enlightenment ideas with computational results (Darnton 1982; Israel 2001; Leemans 2002; Edelstein 2010 etc.).

References

Betti, A. & H. van den Berg, 'Modelling the History of Ideas'. *British Journal for the History of Philosophy* 22 (2014) 4: 812-835.

Blei, D., 'Probabilistic Topic Models'. *Communications of the ACM* 55 (2012) 4: 77-84.

Bolla, P. de, *The Architecture of Concepts. The Historical Formation of Human Rights*. New York 2013.

[13] 'term frequency – inverse document frequency'.

[14] *Korte verhandeling van God, de mensch en deszelvs welstand* (1660-1661); *Renatus Des Cartes Beginzelen der wysbegeerte, I en II bewezen* (1664); *Aanhangzel, over-natuirkundige gedachten* (1664); *Handeling van de verbetering van 't verstant* (1667); *Zedekunst, In vijf delen onderscheiden* (1677); *Brieven Van verscheide geleerde Mannen Aan B.d.S* (1677); *Staatkundige verhandeling* (1677); *De Rechtzinnige Theologant, of godgeleerde staatkundige verhandeling* (1693).

Bové, P.A., 'Discourse'. In: F. Lentricchia & T. McLaughlin, *Critical Terms for Literary Study*. Chicago 1995: 50-64.

Bunge, W. van, 'Philopater, de radicale Verlichting en het einde van de Eindtijd'. *Mededelingen van de Stichting Jacob Campo Weyerman* 26 (2003): 10-19.

Darnton, R., *The literary underground of the Old Regime*. Cambridge (MA) 1982.

Elias, W., 'Het spinozistische erotisme van Adriaan Beverland'. *Tijdschrift voor de Studie van de Verlichting* 2 (1974): 283-320.

Edelstein, D., *The Enlightenment. A genealogy*. Chicago 2010.

Foucault, M., 'The Archeology of Knowledge and the Discourse on Language'. Trans. A. Sheridan. New York 1977.

Gemert, L. van, 'Stenen in het mozaïek. De vroegmoderne Nederlandse roman als internationaal fenomeen'. *Tijdschrift voor Nederlandse Taal- en Letterkunde* 124 (2008) 1: 20-30.

Herbelot, A., E. von Redecker, J. Müller, 'Distributional techniques for philosophical enquiry'. *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Avignon 2012: 45-54.

Israel, J., 'The banning of Spinoza's works in the Dutch Republic'. In: C. Berkvens-Stevelinck e.a. (red.), *The emergence of tolerance in the Dutch Republic*. Leiden 1997.

Israel, J., *Radical Enlightenment*. New York 2001.

Jacob, M.C., *The radical Enlightenment. Pantheists, freemasons and republicans*. London 1981.

Jockers, M., *Macroanalysis. Digital Methods and Literary History*. Urbana 2013.

Kenter, T.M., M. Wevers, P. Huijnen & M. de Rijke, 'Ad Hoc Monitoring of Vocabulary Shifts over Time'. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. Melbourne 2015.

Krop, H., *Spinoza. Een paradoxale icoon van Nederland.* Amsterdam 2014.

Kuukkanen, J.M., 'Making Sense of Conceptual Change'. *History and Theory* 47 (2008): 351-372.

Leemans, I., *Het woord is aan de onderkant. Radicale ideeën in Nederlandse pornografische romans 1670-1700*. Nijmegen 2002.

Lovejoy, A. O., 'The Historiography of Ideas'. *Proceedings of the American Philosophical Society* 78 (1938): 529-543.

Lovejoy, A.O., *The Great Chain of Being. A Study of the History of an Idea*. Cambridge, MA / London 2001 [1964].

Mandelbaum, M., 'The History of Ideas. Intellectual History, and the History of Philosophy'. *History and Theory* 5 (1965): 33-66.

Nelson, R.K., 'Mining the Dispatch', 2010. [http://dsl.richmond.edu/dispatch/pages/home]

Olsen, M. & L.G. Harvey, 'Computers in Intellectual History: Lexical Statistics and the Analysis of Political Discourse'. *The Journal of Interdisciplinary History* 18 (1988) 3: 449-464.

Ramsay, S., *Reading Machines. Towards an Algorithmic Criticism*. Urbana: University of Illinois Press, 2011.

Siebrand, S.J., *Spinoza and the Netherlands. An inquiry into the early reception of his philosophy*. Dissertation Rijksuniversiteit Groningen 1980.

Steenbakkers, P.M.L., 'Benedictus de Spinoza. Een overzicht.' *Filosofie* 9 (1999) 6: 4-14.

Tangherlini, T.R. & P. Leonard, 'Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research'. *Poetics* 41 (2013) 6: 725-749.

Thijssen-Schouten, C.L., *Uit de Republiek der Letteren. Elf studiën op het gebied der ideeëngeschiedenis van de Gouden Eeuw*. Den Haag 1967.

Underwood, T., 'Topic modeling made just simple enough'. Online 2012. [https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/]

Wortel, D., 'Vrouwen in mannenkleren en Spinoza. De *Kloekmoedige Land- en Zee-Heldin* (1682) als verpakking van de filosofie van Spinoza'. In: *Spiegel der Letteren* 48 (2006): 27-55.

Session E

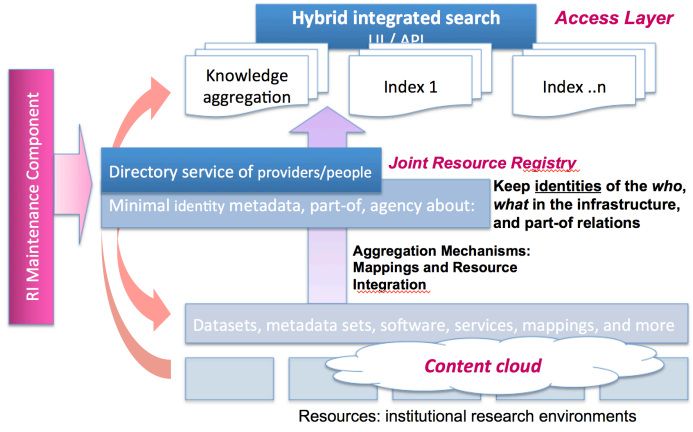
1. Building a Conceptual Architecture and Data Model to address the Sustainable Data Integration Problem

George Bruseker, Maria Theodoridou, Martin Doerr (ICS-FORTH)

Research Infrastructures (RI) seeking to provide a unified resource set to their user community tend to begin with the elaboration of a new model for unifying a domain of discourse and then seek out the institutional and political support to undertake mappings to the defined common structure. These projects are undertaken with the critical aim of facilitating broad resource access within the domain of interest. Such projects, however, notably face strong challenges both in terms of defining an adequate model and, then, in sustaining a mapping and aggregation process which is unavoidably time consuming and expensive. While such resource integration projects undoubtedly serve a crucial role in research environments, an essential aspect of this process seems to be consistently overlooked. Data are fundamentally heterogenous in nature - a state that cannot be avoided - and are in a process of continuous potential or actual change. Further, actors managing resources change composition, status and activities. This quickly creates the potential for obsolesence of any integrated data environment as the indexed resources inevitably change.

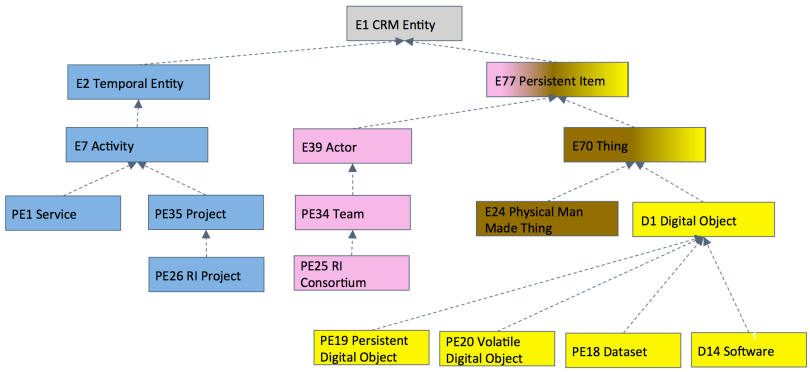
It seems, then, that value can be had from a new approach that focuses on making integration sustainable and useful in the long run by modelling and managing the integration process itself. By modelling this meta meta level and providing a data structure for the tracking of the same, we argue, it is possible to provide the necessary management structures for building ground up and on-demand aggregation which will meet the aims of this process both in the present and into the future. This paper will outline the proposal of a new conceptual architecture to support highly scalable integration activities for developing ever more integrated pools of resources and a conceptual model capable of representing the data required to drive this process.

The proposed conceptual architecture has at its core a registry that is a logically if not physically distinct data structure that holds data pertaining to the activities of RIs and their members themselves, the resources they provide and the manner in which they do so. The registry maintains the picture of who has and does what and where resources are, as well as their level of compatibility with other resources. The data requirements of this registry are extremely light in order to form as little a barrier as possible to participation in such a service by potential partners. The basic functional relationships that are tracked to allow the long-term management and control of resources are: part-of, metadata-of and indexed-by. Additional metadata is only requested in order to help disambiguate entities in the registry and to support its readability by the operators. In the proposed architecture, source metadata and data as well as their multiple mappings remain in a content cloud which can be either data held by trusted providers who guarantee their maintenance or, otherwise, can be copied into a stable storage facility at the time of registration. The registry has the intention to enable decisions with regards to the management of data, based on the high level view of resources, wherever they may reside across the data cloud. Such decisions could include: identifying datasets for an integration, identifying gaps in coverage, connecting orphaned datasets to



appropriate curators, following up with service providers with regards to availability/quality of service etc.

In order to support the proposed architecture, it is necessary to propose a new conceptual model describing integration processes themselves. This is the function of the Parthenos Model. Built off an analysis of the registries of existing RIs, it aims to model the fundamental resources and relations that are of interest to manage in integration. Identified through this process were a number of fundamental entities the study of whose relations drove the model development. These are: services, project, datasets, software, and actors. What was of interest in the model was to understand the nature of these objects not as such but as they play a role within RIs. Taking this scope into account allowed for strong analytic distinctions of the high level entities of interest delivering a compact model of +- 38 classes and 50 relations.



Particular modelling challenges include defining the functional role of services and collections. Service plays a central, if often overlooked, role in RI discourse. It is what binds assets to actors and allows for effective communication between agents on a scientific and technical level. A particular challenge was to model service beyond the scope of e-services and to understand the full range of its meaning. This lead to the definition of service as a willingness and ability for someone to take action to the benefit of some other agent. Modelling service at this generic level and then providing high level classes for hosting, curating and e-services allows a highly flexible description of the various kinds of service RIs provide to their members, notably including non-IT related services. Another particular modelling challenge faced was to address the perennial question of what constitutes a 'collection'. The fact of the plurality of an object is easily modelled through part of relations, but this misses an aspect of the phenomenon that 'collection' tries to express. Consideration of this question in relation to the context of service allowed for a highly useful new conceptualization, distinguishing persistent and volatile digital objects. The former are static information objects whose identity is fixed at the bit level and have an objectively identifiable existence over time from their structure. A volatile digital object, however, has no fixed identity in itself, since it undergoes continuous change and modification. It inherits an identity from the fact that it is an object under curation, the activity of a curation service, undertaken with some specific plan. By making reference to the service of curation and its plan, we can identify volatile digital objects or 'collections' over time.

The proposed shift in focus from domain modelling to modelling of RI integration processes themselves is currently being tested within the Parthenos Project where the architecture and model are being implemented. The model is being developed and validated through an iterative process of mapping from the participating RIs registries to the model for integration in the registry. The mapping process is being undertaken using the X3ML toolkit for writing declarative mappings. Once populated the registry will be used to get an overview of the integrated resource capacities of the joined RIs and determine appropriate deep level integrations. The technologies to run the aggregation and the subsequent VREs are provided through the GCube and D4Science systems. To date, the model has shown itself robust against basic revision and flexible enough to describe this high-level management picture of integration.

2. Improving data quality in Europeana by designing extensive EDM records - The Universitätsbibliothek Heidelberg study case

Pierre-Edouard Barrault, Valentine Charles, Antoine Isaac (Europeana Foundation, Prins Willem-Alexanderhof 5, 2595 BE, The Hague, The Netherlands)

Introduction

For this paper, we have worked on improving the results of mapping process from the METS[15] to EDM[16] schemas, for metadata records associated with cultural heritage objects. We chose to present the case of the Universitätsbibliothek Heidelberg[17], which was founded in 1386 and is Germany's oldest university and one of the world's oldest surviving universities. Its magnificent collection of about 25000 records[18] contains parchments[19] and early printed books from the 14th century until Modern Age, or books, magazines and newspapers from the 19th and onward, in various languages including French[20], German, Italian or Spanish. It is without any doubt a solid accomplishment for an old book digitization project, demonstrating the value added from respecting both content integrity thanks to high digitization standards coupled with the IIIF framework, and informational quality through rich, highly-structured, open data. In addition, the institution proposes its collection under the Creative Commons - Attribution, ShareAlike (BY-SA) open license, allowing for free re-use[21].

On the other hand, the Europeana Collections[22] is an European platform partnering with cultural institutions to centralize, in an open online database, all metadata and content related to cultural heritage objects available across Europe. The platforms acts as a search engine to explore these collections, offers a set of curated channels focused on specific thematics, and also makes several Web services available that can be used by developers, creatives and researchers for tackling and re-using digital cultural resources.

Previously to this experiment, the collection of the Universitätsbibliothek Heidelberg in Europeana was based on harvests of the OAI-PMH server of the institution exposing metadata under the ESE schema. We used to receive limited metadata records in which multiple values for a given field were mapped in only one instance of this field. Fields such as dc:date, dc:type and dc:subject were biased. Having single strings introduced in a single metadata field with separators prevents the Europeana automatic semantic enrichment from detecting the appropriate string and enriching the record based on the matching string. Other shortcomings were based on the lack of language attributes or relevant hierarchical data.

---

[15] See http://www.loc.gov/standards/mets/mets-schemadocs.html

[16] See http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation

[17] See http://www.uni-heidelberg.de/index.html

[18] Europeana records for this institution http://www.europeana.eu/portal/en/search?view=grid&q=PROVIDER%3A%22Universit%C3%A4tsbibliothek+Heidelberg%22&per_page=96

[19] See *Heidelberger Schicksalsbuch (Heidelberg Book of Fate)*, 1491 http://www.europeana.eu/portal/en/record/07932/diglit_cpg832

[20] See *Le Sifflet: journal humoristique de la famille (Le Sifflet: humorous family newspaper), 1872* http://www.europeana.eu/portal/en/record/07931/diglit_sifflet1872.html?q=PROVIDER%3A%22Universit%C3%A4tsbibliothek+Heidelberg%22

[21] See http://creativecommons.org/licenses/by-sa/4.0/

[22] See http://www.europeana.eu/portal/en

IIIF implementation

We focused our work on this specific provider with the hope for improving its collections, which were already available in Europeana Collections, with the IIIF[23] features they had implemented on their side. This open technological framework can be implemented within content management systems to enable deep visualisation features (zoom, crop, effects), and to make image sharing easier on the Web.

The main target of this experiment was about implementing IIIF metadata elements, which were not present in previously submitted data from this institution to the Europeana Collections database. After investigating the available data on the institution's side, we decided to harvest METS records as this was a much richer metadata source, regarding both IIIF core elements and metadata range and quality.

Data quality

Even if metadata improvements are not always obvious on a result page in the Europeana Collections portal, they nevertheless have a strong impact on search and overall findability. Ingestion of reliable data therefore participates in ensuring a cohesive experience for its users, from

In the case of digital cultural heritage, qualitative datasets could be defined as ensemble of standardised (such as LOD resources), granular, specific, relevant and consistent metadata, associated with high quality visualisation standards. The nature of the records itself should obviously be in consideration when defining the overall strategy. For instance, OCR[24] techniques would make sense in the case of text documents while focusing on high digitization standards would better suit photographs. Data quality is yet critical to support users focused discovery scenarios[25], and long-term strategy to improve it should be considered *de facto* by any cultural institutions, as a leverage to reach a wider audience.

By using another metadata source from The Universitätsbibliothek Heidelberg, we refined and improved the overall data quality by relying on Linked Open Data resources from the GND authority vocabulary maintained by the German National Library[26], which were available in the original METS records. We therefore included, as systematically as possible, the provided URIs of resources related to agents, concepts and places. This approach follows LOD implementation best practices: only links to resources are provided in the ingested records, and then Europeana de-references them, fetching all the available metadata for each provided URI.

We also applied stricter conditions to the mapping in order to preserve the semantic precision and granularity of the original data as much as possible. This was done by choosing more specific metadata fields, and rejecting irrelevant ones. We focused on core metadata elements related to typology, format, temporal and geographical information. We also created an *ad hoc* description field in order to provide more physical location information to users.

Further normalization was done for agents related to these records (e.g. creators and contributors), which were previously sent without any role distinction. We disambiguated the mapping of these

---

[23] See http://iiif.io/

[24] See https://en.wikipedia.org/wiki/Optical_character_recognition

[25] Most of Europeana users rely on the search functionality, and 59% of them use extra filtering options to refine their search. More than half of the users search items based on specific geographical location. *(Source: Europeana Collections Online Survey, April 2016)*

[26] See http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html

elements using the MARC Relators codes[27] originally embedded in the METS records, such as *"aut"* that represents *"Author"*. The codes were used to identify the agents as creators or contributors, and then were normalized into strings to be directly incorporated into the resulting EDM records as additional metadata.

Finally, hierarchical relationships that were not made available in the original conversion were represented in the new metadata. We focused on records for individual journals encompassed in bigger volumes, and mapped the relevant metadata - references to parent and children records - within hierarchical fields. This enabled a better experience for end users thanks to the display of a widget dedicated to browse hierarchical resources by following their cardinality or their appartenance.

Results

The first outcome of this work is an extensive report presenting this study case, standing as data guidelines available in the Pro section of Europeana Collections[28]. However, our results rely on both qualitative and quantitative achievements.

The overall data improvement empowers the Europeana users - creatives, searchers, curious - with higher quality results, allowing them to tailor their experience even further from the main public access. Specific data reuse or data mining scenarios also benefit from such experiment, thanks to the Europeana's REST API[29]. In addition, the compatibility with the IIIF framework ensure a seamless user experience carried out through extended visualisation features. This can be transposed into more advanced applications by directly reusing the aggregated IIIF metadata from Europeana, e.g. within Digital Humanities visualisation projects.

Finally, the updated datasets didn't necessarily grow in size, records wise. But instead of the former 1 thumbnail per record rule (for about 25K records), the newly added IIIF metadata enables the Europeana's viewer to fetch now more than 3.5M high-resolution pictures (+1600px wide) from all the connected JSON manisfests[30].

---

3. Easing Access to Linked Data Resources for Digital Humanities Scholars

Albert Meroño-Peñuela[1] and Rinke Hoekstra[1,2]
[1] Computer Science Department, Vrije Universiteit Amsterdam, NL
{albert.merono, rinke.hoekstra}@vu.nl
[2] Faculty of Law, University of Amsterdam, NL

Abstract.

Semantic Web technology comprises a variety of languages, standards and practices that, over the last two decades, has facilitated the emergence of the Linked Open Data (LOD) Cloud – a global Web graph of more than 100 billion interconnected statements [1]. Datasets in this LOD cloud cover a

---

[27] See http://www.loc.gov/marc/relators/

[28] See http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies/the-universitaetsbibliothek-heidelberg-case-study

[29] See http://labs.europeana.eu/api/introduction

[30] See http://iiif.io/api/annex/notes/jsonld/#greedy-compaction-of-terms

variety of domains, including geography, government, life sciences, linguistic, media, publications and social networking. Despite this success integrating data on the Web, Semantic Web technology is still very present at every level of the LOD cloud. This includes the early layer of accessing Linked Data; this is, the mechanism by which users select and grab the data they consider for their applications or analyses. Accessing Linked Data requires certain technical skills –mostly involving understanding of the Resource Description Framework (RDF) [6] and the SPARQL [7] query language, but also others such as SQUIN [3] or Linked Data Fragments [8]– that very often exclude potential users. In the digital humanities, many scholars lack this technical knowledge, and consequently miss a great deal of LOD sources of their interest. This includes, but is not limited to, multiple linked datasets on historical statistics (e.g. CEDAR [2], CLARIAH [4]), museum collections (e.g. Amsterdam, British Museum, Smithsonian), linguistic resources (e.g. lexinfo, BabelNet), and media (e.g. MusicBrainz, BBC, New York Times, Linked Movie Database)). Although these scholars are becoming more and more tech savvy, deep knowledge of technology should not be a strict requirement for accessing Linked Data. In order to address this issue, we propose grlc [5], an Linked Data accessing server that uses SPARQL queries stored anywhere on the Web to generate comprehensive, well documented, neatly organized, and provenance-trusted API specifications. Such APIs make any Linked Data actionable, making access to Linked Data sources easy, repeatable and shareable with one single URI entry point. grlc relies on the Swagger UI[31], an OpenAPI[32] frontend, to present these APIs to the user as an intuitive user interface. In this demo, we will show how grlc can help on easing the traditionally high technical requirements to access Linked Data. We will illustrate this with several running use cases in CLARIAH[33], a Dutch national project to build digital infrastructure for the humanities.

Keywords: Linked Data, API, REST, SPARQL, #LD, Web Data access, middleware, OpenAPI

References

1. Abele, A., McCrae, J.P., Buitelaar, P., Jentzsch, A., Cyganiak, R.: Linking Open Data cloud diagram. http://lod-cloud.net/ (2017)

2. CEDAR Project, http://www.cedar-project.nl/

3. Hartig, O.: Squin: A traversal based query execution system for the web of linked data. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 1081–1084. SIGMOD '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2463676.2465231

4. Hoekstra, R., Meroño-Peñuela, A., Dentler, K., Rijpma, A., Zijdeman, R., Zandhuis, I.: An Ecosystem for Linked Humanities Data. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016), ESWC 2016 (2016)

5. Meroño-Peñuela, A., Hoekstra, R.: grlc Makes GitHub Taste Like Linked Data APIs. In: The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016, Revised Selected Papers. pp. 342–353. Springer (2016)

6. The World Wide Web Consortium (W3C): Resource Description Framework (RDF). http://www.w3.org/RDF/

[31] See http://swagger.io/swagger-ui/

[32] See https://www.openapis.org/

[33] See http://www.clariah.nl/en/

7. The World Wide Web Consortium (W3C): SPARQL Query Language for RDF. http://www.w3.org/TR/rdf-sparql-query/

8. Verborgh, R., Sande, M.V., Colpaert, P., Coppens, S., Mannens, E., van de Walle, R.: Web-Scale Querying through Linked Data Fragments. In: Proceedings of the 7th Workshop on Linked Data on the Web (LDOW2014), WWW2014 (2014)

Session F

1. The Nederlab research environment: an update

Hennie Brugman& Antal van den Bosch
Meertens Institute, Amsterdam
hennie.brugman@meertens.knaw.nl

Nederlab[34] (Brugman, 2016) is a five year long 'NWO groot' project building a research infrastructure for primarily historians and literary, linguistic and cultural scholars. Building this infrastructure involves activities in three main tracks:

1. Acquisition, harmonisation/semantic mapping, text enrichment and metadata curation of a substantial number of existing (historical) Dutch digital text collections of our academic and cultural heritage partners in the Benelux.
2. Improving the quality of the output of existing language processing tools when they are applied to historical Dutch texts from 800 until present.
3. Building a virtual research environment with a powerful search backend for exploration, search and analysis of metadata and annoted text from our very large aggregated and integrated collections (Brouwer, 2016).

We are currently in the last year of our project. Therefore, in our contribution we would like to take the opportunity to evaluate to what extent we have been able to implement our original, ambitious, project use cases. We intend to support this evaluation with a demonstration at DHBenelux 2017.

In general, we expect to have processed between twenty and thirty collections by the end of our project and to have made those available to the research community. At the moment of writing this, we have reached a total of almost ten billion words of annotated text, accessible through our online Virtual Research Environment, the 'research portal'[35]. During the last year of our project we are carrying out a number of scientific pilot projects in an open call, to test the usability of this VRE and the Nederlab collections, and to add extensions based on real user needs.

Below we will zoom in on our original categories of use cases.

1. Detecting the onset of change

When do new concepts occur for the first time? Or new wordforms? Or word combinations (collocations)?

By the end of our project we will have collection data for all periods between 800 and present time, thereby enabling full diachronic searches. Our Nederlab research portal is able to visualise time distributions over all hits found for specific queries, both document and hit counts and showing absolute as well as relative frequencies (for example, show the number of occurrances of 'vliegtuig' -airplane- for each year). The system supports complex queries for sequential patterns over multiple parallel annotation layers using the Corpus Query Language (2), a query language introduced by the Corpus WorkBench (CWB) and regularly used in our domain (e.g. by Sketch Engine, MTAS, BlackLab). Nederlab uses Multi Tier Annotation Search (MTAS) [36]. Searching for patterns using CQL, in combination with grouping of results enables researchers to investigate word combinations and how

[34] www.nederlab.nl

[35] www.nederlab.nl/onderzoeksportaal

[36] https://meertensinstituut.github.io/mtas/

often they occur, for specific periods in time. For example, it is possible to query for the most frequent nouns used in sentences containing the lemma 'varen', for each century, to investigate potential shifts in meaning over time (in this case from 'go' to 'go by boat').

2. Establishing the spread of changes

How do such changes spread, over time, over places, from one text type to another, from one author to another?

Our system allows users to search for words or patterns and visualise the results as distributions over many metadata dimensions, even over multiple dimensions simultaneously (e.g. time and genre). It is also possible to directly compare time distributions for different search terms simultaneously (using a 'trends' visualisation, e.g. 'mensch' versus 'mens') (Tjong Kim Sang, 2016).

3. Finding connections and networks

Find and investigate motives using semantic word fields around concepts. Establish relations between persons and places.

We currently already support expansion of queries with historical variants using a web service built around the Dutch historical lexicon by the Instituut voor de Nederlandse Taal (INT). We intend to generalize and extend this query expansion mechanism to include semantic expansion and expansion with user defined domain lexica. We will do this in collaboration with a number of our ongoing scientific pilot projects. An example of such a domain lexicon is a semantic lexicon containing emotion words.

Networks of persons and places can be charted on basis of the named entities that were added to our corpus during the enrichment process. We use CQL searching in combination with grouping functionality to do this (e.g. list the most frequently mentioned persons in sentences or paragraphs containing the location 'Deventer').

4. Detecting similarities and differences between texts

Investigate reuse of text fragments among authors. Compare texts or text collections with corpus analysis tools.

For individual texts or for any subcollection of texts from our complete corpus we can query for statistics. We can determine total numbers of documents, tokens and types, but also mean and median number of words per document, in fact, our system can return complete word count distributions that can be directly visualised. Other statistics that are supported: numbers of sentences, paragraphs, divisions, heads, frequency lists over words or over any of the annotation layers, for any subcollection of our corpus. All of these statistics and lists can in principle be used to compare text documents or complete document collections. All statistics can also be exported for further analysis in external tools, like for example R.

Conclusion

After a number of years of constructing the foundations of our infrastructure, the project is now at a stage where we can start using it for real research pilots or projects. Although there is substantial room for improvement on many aspects of our products, our initial aims are within reach.

References

Brouwer, Matthijs, Hennie Brugman, Marc kemps-Snijders (2016). 'MTAS: A Solr/Lucene based Multi Tier Annotation Search solution', *CLARIN Annual Conference 2016*, Aix-en-Provence, France, 26-28 October 2016.

Brugman, Hennie, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, Antal van den Bosch (2016). 'Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora', in: *Proceedings of LREC (10th edition of the Language Resources and Evaluation Conference,* 23-28 May 2016, Portorož (Slovenia), pp. 1277-1281.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of COMPLEX '94: 3rd Conference on Computational Lexicography and Text Research, Budapest)*.

Tjong Kim Sang, Erik (2016). 'Finding Rising and Falling Words', In: *Proceedings of the COLING 2016 workshop Language Technology Resources and Tools for Digital Humanities*, ACL, Osaka, Japan, 2016. http://ifarm.nl/erikt/papers/lt4dh2016.pdf

---

2. Modeling the evolution of languages through text mining

A proposed methodology applied to the transition between Latin and romance vernaculars

Florian Cafiero and Remy Verdo

The mechanisms at stake in the passage from a "dilated diasystem", where a language becomes more and more complex, to a "disconnected diasystem", where two distinct linguistic systems appear in the same cultural system, seem to be a well-studied problematic.

For instance, several models have been presented to describe the evolution from Latin to romance vernaculars in the past decade. The first model proposed to address this question is Ernst Pulgram's (Pulgram, 1950: 462). In this pioneering work Latin, language is however represented according to the traditional written vs. oral distinction, and does not allow a very detailed analysis. Its deterministic approach might also lead to some inaccuracies, the language being considered as always further from "old Latin" as the time goes by. In 1986, Walter Berschin (Berschin, 1986: 148) proposed a more comprehensive modeling. Berschin proposes a two-sided diachronic modeling. The concept of vulgar Latin is more refined here, as it includes both written and spoken language. Yet, here too, vulgar Latin is separated from literary Latin, the "styl istical level" (*Stilhöhe*) of which, even when it is at lowest, never crosses, or even touches, the curve of oral language. What is more, this Vulgar Latin is supposed to evolve linearly, as in Pulgram's works. The curve modeling literary Latin seems to represent the sole evolution of the higher registerof language that is observed. It disregards the co-existence of different registers of language in literary Latin, and ignores their articulation to vulgar Latin. Last, we can only regret the absence of data taken from "diplomatic" texts, in which stylistic and pragmatic efforts are also to be noticed.

Hence, those studies raise a few problems. They do not address well the problem of registers, using very broad distinctions, and forgetting the possibility that different language registers could be used at the same time, even in the same text. They also are "experts view", based on the author's extensive experience, rather than on a systematic analysis of the texts.

We thus propose a methodology to systematically study the evolution of a language from a form to another, taking into account our remark on registers. This methodology involves computerized statistical analysis and artificial intelligence but should not be seen as an automated process disconnected from the linguist's analysis. On the contrary, it has been designed as a way to extend the way of thinking of a particular expert. It enables to partially re-create his own point of view, and to apply to a large amount of text, that would take too long to analyze otherwise.

The **first step** consists in **"traditional" linguistic analysis on a selection of texts** , aimed at differentiating several registers used inside the texts of one's period of interest.

Our sample corpus consists in three hagiographical texts and twenty-one diplomatic texts. Our three hagiographical texts were written in later Merovingian or in early Carolingian ages (ca. 650-780), then rewritten during the Carolingian Renaissance (from 780 to the death of Charles the Bald in 877, or so). The diplomatic texts are 21 original Frankish royal charters dating from ca. 665 to 868. Most of them are accounting for a judgment. Originally part of the great collection of the monastery of Saint-Denis, they are kept in the French national Archive.

We isolate five language registers in this sample corpus, consistent with Michel Banniard's works (Banniard,2008), and we design a table of criteria to characterize them.

We then go through a **calibration phase**. We try to apply various computing methods that can help isolating different language registers used in the various texts of the corpus - or inside on text of the corpus. This initially calls for unsupervised methods, as we would not want to influence the computations' outcome. The statistical analysis could reveal divisions we ignored, highlight unnoticed phenomena... We try to implement clustering algorithms such as ***k-means, hierarchical clustering,*** and various ***neural networks.*** We then compare the performance of those algorithms with ***supervised algorithms***, where our sample corpus is used as **training data** .

Crucial for those analysis is the way we choose to present the texts to our algorithms. Lemmatizing the texts would remove too much information. Here, even small variations, such as written form variations, are likely to be significant. It can sometimes be even more significant than the grammatical structure of the texts itself. This is why we apply our computations to two types of version of our corpus' texts. In the first versions, the texts are treated as a list of **words** , without any further treatment, or with a selection of the most frequent words. In the second versions, the texts are treated as **n-grams** (for 8>n>3), without any further treatment, or with a selection of the most frequent forms. N-grams can demonstrate great performance here, as they allow to take implicitly into account the structure of the sentences - here, which word comes after which.

We compare all those findings with our own "expert" model designed on our sample, and select the solution that gives the most accurate division in registers.

We then run the selected algorithm on an extended corpus , formed by a large selection of texts written during the same period (650 - 877). We then follow the register's evolution across time on this broader corpus. We then conclude on the global consistence of these results with the model we designed by analyzing our first sample.

BIBLIOGRAPHY

Michel Banniard, « Du latin des illettrés au roman des lettrés : la question des niveaux de langue en France (viii^{e} -xii^{e} siècle) », in *Zwischen Babel und Pfingsten : Sprachdifferenzen und Gesprächsverständigung in der Vormoderne (9.-16. Jh.) : Akten der 3. deutsch-französischen Tagung des Arbeitskreises « Gesellschaft und*

individuelle Kommunikation in der Vormoderne » (GIK) in Verbindung mit dem Historischen Seminar der Univ. Luzern, Höhnscheid (Kassel), 16-19 nov. 2006, Peter von Moos ed., Münster, 2008 (« Gesellschaft und individuelle Kommunikation in der Vormoderne », 1), p. 269-286.

W. Berschin, Biographie und Epochenstil im lateinischen Mittelalter, Stuttgart, t. 3 : Karolingische Biographie, 750-920, 1991.

Piera Molinelli, « Per una sociolinguistica del latino », in *Latin vulgaire – latin tardif : actes du VII^{e} colloque international sur le latin vulgaire et tardif (Séville, 02-06 septembre 2003),* éd. Carmen Arias Abellán, Séville : Universidad de Sevilla, 2006, p. 463-474.

Giovanni Polara, « Problemi di ortografia e di interpunzione nei testi latini di età carolina », *Grafia e interpunzione del latino nel Medioevo (Roma, sept. 1984)*, éd. Alfonso Maieru, Rome, 1987.

Ernst Pulgram, « Spoken and written Latin » , *Language. Journal of the Linguistic Society of America,* t. 26, 1950.

3. Experiments in fine-grained entity typing for Dutch

Marieke van Erp and Piek Vossen
Computational Lexicology and Terminology Lab, Vrije Universiteit Amsterdam

Introduction

Many entity recognition approaches classify recognised entities into a limited set of coarse-grained entity types [1]. However, fine-grained entity types are more useful for deeper natural language analysis and end-user tasks, in particular in the digital humanities domain where entity linking (grounding an entity in a knowledge base) is not possible. For example, while standard named entity recognition may determine that an entity is a person knowing whether that entity is a writer or a politician is important for populating a database of persons with particular occupations. Currently, fine-grained entity typing has only been investigated for English. In this abstract, we present a fine-grained entity typing system for Dutch using training data extracted from Wikipedia and DBpedia. Our system achieves comparable performance to English with an F_1 measure of .90 on 59 types and .57 on 269 types.

Approach

Our approach to generate training data is inspired by [2] and [3]. In [2], the training data is generated using Wikipedia, where the wikilink anchor text is extracted as an entity mention which map it to its corresponding Freebase entity types. We also take the Wikipedia wikilinks, anchor text and surrounding text, but instead of linking it to Freebase, we link it to DBpedia[4]. The advantage of DBpedia is that it is based on Wikipedia, therefore there is a direct link available between a wikilink and DBpedia through a mappings file.[37]

Feature name	Description	Example
Mention	The entity phrase	San Francisco
Head	The syntactic head of the entity phrase	Francisco
Non-head	The non-head tokens in the entity phrase	San
Entity-shape	The word shape of the words in the entity phrase	Aaa Aaaaaaaa
Trigrams	Character trigrams in the entity head	_Fr Fra ran anc nci cis isc sco co_
Word before	The word before the entity phrase	te
Word after	The word after the entity phrase	Californië

[37] http://downloads.dbpedia.org/2016-04/core-i18n/nl/wikipedia_links_nl.ttl.bz2

Table 1: Description of the extracted features

We base our feature vectors on [3], where we leave out the dependency and topic related features due to processing constraints. This results in the features displayed in Table 1.

To compare our results to those in previous work, we mapped the DBpedia type hierarchy to the entity typing hierarchy used in [2] and [3]. Out of the 86 types that were present, 9 types could not be mapped to the DBpedia type hierarchy.[38] As not all types are present in the dataset, we only find 59 of the types from previous work in our dataset. We also perform a series of experiments with the full DBpedia type hierarchy, resulting in an experiment with 269 types to predict.

As there are no fine-grained entity typing datasets available for Dutch yet, we split the generated dataset into ⅔ parts for training and ⅓ parts for test. This results in about 1 million instances for training on the set with 59 entity types, and 2 million on the set with 269 entity types.

We use the FastText algorithm [5,6][39] to train our type prediction model. This algorithm learns representations for character n-grams and words are represented as the sum of the n-gram vectors. This helps in covering morphologically rich languages, words that do not occur often and potentially entity mentions that do not occur in the training corpus.

Experiments and Results

We first evaluate our approach on the entity types from previous work (rows 2-6 in Table 2). At Level 1, coarse-grained entity types (person, location, organisation, and other) are evaluated. These are the same high-level types that are present in most named entity classification tasks. At Level 2, the finer-grained entity types that are directly below these are evaluated (e.g. person/artist and organisation/company). At Level 3, super fine-grained types are evaluated, for which we still achieve a macro F_1 of .90 (e.g. person/artist/music and organisation/company/news).

Types	Precision	Recall	F1
Level 1: 4 types	.98	.98	.98
Level 2: 33 types	.92	.90	.91
Level 3: 24 types	.89	.91	.90
Overall (59 types)	.93	.88	.90
Overall only dark entities (59 types)	.67	.56	.60
DBpedia types (269)	.68	.52	.57
DBpedia types, only dark entities (269 types)	.50	.41	.44

Table 2: Precision, recall and macro-average F_1

[38] The types we could not map were the following: location/structure/government, organization/stock exchange, other/health, other/living thing, other/product/car, other/product/computer, person/education, person/education/student, person/education/teacher

[39] https://github.com/facebookresearch/fastText

We also evaluated the approach on only dark entities (i.e. entity mentions that were not present in the training data).[40] Here we see that the scores drop to and F_1 of .60 which is in line with previous research [7]. It is unlikely that there is no overlap between the training and test data, but this issue deserves further investigation.

Furthermore, we see that the results for the DBpedia type hierarchy containing 269 types are significantly lower, but there is less training data available for those and not all 685 DBpedia types are covered. This is partly a result of the mappings file only containing the most specific DBpedia type, for example http://nl.dbpedia.org/resource/Old_ Amsterdam is listed as having type 'Cheese' in the mappings file, but its superclass 'Food' is not present.

Conclusions and Future Work

We have presented an approach and experiments for fine-grained entity typing for Dutch which can be particularly interesting for collecting information about entities in digital humanities sources. Our results are on par with previous work for English and our software is available at https://github.com/cltl/multilingual-finegrained-entity-typing.

For future work, we aim to test the approach on historical datasets such as the NIOD "Getuigen Verhalen" dataset and Biografisch Portaal. We also intend to compile a subset of most relevant types for the digital humanities domain and provide a trained model for reuse by humanities researchers.

References:

[1] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)

[2] Ling, X., Weld, D.S.: Fine-grained entity recognition. In: AAAI (2012)

[3] Gillick, D., Lazic, N., Ganchev, K., Kirchner, J., Huynh, D.: Context-dependent fine-grained entity type tagging. In: arXiv (2014)

[4] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. Web Semantics: science, services and agents on the world wide web 7(3), 154–165 (2009)

[5] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Tech. rep., Archiv (2016), https://arxiv.org/abs/1607. 04606

[6] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. Tech. rep., arXiv (2016), https://arxiv.org/abs/1607.01759

[7] Yaghoobzadeh, Y., Schütze, H.: Corpus-level fine-grained entity typing using contextual information. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 715–725. Association for Computational Linguistics, Lisbon, Portugal (17-21 September 2015)

[40] Whilst we made sure the training data and test data were separate on the instance level, popular entities can still be mentioned in both datasets

Session G

1. Predicting familial risk of dyslexia by applying machine learning to infant vocabulary data

Ao Chen*[1,2], Frank Wijnen[2], Charlotte Koster[3], Hugo Schnack[1]
[1] Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands
[2] Institute of Linguistics, Utrecht University, Utrecht, the Netherlands
[3] Center for Language and Cognition Groningen, University of Groningen, Groningen, the Netherlands

Background

The combination of rapid progress in the development of computational tools, such as machine learning, and the growing availability of digitized data in language research (e.g., the DANS data archive) and tools to assess these data (e.g., via CLARIAH), has made it possible to investigate language acquisition in an automated way and on a large scale (we used 22,000 vocabulary scores in our study). In this study, we applied a machine learning algorithm to vocabulary data to map the pattern of vocabulary development in individual children. We investigated whether individual differences between children in the word knowledge in different word classes (e.g., nouns, pronouns, helping verbs) can be used to detect if a child is at risk of developing dyslexia. Early detection of developmental dyslexia, a specific reading disorder, will enable interventions at an early age, before the onset of formal reading and spelling instruction. Although deviations in early speech/language development have frequently been related to (risk of) dyslexia (van der Leij et al, 2013), none of these markers have been successfully used to predict later language/literacy performance at the individual level. Machine learning is a technique capable of discovering patterns in data to make such predictions. In the past decade machine learning has been successfully employed in, e.g., medicine and the humanities. Recent examples include the prediction of disease-course in psychosis (Koutsouleris et al, 2016) and the attribution of a writer who was previously not considered, as author of the Dutch anthem (Kestemont et al, 2016). The aim of this study was to investigate if early vocabulary development can be used to predict whether or not an infant is at risk of dyslexia.

Method

We investigated early vocabulary development in two large, independent samples of children at familial risk of dyslexia (FR; *N*=495) and typically developing children (TD; *N*=498) between 17 and 35 months of age. The Dutch version of the McArthur-Bates Communicative Development Inventory (Words and Sentences) (N-CDI; Fenson et al, 1993) was used to measure each infant's vocabulary development. This was done by counting the number of words he/she knew in 22 word categories. These so-called 22 features formed together the feature vector representing this subject. We trained a linear support vector machine (SVM; Vapnik, 1999) to predict the status of at-risk at the *individual* level, based on these feature vectors. SVM is a supervised machine learning technique that is able to find patterns in the input data (word counts in 22 categories, in our case) that are related to some output measure (in our case: belonging to the FR or TD group). The training procedure results in a model that optimally predicts for (new) subjects to which group they belong. This prediction is based on the weighted sum of the input variables, where the weights are the result of the optimization procedure during training.

Performance of our prediction model was assessed by the percentages of FR subjects that were correctly classified as FR (sensitivity), the percentage of TD subjects correctly classified (specificity) and the balanced accuracy (mean of sensitivity and specificity).

The model's generalizability was tested using cross-validation. In this setup the model is trained and tested in different subsamples.

Results

There was a specific age period, 18-20 months, in which the model was sensitive to predict the status of being at risk (FR). At 19-20 months of age, the cross-validation accuracy was 68% (p<0.01), with sensitivity being 70% and specificity being 67%. In the other age groups the accuracy was lower and not significant.

Not all 22 features contributed to the same extent to the discrimination between the FR and TD subjects at age 19-20 months. The weights of 5 word categories were significantly different from zero. The categories *helping verbs* and *prepositions and locations* contributed most. The model had learnt from the data that knowing fewer words in these categories at this age is a significant marker for being at family risk.

Conclusion

Machine learning methods are promising techniques for separating FR and TD children at an early age, before they start reading. There is a sensitive window in which the difference between FR and TD is most evident. The model also indicated the word categories in which FR infants know (on average) fewer words as compared to TD infants. It should be noted that we did not predict the manifestation of dyslexia, but only elevated risk. We will follow these children up, and the ultimate goal is to train a model that is able to discriminate between the FR children who develop dyslexia and who do not at an early age.

References

CLARIAH. http://www.clariah.nl

DANS. https://dans.knaw.nl

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., et al. (1993). The MacArthur Communicative Development Inventories: User's Guide and Technical Manual. San Diego, CA: Singular Publishing Group.

Kestemont M, Stronks E, De Bruin M, De Winkel T. Van wie is het Wilhelmus? (2016 Dec) Amsterdam University Press.

Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, Derks EM, Fleischhacker WW, Hasan A. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry. 2016 Oct;3(10):935-946. doi: 10.1016/S2215-0366(16)30171-7.

van der Leij, A., van Bergen, E., van Zuijen, T., de Jong, P., Maurits, N., and Maassen, B. (2013). Precursors of developmental dyslexia: an overview of the longitudinal dutch dyslexia programme study. Dyslexia 19, 191–213. doi: 10. 1002/dys.1463.

Vapnik, VN. (1999). An overview of statistical learning theory. Neural Networks, IEEE Transactions on, 10(5), 988-999. doi:10.1109/72.788640.

2. The Dictionary of the Southern Dutch Dialects (DSDD): Designing a Virtual Research Environment for digital lexicographical research

Prof. dr. Jacques Van Keymeulen
Ghent University, Belgium

The southern Dutch dialect area consists of four dialect groups: (1) the Flemish dialects, spoken in French Flanders (France), West and East Flanders (Belgium) and Zeeland Flanders (The Netherlands); (2) the Brabantic dialects, spoken in Antwerp and Flemish Brabant (Belgium) and Northern Brabant (The Netherlands); (3) the Limburgian dialects (spoken in the Limburg provinces of Belgium and The Netherlands); (4) the Zeeland dialects, spoken in Zeeland and Goeree-Overflakkee (the Netherlands).

The dialect vocabulary of the Flemish, Brabantic and Limburgian dialects is collected in three regional dictionaries (WVD, WBD and WLD respectively), which are set up according to the same plan, conceived by prof. A. Weijnen (Nijmegen): they are onomasiologically arranged and published in thematic fascicles. Contrary to their titles, these dictionaries are to be considered as geographically-orientated inventories of word usage, and not as dictionaries proper, since it is impossible to describe meaning in an onomasiologically arranged dictionary. They are atlasses, not dictionaries! We retain the word dictionaries – however – since the three projects are traditionally known as such.

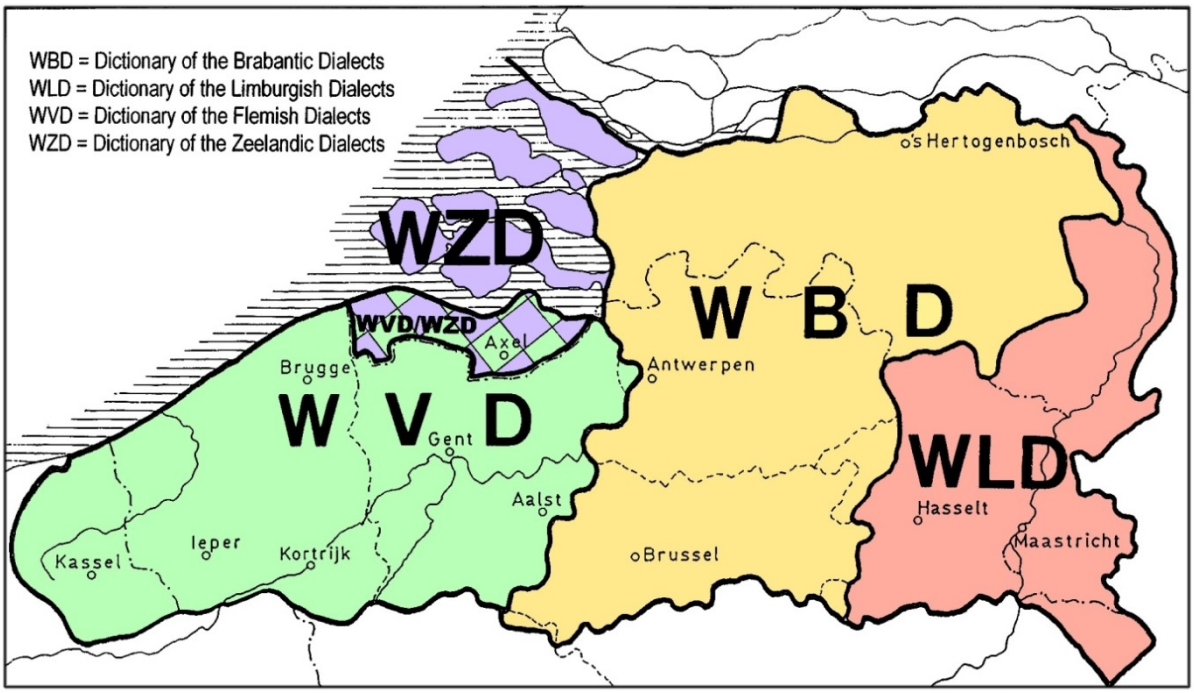


Figure 1: Research areas of the 4 regional dialect dictionaries of the southern Dutch area

The three dictionaries describe the vocabulary of the traditional dialects of the first half of the twentieth century in the southern part of the Dutch language area, in a joint **international** and **inter-university** project. The WBD, the 'mother' of the two other projects, was finished in 2005; the WLD was completed in 2008. They were compiled at the University of Nijmegen and the University of Leuven. The WVD started 12 years later than its sister projects (in 1972 at the Ghent University, by prof. W. Pée) and will continue until about 2019.

The dictionaries were set up in parallel in order to make possible the aggregation of the data, thus fulfilling the objectives of the founders of the projects. To that effect, in 2016 a consortium of 11 linguists, computer scientists, digital humanities experts and geographers was created supporting the project "Dictionary of the Southern Dialects" (DSDD). It aims at the aggregation and standardization of the three comprehensive dialect lexicographic databases into one DSDD-database (to which hopefully the alphabetically arranged WZD will be added in the future). In particular, dialectologists from Ghent University work closely with the Ghent Centre for Digital Humanities (GhentCDH) to prepare the ground for the aggregation of the three Southern Dutch dialect databases and their exploitation via a Virtual Research Environment for digital lexicographical research. The Ghent team will work closely with the *Instituut van de Nederlandse Taal* with regard to the technical and linguistic sustainability of the DSDD. Through this collaboration interoperability with CLARIN will also be ensured. The DSDD is additionally a pilot project of DARIAH-BE Belgium.

The *DSDD* Virtual Research Environment will enable a research programme with new research questions, particularly in the field of quantitative lexicology and geographical analysis. During the project 2-3 research use cases will be developed to test the applicability of the newly aggregated *DSDD* for digital scholarship. For example:

1. What systematic lexico-geographical patterns do the southern Dutch dialects show? Do they coincide with the traditional ones, based on phonology? (see De Vriendt 2012). Are there geographical patterns in semantics?
2. In order to explore the geographical spreading of several dialectology concepts and to link them to "Kloekeplaatscodes" (which are used in linguistic research for mapping/linking dialectology concepts to geographical regions), a set of generic building blocks for automatic atlas/heatmap generation will be developed. Segmentation and clustering techniques can be run over the generated atlases/heatmaps in order to automatically detect the homogeneity (or heterogeneity) of a particular dialectology concept. Furthermore, spatial querying techniques will be supported in order to geographically search/explore this kind of dialectology concepts.
3. Cluster analysis and exploration of the linkage (and visualization) of linguistic data with synchronic and diachronic extralinguistic data of all kinds.

By the end of the project, the DSDD will a) make the newly aggregated *DSDD* available via a user-friendly website and b) enable the *DSDD* for digital scholarship. To enable this, a professionally designed user-friendly web application, or Virtual Research Environment, (including application programming interface (API) for data export) will be created. The exported data will use existing digital research tools (e.g. for geo-visualisation, qualitative lexicology and dialectometry) to validate the research case studies described above.

At the DH Benelux Conference, we will propose the plan for the aggregation, the structure of the database and dwell on the different 'editorial' problems that have to be solved. The different dictionaries / database were indeed composed over a very long period of time, at different places (Nijmegen, Leuven, Ghent) and by different editors, hence a great number of inconsistencies arose over time. In order to compose an aggregated DSDD-database, a number of standardization activities have to be carried out. Additionally, we will present the initial results of the Virtual Research Environment requirements analysis.

3. Establishing interdisciplinary dialogue: conducting a qualitative investigation into linguistic requirements for Natural Language Generation

Emma Clarke and Owen Conlan

Background

Dialogue systems, commonly referred to as chatbots are becoming increasingly popular. In 2016, chatbot was shortlisted as word of the year by Oxford Dictionaries[41] and platforms such as Facebook Messenger[2] are frequently utilised to communicate updates or information, sell products or provide services. While the goal of a dialogue system which communicates naturally with its user appeared to have been 'within reach' as far back as 2001 (Rambow et al., 2001), current Natural Language Generation (NLG) research approaches continue to have limitations when it comes to the *'natural-ness'* of their interactions (LeCun et al., 2015) (Reiter and Dale, 2006) (Manning and Schütze, 1999). Thus, the NLG field is looking to move towards more natural conversational interfaces by taking influence from natural human speech and as dialogue systems become more human-like, the interspersion of persuasive language within them will become more applicable. Some prior research has been carried out on the development of persuasive dialogue systems (Prakken, 2009) (Parsons et al., 2003) (Walton and Krabbe, 1995). Most recently, Hiraoka et al. (2016) observed that "these persuasive dialogue systems are in their first stages of development, and are far from the abilities of their human counterparts, both in terms of persuasive ability, and also ability to achieve user satisfaction". The focus of this research project is the language of persuasion, namely rhetorical devices. We believe that in order to understand the requirements of the NLG community in this area, the establishment of cross-disciplinary conversation is essential.

Challenge

The nuances of human speech such as sarcasm, slang and wordplay and the human ability to process and understand these subtleties make them equally fascinating and frustrating for researchers in the areas of natural language processing, understanding and generation. A major challenge faced by Natural Language Generation (NLG) researchers is how to incorporate linguistic understanding into NLG systems in order to generate more natural sounding language. This challenge is expected to continue to pervade in the next generation of natural language systems (Dale, 2016) (LeCun et al., 2015) (Ward and DeVault, 2015) (Gartner, n.d.).

Often lacking in dialogue systems and NLG research is linguistic expertise presented in a form which is understandable, that dissects natural elements of human speech, particularly elements which are difficult for machines to learn. Ward and DeVault (2015) highlight this interdisciplinary engagement in their 'Ten Challenges in Highly-Interactive Dialog Systems'.

As interdisciplinary research becomes more prevalent, the requirement for computer science practitioners to engage with non-technical researchers from diverse backgrounds will increase. Dale (2016) also refers to cross-disciplinary conversations and encourages dialogue systems developers to access the expertise of the computational linguistics community, in which research into discourse phenomena has been on-going since the inception of the field. Dale (2016) presents an encouraging call to action: "If we want to have better conversations with machines, we stand to benefit from having better conversations among ourselves.".

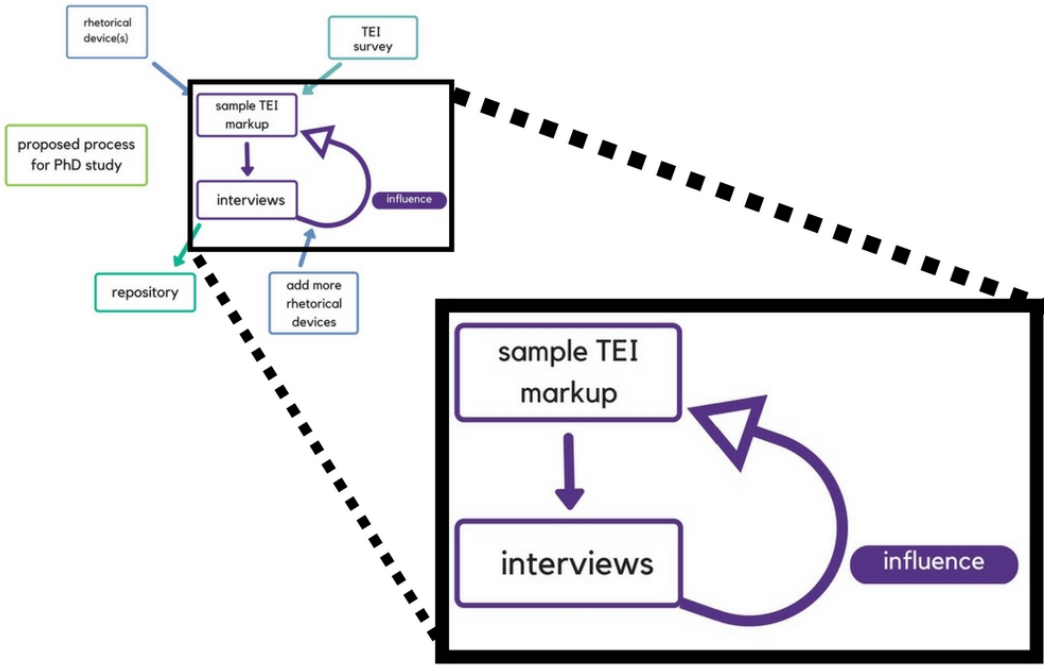
Approach

The overall aim of this research (fig. 1) is to establish an an approach to understanding how rhetorical devices function in natural human speech in order to propose a method which can be built into practical NLG applications such as dialogue systems (chatbots). The work will draw upon structured rather than random influence by observing the usage of these linguistic strategies for persuasion in human speech. From these observations, a TEI schema has been customised in order to

[41] https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016 [2] https://www.messenger.com/

markup a set of rhetorical devices within a corpus.

Figure 1



This paper will present findings on the central component of the diagram above: the cross-disciplinary engagement with NLG practitioners in order to develop a pragmatic approach to incorporating persuasive language into dialogue systems. We explore how a customised TEI schema is used in semi-structured interviews with NLG researchers (an ongoing, iterative process). Based on qualitative findings from the interviews, the schema is revised and amended to incorporate requirements and suggestions. The final schema will ultimately be used to markup and annotate speeches from the corpus in order to be added to NLG as part of the system training.

Method

A series of semi-structured interviews are being carried out in which ten NLG practitioners in are asked questions in order to understand current and future requirements of NLG applications such as dialogues systems.

In the course of each interview, the TEI schema is presented and the suggestions of the NLG practitioners sought. The interviews are recorded and the resulting outcomes are analysed using atlas.ti software. The results are then summarised to create an overall picture of NLG researcher requirements.

Outcomes (to date)

The process outlined above is ongoing at the time of submission. However, preliminary findings from the interviews can be summarised as follows:

- ● Both template-driven and deep learning systems use annotated data. In a rule-based approach, annotations are used to help further engineer features by hand while a deep learning approach uses annotation to help learn and understand structure.

- There is an emerging question in NLG research about how to deal with sentence structure and nuance. Increasingly, researchers are using marked up text to help systems learn higher order structures.
- Pattern-matching alone is not a robust enough approach.
- A very clear annotation schema that marks up features of rhetorical devices would be useful for NLG researchers working in the area of persuasion. Conclusion The aim of this research is to engage in an interdisciplinary conversation with NLG practitioners. The process of engagement and the findings from the interviews will be presented in this paper.

References

Dale, R., 2016. The return of the chatbots. Nat. Lang. Eng. 22, 811–817. Gartner, n.d. Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage [WWW Document]. URL http://www.gartner.com/newsroom/id/3412017 (accessed 11.24.16). Hiraoka, T., Neubig, G., Sakti, S., Toda, T., Nakamura, S., 2016. Construction and analysis of a persuasive dialogue corpus, in: Situated Dialog in Speech-Based Human-

Computer Interaction. Springer, pp. 125–138. LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. Manning, C.D., Schütze, H., 1999. Foundations of statistical natural language processing.

MIT Press, Cambridge, Mass. ; London. Parsons, S., Wooldridge, M., Amgoud, L., 2003. Properties and complexity of some formal

inter-agent dialogues. J. Log. Comput. 13, 347–376. Prakken, H., 2009. Models of persuasion dialogue, in: Argumentation in Artificial Intelligence.

Springer, pp. 281–300. Rambow, O., Bangalore, S., Walker, M., 2001. Natural language generation in dialog systems, in: Proceedings of the First International Conference on Human Language

Technology Research. Association for Computational Linguistics, pp. 1–4. Reiter, E., Dale, R., 2006. Building natural language generation systems, Digitally printed 1st pbk. version. ed, Studies in natural language processing. Cambridge University

Press, Casmbridge, U.K. ; New York. Walton, D., Krabbe, E.C., 1995. Commitment in dialogue: Basic concepts of interpersonal

reasoning. SUNY press. Ward, N.G., DeVault, D., 2015. Ten challenges in highly-interactive dialog systems, in: AAAI

Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction.

Session H

1. Getting the Bigger Picture: Exploratory Search and Narrative Creation for Media Research into Disruptive Events

dr. Berber Hagedoorn, University of Groningen, Research Centre for Media Studies and Journalism
dr. Sabrina Sauer, University of Groningen, Research Centre for Media Studies and Journalism

Introduction

Digital Humanities centres on questions that are raised by and answered with digital tools in the Humanities. At the same time, it interrogates the value and limitations of digital methods in Humanities' disciplines. While it is important to understand how digital technologies can offer new venues for Humanities research, it is equally essential to understand – and therefore, being able to interpret – 'the user side' of Digital Humanities. Specifically, how Humanities researchers appropriate and domesticate search tools to ask and answer new questions, and apply digital methods. Previous user research in Digital Humanities concentrates on assessing, for example, how and why Digital Humanities benefits from studies into user needs and behaviour (Warwick, 2012), user requirement research, as well as participatory design research (Kemman & Kleppe, 2014).

Exploratory search is crucial for Humanities researchers who draw upon media materials in their research. Audio-visual, online and digital sources are in abundance, scattered across different platforms, and changing daily in our contemporary landscape. Supporting researchers' explorations becomes even more important when scholars study media events. A 'media event' is an event with a specific narrative that gives the event its meaning, and is in contemporary societies increasingly recognized as non-planned or disruptive. Disruptive media events, such as the 'sudden' rise of populist politicians, terrorist attacks or environmental disasters, are shocking and unexpected, making them difficult to interpret. This leads to problems for media researchers who analyse how narratives construct different political, economic or cultural meanings around such events. Previous research argues that media events should always be viewed in relation to their wider political and sociocultural contexts. Events, as they unfold in the media, may correspond to long-term social phenomena, and the way in which such events are 'constructed' has particular connotations (Jiménez-Martínez, 2016). Specific actors (newscasters, governments, institutions) use media events to build narratives in line with their own political, economic or cultural purposes. Media researchers also build narratives around events; prior research underlines the importance of visualizing, constructing and storing of narratives during the information navigation to contextualize material (Akker et al., 2011; Kruijt, 2016; De Leeuw, 2012). Offering media researchers the ability to explore and create lucid narratives about media events therefore greatly supports their interpretative work.

This paper proposes to add to this body of research by presenting the insights of a cross-disciplinary user study that involves, broadly speaking, researchers studying audio-visual materials, in a co-creative design process, set to fine-tune and further develop a digital tool that supports Humanities' research through exploratory search. This paper focuses on how researchers - in both academic as well as professional settings - use digital search technologies in their daily work practices to discover and explore digital audio-visual archival material. We focus specifically on three user groups, namely (1) Media Studies researchers, (2) Humanities researchers that use audio-visual materials as a source and (3) Media professionals. These user groups are the foreseen end users of the tool, because they create audiovisual narratives for their respective work purposes. We set-up co-creative design sessions with 74 participants (group 1: 24; group 2: 40; group 3: 10) to observe and reflect on the practices of media researchers in terms of how they interact with search tools to explore, access and retrieve digitized audio-visual material, in order to interpret, and in some cases, re-use this material in new audio-visual productions.

Methodology

In our user study, we employ a user-centred design methodology to evaluate and fine-tune the exploratory search tool DIVE+ media browser. It offers events-driven exploration of digital heritage material, where events are prominent building blocks in the creation of narrative backbones (De Boer et al., 2015) and links a variety of different media sources and collections. DIVE+ offers intuitive exploration of media events at different levels of detail. It connects media objects, subjects ("concepts"), events, and persons to aid in the formulation of research questions, and to contextualize the former into overarching narratives and timelines. Our main research question throughout the case study is how does exploratory search support media researchers in their study of how media events are constructed across different media and instilled with specific cultural or political meanings? To be able to answer this question, we study how media researchers construct navigation paths via exploratory search and - by means of user studies - evaluate the role of narratives in (1) learning and (2) research. In this process, we compare DIVE+ to other online search tools.

The user study observes media researchers as they use DIVE+ to explore media events, across 3 stages: (1) during research question formulation (2) DIVE+ use; and (3) comparative user evaluations of the DIVE+ browser, compared to other online search tools. The collected data, consisting of both qualitative – observational and focus group - data, as well as logging data gathered during user testing, provides insights about how media researchers search and explore digital audio-visual archives. We utilize a case study approach, which combines grounded theory (that fosters an understanding of how researchers interpret and create narratives) with usability methodologies, such as work task evaluations. This, first of all, allows us to draw conclusions about how search tools and digital technologies co-construct the researcher's professional practice. Second, the data helps us probe the question how the 'digitality' of search and retrieval shapes the practice of media research, and, in extension of this, creative processes.

The research presented in this paper takes an interdisciplinary approach: it combines insights from Media Studies, as well as from Information Studies and Science and Technology Studies and integrates ideas about narrative creation, search practices, and overarching notions about how users and technologies co-construct meaning. Therefore the presented research does not focus on how Digital Humanities' tools have an impact on researchers' practices, but rather analyses how researchers make use of search tools. We subsequently (1) draw conclusions about scholarly practice and the role of search technologies for digitized audio-visual materials therein; and (2) present lessons learned on how to optimize the search tool that is used, in order to improve its performance.

Acknowledgments

The authors would like to thank the anonymous reviewers of the first version of this abstract for their helpful comments and suggestions. This research was supported by the Netherlands Institute for Sound and Vision (partially in the context of Berber Hagedoorn as Sound and Vision Researcher in Residence in 2016-7) and the Netherlands Organisation for Scientific Research (NWO) under project number CI-14-25 as part of the MediaNow project. This research was also supported by CLARIAH, Common Lab Infrastructure of Arts and Humanities, in the context of the Research Pilot Narrativizing Disruption: How exploratory search can support media researchers to interpret 'disruptive' media events as lucid narratives (https://www.clariah.nl/projecten/research-pilots/nardis), CLARIAH-project number CC 17-13. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Bibliography

Akker, C. van den, Legêne, S., Erp, M van, Aroyo, L., Segers, R. Meij, L. van der, Ossenbruggen, J. van, Schreiber, G. Wielinga, B., Oomen, J., Jacobs, G. (2011). Digital Hermeneutics: Agora and the Online

Understanding of Cultural Heritage Categories and Subject Descriptors. WebSci 11, Koblenz, Germany.

Boer, V. de, Oomen, J., Inel, O., Aroyo, L., Staveren, E. van, Helmich, W., & Beurs, D. de. (2015). DIVE into the Event-Based Browsing of Linked Historical Media. Web Semantics: Science, Services and Agents on the World Wide Web, 35(3), 152–158.

De Leeuw, S. (2012). European Television History Online: History and Challenges. VIEW Journal of European Television History and Culture, 1(1), 3–11.

Jiménez-Martínez, C. (2016). Integrative disruption: the rescue of the 33 Chilean miners as a live media event. In: Fox, A., (ed.) Global Perspectives on Media Events in Contemporary Society. IGI Publishers, Hershey, USA, 60-77.

Katz, E., and Liebes, T. (2007). 'No More Peace!': How Disaster, Terror and War Have Upstaged Media Events. International Journal of Communication 1, 157-166.

Kemman, M, and Kleppe, M. (2014). "User Required? On the Value of User Research in the Digital Humanities." Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands. No. 116. Linköping University Electronic Press.

Kruijt, M. (2016). Supporting Exploratory Search with Features, Visualizations, and Interface Design: A Theoretical Framework. University of Amsterdam.

Warwick, C. (2012). "Studying users in Digital Humanities." Digital Humanities in practice, 1-21.

2. Bias in the analysis of multilingual legislative speech

Laura Hollink, Astrid van Aggelen, Jacco van Ossenbruggen
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
l.hollink@cwi.nl

In this paper we investigate the application of natural language processing tools to the multilingual proceedings of the European Parliament. This work is part of a study in which we explore (1) how subcorpora in different languages may lead to different conclusions about the political landscape, (2) how to determine what a potential language-related bias originates from, and (3) to what extent we can limit or even prevent an unwanted language-bias.

Parliamentary speech has been used to study party positions [1,2,3], issue selection [4,5,6,7] and the level of disagreement within a debate [8]. Many studies have moved away from manual coding (which is done in e.g. [4,5]) and instead position speech texts on one or more (latent) dimensions in statistical models based on relative word frequencies [1,2,3,6,7,8], often in combination with basic pre-processing steps such as stemming and stopping. These models and tools, while imperative to analyse bigger datasets, add a source of errors and bias. One source of potential bias comes from the fact that the used tools perform differently on different languages. Considering that the aforementioned studies were carried out on the European, Irish, US, Spanish, Norwegian and Swedish legislatures, the comparability and reproducibility of the results for different languages is unclear.

In the European Parliament, the spoken accounts appear in (currently) 24 languages. Here, the uncertainty stems not only from tools that perform differently on each language, but also from the fact that the availability of data in each language varies. Members of Parliament (MEPs) are free to speak in any of the official languages. Speeches are sometimes translated into (some) other languages, depending on prioritization with the EP, specific translation-requests of the members and

(supposedly) budgetary constraints. Thus, we are left with 24 subcorpora of varying size, one per language, including both original and translated speech.

The need to study language-effects in this context has been recognised before. Proksch *et al.* [3] reported a modest language-effect[42] in their study of party positions in the European Parliament, which they ascribed to translation rather than actual differences in position taking between three countries. However, while the overall effect may be small, we argue that specific local effects could still lead to significant biases in the results. For example, French translations of German texts seemed to systematically get a more neutral position than the original text, while the opposite was not the case. It is important to realise that the proceedings of the European Parliament are not only a corpus for researchers. Residents of the European Union have a right to access these documents in order to make informed votes and to hold the MEPs accountable[43]. This right would be compromised when French speaking citizens come to different conclusions about what has been discussed than German speaking citizens. Our aim is to gain insight into how working with subcorpora in different languages may lead to different conclusions about the political landscape.

In this study, we use the data provided by the Talk of Europe project [9], in which speech transcripts and all available translations were crawled from the website of the EP[44], and translated into the semantic web format RDF. Data is available from 1999 to 2015 and contains around 300K speeches in 22K debates. We apply topic detection to six language-specific subcorpora of the proceedings of the European Parliament: German, English, French, Italian, Spanish and Dutch. We use the JEX software developed by the European Commission's Joint Research Centre, which learns multi-label categorisation rules from documents that were previously manually indexed using the multilingual Eurovoc thesaurus [10]. The advantage of using this tool over, for instance, widely used topic modeling approaches such as LDA [11], is that the output is directly comparable across languages: the tool uses a single thesaurus, Eurovoc, to classify documents in each language, and concepts in the Eurovoc thesaurus have labels in all languages. In a later stage of the study, we plan to include other topic detection techniques, and widen the scope to all EU languages.

Over 2000 distinct Eurovoc topics were detected in the six subcorpora. The frequency distributions over topics vary per language. Figure 1 visualises the distance between languages. We use Kullback–Leibler divergence [12], a non-symmetric measure for the difference between two distributions. A higher score, visualized as a redder colour, signifies a greater distance. For example, Italian and French are relatively close, while Spanish and German are far apart. There are four hypotheses as to what these differences originate from:

1. MEPs speaking one language indeed speak about different topics than their colleagues who speak in another language.
2. There is a bias in the selection of speeches that are being translated.
3. There is a bias in how certain topics are translated, e.g. translators use more ambiguous or polarized language.
4. The topic detection tool works differently on one language than on another.

[42] A correlation coefficient ranging between 0.86 and 0.93 when comparing party positions derived from texts in German, French and English [3].
[43] Regulation (EC) No 1049/2001 of the European Parliament and of the Council
[44] http: //www.europarl.europa.eu

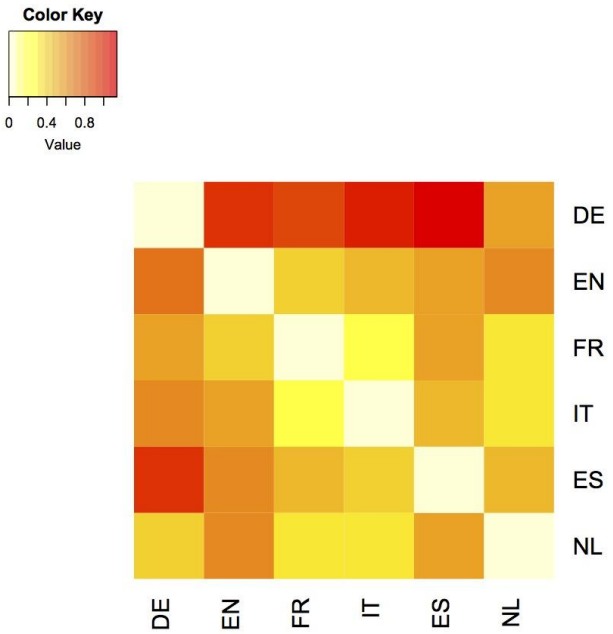


Figure 1: Heatmap of differences between topic distributions in languages.

In our presentation, we will tackle this issue from two sides. Firstly, we compare different subsets of topics based on whether or not speeches were translated, and to which languages, to explore hypotheses 1 and 2. Then, to study hypothesis 4 (and to a lesser extent hypothesis 3) we zoom into topics that appear to be particularly distinctive between languages, and compare the topic annotations to what was actually said in the debates. As an example of the latter method, Figure 2 shows the differences in frequency of the detected topics “nuclear weapons” and “nuclear energy”. Remarkably, only French and Italian speeches seem to be about nuclear weapons, while English and Spanish speeches are often about nuclear energy. As a comparison, Figure 3 plots the occurrences of the phrases “nuclear weapons” and “nuclear energy” (and translations thereof) in the raw speech texts. Here, part of the effect is gone, suggesting an error of the topic annotation software, while part of the effect remains - German texts indeed seem to talk less about both nuclear weapons and nuclear energy.

With this study, we aim to contribute to the discussion about systematic methods for tool criticism and source criticism in a complex multilingual context like the European Parliament.

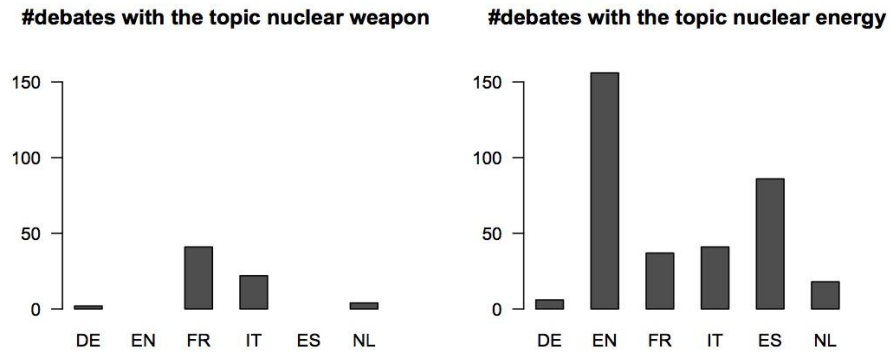


Figure 2: Frequency of topics in debates.

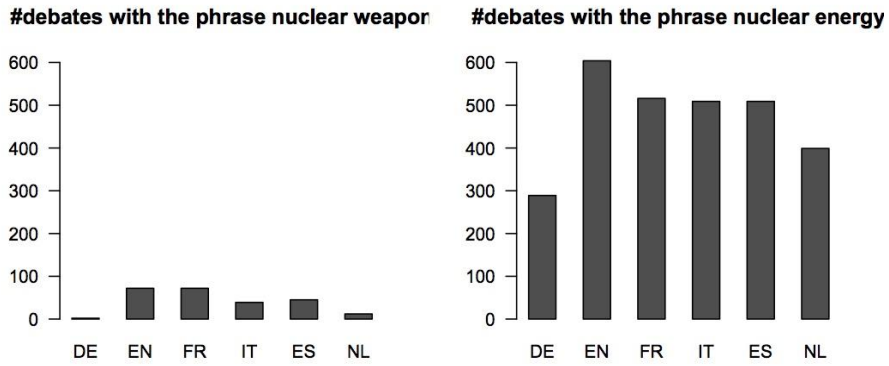


Figure 3: Frequency of phrases in debate texts.

References

[1] Benoit, Kenneth, and Michael Laver Nd. Estimating Irish Party Positions Using Computer Wordscoring: The 2002 Elections. Irish Political Studies Vol. 18 , Iss. 1, 2003.

[2] Laver, Michael J., Kenneth R. Benoit, and John Garry. Extracting Policy Positions from Political Texts Using Words as Data. American Political Science Review 97(2):311–31, 2003.

[3] Proksch, S.-O. and Slapin, J.B. Position Taking in European Parliament Speeches, British Journal of Political Science, 40(3), pp. 587–611, 2010.

[4] Hanna Bäck, Marc Debus & Jochen Müller. Who Takes the Parliamentary Floor? The Role of Gender in Speech-making in the Swedish Riksdag. Political Research Quarterly 67: 504–518, 2014.

[5] Markus Baumann. Constituency Demands and Limited Supplies: Comparing Personal Issue Emphases in Co-sponsorship of Bills and Legislative Speech. Scandinavian Political Studies, Vol. 39, issue 4, pp. 366-387, 2016.

[6] Pardos-Prado, Sergi, and Iñaki Sagarzazu. The Political Conditioning of Subjective Economic Evaluations: The Role of Party Discourse. British Journal of Political Science 46(4), 799-823, 2016.

[7] Kevin M. Quinn , Burt L. Monroe , Michael Colaresi , Michael H. Crespin , Dragomir R. Radev. An automated method of topic-coding legislative speech over time with application to the 105th-108th US Senate. Midwest Political Science Association Meeting. 2006.

[8] Benjamin E. Lauderdale, Alexander Herzog. Measuring Political Positions from Legislative Speech. Polit Anal; 24 (3): 374-394, 2016.

[9] Astrid van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. The debates of the european parliament as linked open data. Semantic Web, 8(2):271–281, 2017.

[10] Pouliquen Bruno, Steinberger Ralf, Camelia Ignat. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology - Its Potential and Practicalities (EUROLAN'2003). Bucharest, Romania, 28 July - 8 August 2003.

[11] Blei, David M., Ng, Andrew Y., Jordan, Michael I. Lafferty, John, ed. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3 (4–5): pp. 993–1022, 2003.

[12] Kullback, S., Leibler, R.A. On information and sufficiency. Annals of Mathematical Statistics. 22 (1): 79–86, 1951.

Session I

Cultural Heritage Data for Research: A Europeana Research Panel

Nienke van Schaverbeke, Head of Europeana Collections
Marjolein de Vos, Europeana Data Partner Services
Dr. Agiatis Benardou, Digital Curation Unit, R.C. "Athena", Institute for the Management of Information Systems

Panel members:

Nienke van Schaverbeke - Head of Europeana Collections - session Chair

Dr. Agiatis Benardou - Digital Curation Unit, R.C. "Athena", Institute for the Management of Information Systems - Researcher Needs Management

1 Member of our Board from a research network (http://research.europeana.eu/blogpost/europeana-research-advisory-board-established) - **TBC**

Marjolein de Vos - Europeana, Digitised Medieval Manuscripts Maps - Data Quality

Dr. Caroline Ardrey - University of Birmingham - Europeana Grants Winner

Dr. Dana Mustata - University of Groningen. Academic in a digital humanities related field, outsider to Europeana - **TBC**

Cultural Heritage Data for Research: A Europeana Research Panel

In this panel members of the Europeana Research Advisory Board, Europeana Data Partner Services, one of the Research Grants winners and, importantly, an academic external to Europeana will present and discuss the value of Europe's cultural heritage data for research in the humanities and social sciences, and the ways in which Europeana Research is promoting and enabling its use. The panel is part of a larger discussion going on about making cultural heritage available for research and the opportunities, challenges, and considerations involved in this.

In short, the panel will focus on the following points:

- Europeana Research - Objectives & Achievements
- Relationship to other research networks and infrastructures (DARIAH, CLARIN, EHRI, Parthenos etc)
- Researcher needs and community engagement
- Data aggregation and quality improvement
- Using Europeana data in research

Europeana Research was established as a link between cultural heritage institutions and researchers. We recognize that undertaking research on the digitised content of Europe's galleries, museums, libraries, and archives has huge potential that should be exploited. But issues with regards to licensing, interoperability, and access can often impede the re-use of that data in research. Europeana Research aims to help with these issues, liberating cultural heritage for meaningful academic re-use. We work on a series of activities to enhance and increase the use of Europeana data for research, and develop the content, capacity, and impact of Europeana, by fostering collaborations between Europeana and the cultural heritage and research sector, as well as liaising with other digital research infrastructures and networks.

Europeana Research is governed by an Advisory Board comprising of renowned digital humanities experts who help us grow and strengthen services for DH researchers. In the first section of the panel

we will highlight our main objectives and greatest achievements, such as the Research Grants Programme.

Following this introduction, one of our panel members, a representative from a research network that we collaborate with and an academic who is not connected to Europeana will expand and elaborate on this relationship between their network and Europeana, and the value thereof.

Since our target audience are research communities in the humanities and the social sciences, it is vital to understand their heterogeneous needs vis à vis their information behaviour and their interaction with digital content. In this part of the panel, we will go into detail about how we come to understand the needs of our users, how to cater to them, and how we continuously develop and further this understanding and adapt to the requirements.

With more than 54 million objects from 40 countries and in a variety of languages, the Europeana portal contains a substantial amount of data to manage. The Data Partner Services team does not only work continuously on ingesting new data for the portal, but also invests time into evaluating and improving existing data. We make data quality plans with aggregators and direct providers to further findability and granularity of the records in the portal. Furthermore, there is a special assigned Data Quality Committee that works on refining and expanding the Europeana Data Model. During this part of the panel, we will talk about the work that is being done from the metadata perspective on data quality, the importance of understanding researchers needs for this, and the value of cultural heritage data for research.

In 2016 the Europeana Research Grants Programme was launched, in which Digital Humanities researchers were encouraged to apply with a project where Europeana data would be central in answering their research question. The unprecedented success of this call for proposals shows us how important it is to make heritage data available; the variety in ideas showing us the range of potential of what is in the portal. To further illustrate and strengthen the points that will be mentioned in the panel one of the winners of the Europeana Research Grants Programme 2016 will discuss her project as a showcase of Europeana data re-use for research and the potential offered to research communities through open access, clear licensing, and adequate digital tools.

After providing short explanations on the points mentioned in this proposal, we will encourage discussion from the panel and the audience on these matters. These could lead to valuable insights for Europeana Research in the wider discussion of opening up cultural heritage for the research community. We also welcome suggestions for Europeana Research's future activities and improving services.

Session J

Text mining in practice: A discussion on user-applied text mining techniques in historical research.

Language: English, Duration: 60 minutes

In this panel we look at the application of text mining techniques in historical research. In recent years, text mining has come within reach of any vaguely computer-literate scholar. The growing availability of large digital text collections leads to growing abilities to apply digital and quantitative approaches to the study of historical texts. Commonly used languages and statistical environments such as Python and R, offer applicable software solutions for free. This has liberated historians and other humanities scholars from the shackles of time-consuming and often expensive programming work by hired external programmers.

Techniques like topic modelling, word embeddings, sentiment and emotion mining are increasingly being used in the humanities and social sciences. Historians, political scientists, sociologists and others now have the opportunity to use advanced text mining techniques on large datasets from their desktops. Although still mostly experimental, the potential gains now appear enormous.

It is often claimed that this enables researchers to study concepts and developments in longitudinal, systematic and quantitative ways that were impossible before. But what do these digital techniques really add to more traditional approaches? How can traditional approaches and innovative digital methodologies be paired in a meaningful and enriching manner? Does quantitative text analysis primarily provide context to existing knowledge, or is it a radical departure from what went before?

We believe that quantitative text analysis could well prove to be a dramatic, agenda-setting change. As yet, however, several problems need to be addressed. First, most of the techniques involved are less than a decade old, researchers are scattered among departments and disciplines, and there is as yet no overarching discussion about best practices, pitfalls and problems with methodology, or even a shared platform to discuss basic technical problems has been established. There is a distinct need for a better exchange of information and sharing of experience, both inside and outside the world of digital humanities.

A second problem that needs to be addressed is the slow advancement of new techniques in published research outside the narrow digital humanities world. Anecdotal evidence suggests that leading journals in the humanities, political and social sciences are not particularly keen on papers using text-mining methodologies. This unwillingness is at least in part inspired by the problem mentioned above. There are few established norms to evaluate the validity of new techniques. On the other hand, conservatism may also play a role.

A third problem, which also impacts publication opportunities, is that the bulk of publications sing text-mining techniques are still primarily *about* text mining. The corpora used, and the research questions asked, in many cases still seem peripheral to technological glitz. It is of course useful to investigate the technical opportunities that new techniques have to offer, but for the wider dissemination of these techniques it will probably prove necessary to tackle existing research problems in various fields and show that this particular field of the digital humanities has something to offer to the study of history.

We propose to discuss these problems with a mixed panel of experienced text mining researchers from different (sub-) disciplines. Our central goal is to discuss practices for validation of techniques and methodologies. We want to come up with a proposal for integrating text mining techniques in

historical research practice in a meaningful, substantive, and contributive way, and pave the way for the move of text mining into common research practice, beyond the current hype.

Chair:

- Dr. Ralf Futselaar (EUR/NIOD)

Panel members:

- Dr. Jesse de Does (IvdNT)
- Prof. dr. Yasuto Nakano (KGU, Japan)
- Dr. Martijn Schoonvelde (VU)
- Milan van Lange, MA (NIOD/UU)

Session K

Mapping Historical Leiden: The Creation of a Digital Atlas

Organiser: **Arie van Steensel, University of Groningen (a.van.steensel@rug.nl)**

Panellist: **Jaap Evert Abrahamse, Cultural Heritage Agency (j.abrahamse@cultureelerfgoed.nl)**

Speakers: **Ellen Gehring, Erfgoed Leiden en Omstreken (e.gehring@erfgoedleiden.nl)**
Roos van Oosten, Leiden University (r.m.r.van.oosten@arch.leidenuniv.nl)
Arie van Steensel, University of Groningen (a.van.steensel@rug.nl)

The digital revolution has rendered maps even more useful for all kinds of purposes, such as navigating, locating services, or geotagging activities. Moreover, a growing array of digital technologies, applications and platforms offer new research opportunities for scholars in the humanities, for whom maps are both a source about the past and a tool to study the past, and they allow heritage organisations to unlock, visualise and analyse diverse historical and archaeological data and objects innovatively on the basis of geographical relations. It is beyond doubt that the spatial encoding of objects and textual information offers a new framework of analysis and enables us to better explore the experiences and meanings of space and place in the past.[45] Tools, maps and data are often readily available for the study of the more recent past, but this is less the case for the pre-modern period. In general, it requires a considerable time investment to develop historical Geo Information Systems (GIS) and online mapping platforms. These efforts, however, pay off in the long run, since these applications open a whole range of new research opportunities and novel ways to present and visualise research results.[46]

This panel presents and critically discusses the first results of the *Mapping Historical Leiden* project, which aims to develop a dynamic digital atlas of the pre-modern city of Leiden. The first phase of this project – a collaboration between historians, archaeologists and Leiden's heritage organisation (Erfgoed Leiden en Omstreken) – was recently completed (the first version of the atlas is accessible online at hlk.erfgoedleiden.nl, in Dutch). The mapping tool still requires further technical improvements to make it easier to upload and analyse additional data, and more geocoded datasets will become available in the coming months. The tool enables users to link, identify and search data across place and time, rather than providing static snapshots of the urban space in the past.

Apart from its technical resources and aspects, the mapping tool's research possibilities will be demonstrated by two case studies: one on the relation between space and wealth in sixteenth-century Leiden, and the other on the city's sanitary infrastructure in the early modern period. Together, these presentations will offer an opportunity to discuss the possibilities of digital mapping tools and the value of collaboration between scholars and specialists from the heritage sector in the

[45] See, for example, Anne Kelly Knowles and Amy Hillier, eds., *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (Redlands, Calif: ESRI Press, 2008); David J. Bodenhamer, John Corrigan, and Trevor M. Harris, eds., *The Spatial Humanities: GIS and the Future of Humanities Scholarship* (Bloomington: Indiana University Press, 2010); Alexander von Lünen and Charles Travis, eds., *History and GIS: Epistemologies, Considerations and Reflections* (Dordrecht: Springer, 2013); Ian N. Gregory and A. Geddes, eds., *Toward Spatial Humanities: Historical GIS and Spatial History* (Bloomington: Indiana University Press, 2014).

[46] See, for example, Onno Boonstra and Gerrit Bloothooft, eds., *Tijd en ruimte: nieuwe toepassingen van GIS in de alfawetenschappen* (Utrecht: Matrijs, 2009); a theme issue of *PCA. Post Classical Archaeologies* 2 (2012) on GIS for archaeologists and historians; Hélène Noizet, Boris Bove, and Laurent Jacques Costa, eds., *Paris de parcelles en pixels: analyse géomatique de l'espace parisien médiéval et moderne* (Saint-Denis: Presses Universitaires de Vincennes, 2013); Nicholas Terpstra and Colin Rose, eds., *Mapping Space, Sense, and Movement in Florence: Historical GIS and the Early Modern City* (New York: Routledge, 2016).

field of digital humanities, but also the practical and technical challenges of historical GIS and potential pitfalls of partnerships.

Presentation 1 (Ellen Gehring): ***One Size Fits All? Developing a Multi-Functional Digital Mapping Tool***

Building a cutting-edge map application for scholars, heritage managers and the general public is a major challenge in technical and methodological terms. *Mapping historical Leiden* has overcome some of the barriers, and this presentation focuses on the technical aspects of the mapping tool. Crucial for the project, for example, was the development of a so-called historical geocoder, which allows to link different geometric forms and to define their relations. Apart from technicalities, it will be further shown how very diverse data can be standardised through an advanced use of databases to ensure meaningful spatial analyses. The code of the mapping tool is available as open source, and since it is unnecessary for others to reinvent the wheel, it will be finally explained how the tool can be utilised in other contexts.

Presentation 2 (Arie van Steensel): ***Wealth and Place in Late Medieval Leiden: a Parcel-Based Analysis***

Leiden has a unique source, the so-called *Book of Waterways and Streets*, which contains about a hundred cadastral maps that were drawn for fiscal purposes in the second half of the sixteenth century. In this presentation, it will be first demonstrated how these maps were turned into a georeferenced base map. Secondly, it will be shown how this sixteenth-century pre-cadastral map can be used to analyse the relation between wealth and space in the city of Leiden at a parcel level, resulting in a more refined understanding of the complex relationship between occupation, wealth and place, which challenges common assumptions about the social geography of premodern cities and towns. The main point to be made is that historical GIS makes it possible to reinterpret sources that inform us about the importance of space and locality in structuring human interactions, as well as to present these data in an attractive and accessible way.

Presentation 3 (Roos van Oosten): ***Was sanitary infrastructure a privilege?***

Scholars have generally accepted that sanitary infrastructure was the privilege of the wealthy few. However, with the uncovering of hundreds of cesspits and water supply facilities in the town of Leiden in the past decades, this assumption can now be tested for different time periods. In order to investigate the question of accessibility to sanitary arrangements, the archaeologically documented sanitary structures must be plotted and financial valuation attached to them. Socio-economic data based on tax registers are available from about 1600, which will be most useful in this venture. Furthermore, thanks to HISGIS, we also have access to socio-economic data from 1832, which will allow us to establish a long-term perspective on the development of Leiden's sanitary infrastructure.

Session L

1. Was the *Ferguut* written by one or two authors?

Theo Meder, Gosse Bouma, Hannah Mars, Trudy Havinga (RUG)

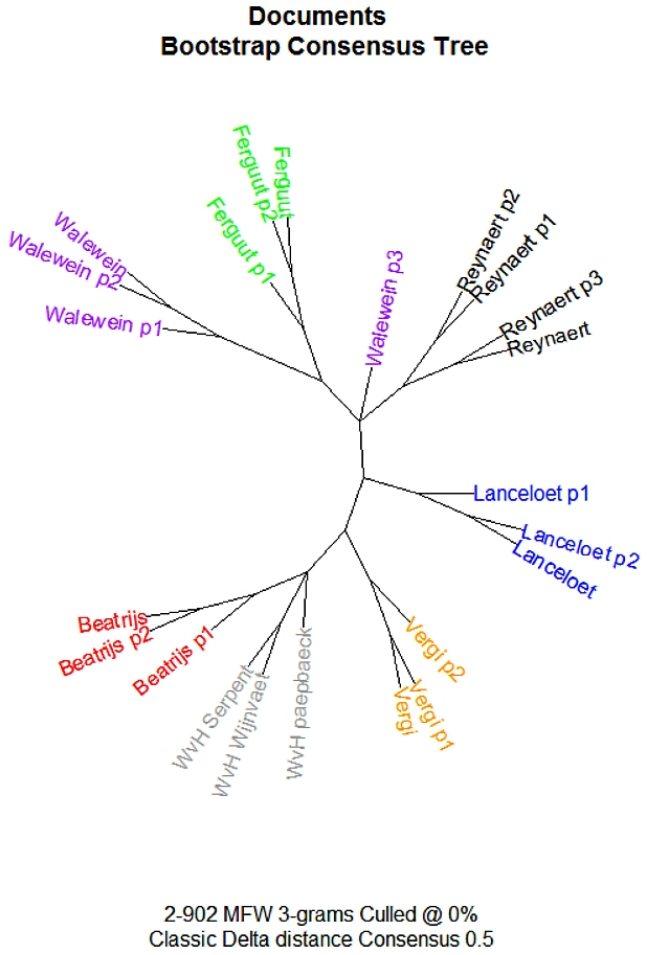
In 1989, Willem Kuiper published his thesis on the Middle Dutch romance *Ferguut* in which he concluded that the romance is written by two authors. Kuiper showed differences in writing style at all levels (rhyme, syntax, vocabulary, spelling) and concluded this was no coincidence. According to Kuiper, the first author translated the Old French *Fergus* by Guillaume le Clerc, approximately until vs. 2592, whereafter the second author completed the second half without French example - in the spirit of *Fergus*, but in his own words. Nowhere in the text there is a clear reference to a dual authorship (cf. the *Roman van Walewein*), but the style break halfway through the text was nevertheless something that a scholar like Eelco Verwijs noticed as well. Other researchers questioned or denied the finding that the *Ferguut* was written by two authors, like W.J.A. Jonckbloet, and after the appearance of Kuiper's thesis also Bart Besamusca and Mike Kestemont. With the thesis of Kestemont we entered the era of e-humanities. Whereas Kuiper had to do his quantitative style analysis by hand, today the programming language R in collaboration with the stylometric program Stylo can perform the job much faster, more thorough and completely unbiased (Stylo doesn't know or care what texts it gets presented and what the outcome may be, whereas human researchers may be influenced by preconceived ideas). In its analysis, the software not only takes all the differences into account (like Kuiper did), but all the similarities as well, even at levels where writers and readers are hardly aware of, such as word order and the use of function words. At this level every author leaves his most personal fingerprint behind.

Somewhat cautious Kestemont finally assumes that *Ferguut* was written by one author, who as a translator pulled open another register than as a free writer. Because *Ferguut* plays no prominent role in the investigation of Kestemont, we want to zoom in more focused on this particular romance. The central question: is the *Ferguut* written by one or two authors?

In order to investigate whether the two parts of the *Ferguut* are stylistically similar, we compare the similarity between the two parts of the *Ferguut* with the similarity between two or three parts of other 'randomly' selected Middle Dutch texts from around the same period and region, most of them dealing with courtly life. Seven texts we know to have been written by a single author, an eighth text we know that it is written by two authors. We involve the following texts in the analysis: *Ferguut*, *Beatrijs*, *De Borchgravinne van Vergi*, *Lanceloet en het hert met de witte voet*, *Van den vos Reynaerde* by Willem (the Aernout mentioned in the preface is the author of an Old-French *Renart* tranche), three poems (a deliberate misfit) by Willem van Hildegaersberch (*Vanden Serpent*, *Vanden Paep die sijn Baeck gestolen wert*, *Vanden Wijnvaet*) and *De Roman van Walewein* – for this experiment we looked at the complete texts, and cut up the longer texts into two or three even pieces in case there were no clear textual divisions. All the editions had to be thoroughly cleaned and converted to txt format.

Only *De Roman van Walewein* is most certainly written by two authors: to about two-thirds of the total number of verses, the story is written by Penninc (vs. 1 – 7.880), the last part is written by Pieter Vostaert (vs. 7.881 – 11.198). For the analysis we therefore cut this text into three pieces, so that the third part is written by Vostaert. As an experiment, we cut *Van den Vos Reynaerde* in three even pieces. The other longer texts we cut into two even pieces. *Ferguut* is cut at the location where the style transition should occur, so the place where the second author took over from the first, according to Kuiper. All these texts and fragments are then presented to Stylo for analysis. In this way, we can compare the similarity between the two parts of the *Ferguut* with the similarity between the two / three parts of a number of texts that we know are written by a single author, and the three parts of a text which we know that it was written by two authors. If the stylometric analysis

shows that the two parts of the *Ferguut* look as much alike as two parts of the texts of one author, and resemble each other more than the first two and the third part of the *Walewein*, that indicates that the *Ferguut* was also written by one author. If the analysis shows that the two parts of the *Ferguut* look less alike than the two parts of the texts of one author, and just as much, or less than the three parts of the *Walewein* together, this may indicate that the *Ferguut* is written by two authors.



In above graph, based on word tri-grams, Stylo shows what many already expected: all novels and writers are clustering neatly together (N.B.: the same happens with word bi-grams and with character bi-grams and tri-grams. As one can see in the graph, in hindsight the full texts need not have been included, but we wanted to be very sure we would not encounter any nasty surprises). The three parts of the *Reynaert* are stylistically most alike, the two parts of the *Beatrijs* mostly resemble each other, *Vergi* part 1 looks most like *Vergi* part 2 etc. Also the two parts of *Ferguut* stylistically match each other rather than any other text. Even the exemplum, the jest and the song of Hildegaersberh share the style of one and the same author. Only *Walewein* exhibits the expected deviation: Part 3 wanders off and positions itself somewhere between *Ferguut* and *Reynaert*, rather than next to the other parts of the *Walewein*. This graph of the stylometric analysis justifies no other conclusion than that the *Walewein* is written by two authors, but *Ferguut* by one author. Furthermore, it shows that the three Arthurian romances and *Reynaert* cluster together, and the courtly, religious and moralistic texts stand together separately. We experimented with all kinds of different parameters, but the results (practically) remained the same. Rolling delta resulted into nothing conclusive. Only cutting up the *Ferguut* in even smaller pieces and clustering them resulted in the style differences that Kuiper discovered, based on small pieces of comparison, but deprived of a long-term similarity overview over the text material.

Reservations can be made for the techniques used: stylometrics works better with longer texts than shorter ones, stylometrics works better on Standard Modern Dutch than on Middle Dutch texts with its unstable spelling, stylometrics works better on Middel Dutch rhyme pairs, all the editions should be either diplomatic or critical or in any other way normalized/standardized et cetera.

Still, all things considered, based on multiple stylometric examination, Stylo sees more similarities than differences between the two parts of the *Ferguut*, both on the level of word order and the use of function words – traits that are considered to be rather personal for each author. The *Ferguut* is most probably written by one author. In writing the second half of the text, the author may – also stylistically – be inspired by the fairy tale known as ATU 314A *The Shepherd and the Three Giants*, that was present in the Old-French *Fergus* as well. What we already knew about *Walewein* is confirmed: the last part of the romance shows more stylistic differences than similarities compared to other romances like the *Reynaert* and even *Ferguut*, and therefore *Walewein* was written by two authors. Finally, it is good to know now that one author could have several stylistic registers: one for when he translated, and one for when he freely retold a story.

References

B. Besamusca: 'De Vlaamse opdrachtgevers van Middelnederlandse literatuur: een literair-historisch probleem', in: *De nieuwe taalgids* 84 (1991), p. 150-162.

A.Th. Bouwman: *Reinaert en Renart. Het dierenepos Vanden vos Reynaerde vergeleken met de Oudfranse Roman de Renart*. 2 parts, Amsterdam 1991.

W. Bisschop & E . Verwijs (eds.): *Willem van Hildegaersberch: Gedichten.* 's-Gravenhage 1870.

K.H. van Dalen-Oskam: *De stijl van R.* Amsterdam 2013.

T. Dekker, J. van der Kooi & T. Meder: *Van Aladdin tot Zwaan kleef aan. Lexicon van sprookjes: ontstaan, ontwikkeling, variaties*. Nijmegen 1997.

M. Draak (ed.): *Lanceloet en het hert met de witte voet.* 6th imprint, Den Haag 1979.

M. Eder, J. Rybicki & M. Kestemont: *'*Stylometry with R: a package for computational analyses*',* in: *The R Journal* (2016), as download: https://journal.r-project.org/archive/accepted/eder-rybicki-kestemont.pdf

G.A. van Es (ed.): *De jeeste van Walewein en het schaakbord*. Zwolle 1957.

J.D. Janssens, R. van Daele & V. Uyttersprot (eds.): *Van den Vos Reynaerde. Het Comburgse handschrift.* 2nd imprint, Leuven 1998.

W.J.A. Jonckbloet (ed.): *Beatrijs. Eene sproke uit de XIII eeuw.* Den Haag 1841.

W.J.A. Jonckbloet: *Geschiedenis der Nederlandsche letterkunde*. 4th imprint, Groningen 1888, part 1.

M. Kestemont: *Het gewicht van de auteur. Stylometrische auteursherkenning in Middelnederlandse literatuur.* Gent 2013.

P. de Keyser (ed.): *De Borchgravinne van Vergi.* Antwerpen 1943.

W. Kuiper: *Die riddere metten witten scilde. Oorsprong, overlevering en auteurschap van de Middelnederlandse Ferguut, gevolgd door een diplomatische editie en een diplomatisch glossarium.* Amsterdam 1989.

E. Rombauts, N. de Paepe & M.J.M. de Haan (eds.): *Ferguut*. Den Haag 1982.

E. Stamatatos: 'A survey of modern authorship attribution methods', in: *Journal of the Association for Information Science and Technology* 60 (2008) 3, p. 538–556.

H.-J.Uther: *The Types of International Folktales. A Classification and Bibliography*. 3 volumes. Helsinki 2004.

2. Stylometry applied to book preferences

Peter Boot, peter.boot@huygens.knaw.nl

Introduction

One of the oldest and most active fields in Digital Humanities is authorship attribution. It has been shown many times that writers have a characteristic style that can be used to tell them apart (e.g. Burrows, 2002). It is also well known that word usage can be used to predict personality characteristics (e.g. Noecker, Ryan, & Juola, 2013). Personality characteristics in turn are related to preferences in different art forms (e.g. Cantador, Fernández-Tobías, Bellogín, Kosinski, & Stillwell, 2013). This suggests that, as one would hope, the stylistic differences whereby we tell authors apart (such as differences in function word usage) are not just meaningless preferences for one function word over another, but are related to artistic preference, in a way that is still to be clarified.

This paper, continuing earlier work (Boot, 2014), tries to contribute to that clarification, in that it will remove the middle term (the personality characteristics) and show that there is a direct relation between the words that people use and their preferences in art, in this case, for books. The writers that I study here are the writers of book reviews, not books. In the first section, I will use book reviews and ratings from book discussion sites and show correlations between word usage and book ratings. In the second section, I will take an exploratory approach and create a clustering of reviewers by word usage. For the two clusters, I will then look at their preferred word usage, as well as the word usage in the book descriptions of their preferred books.

Correlations between word usage and ratings

The data that the paper uses were collected from a number of Dutch book discussion sites. These sites include hebban.nl, lezerstippenlezers.be, bol.com and the now defunct sites watleesjij.nu and dizzie.nl.

The correlations were computed as follows: I selected reviews from users who had written at least 100000 characters, excluding some users with multiple accounts. I computed relative word frequencies in their reviews, and normalized the results (center around zero and divide by the standard deviation). In order to remove words with thematic links to books (murder, war, castle, love) I limited the computation to words defined as function words in the Dutch LIWC 2007 dictionary (Boot, Zijlstra, & Geenen, 2017, in press). For the same users I retrieved the book ratings and created a matrix of users by rating, excluding books that were rated only once. I computed the bias corrected distance correlation (a multivariate generalization of the correlation coefficient, see Székely & Rizzo, 2013) between the two matrices, and repeated that computation for reviews in all genres, in literature and in the literary thriller. The results are given in the first row of Table 1.

To be absolutely sure that no content-aspects of the reviews were reflected in the word usage, I repeated the computation using Part-of-speech-tags . The texts were tagged using Treetagger and instead of the relative word frequencies I used relative frequencies of POS bigrams. The results are given in the second row of the table.

Table 1

Correlations with p-values	All genres 189 reviewers 166 reviews (avg.)	Literature 41 reviewers 126 reviews (avg.)	Literary thriller 32 reviewers 88 reviews (avg.)
function words (200) vs. ratings	0.20 (0.000)	0.16 (0.000)	0.41 (0.000)
POS bigrams (100) vs. ratings	0.16 (0.000)	0.10 (0.002)	0.22 (0.000)

It is hard to interpret these correlation sizes, but it is clear that there are very significant correlations between function word usage and book ratings. The fact that these correlations persist even when looking at POS bigrams shows that the relation is to some extent based purely on linguistic style, not on content. Why sequences of POS-tags should be related to literary preference is an intriguing question that this paper will not solve.

Exploratory analysis

To get a feel for what this correlation might mean in terms of real reviews and ratings, I created a clustering based on function word usage for a group of reviewers. I removed a few outliers and was left with two clusters, cluster 1 containing 20 reviewers and cluster 2 containing 11.

I then looked at their reviews and preferred books. A sample of reviews from cluster 1 showed their informal, direct and very personal writing, characteristics that were much less prominent in cluster 2. This impression is confirmed when looking at contrastive keywords in the reviews of both clusters. The 20 key words with the largest effect size (Gabrielatos & Marchi, 2011) for both clusters are shown in table 2. It is clear cluster 1 prefers the first person, cluster 2 has more interest in writing.

Table 2

Cluster	Preferred review words
1	thought (was of the opinion), very, because, completely, me, actually, therefore, read (past part.) , beautiful, after all, had, have (1st pers. sing.), am, I, very, all, good, otherwise, yet, again
2	writer (fem.), writer, novel, reader, years, under, know, these, characters, one, between, gives, second, the, them, of, until, end, in, who

Turning to the ratings, while there were many books that were rated significantly higher by one of the groups, the preferences were hard to understand in terms of taste. Ratings summed by genre didn't show a very clear picture either. It was only when looking at contrastive word usage in the (publisher-provided) book descriptions for books read by either cluster that a clearer picture emerged.

Table 3

Cluster	Key words in preferred book descriptions
1	thriller, investigation, police, murdered, murder, case, body, someone, further, secret, above, know, very, sits, very, disappeared, within, nothing, appears, found, become, part, truth, books, there, something, else
2	in which, without, about, parents, family, city, big stories, last, exist, us, we, writer, history, love, country, tells, century, novel, Netherlands, war

Here it becomes clear that cluster 1 prefers thrillers and police novels, while cluster 2 has a less-focussed interest in family, writing and the country. It is worthwhile to repeat that these clusters of content words result from clustering reviewers on the basis of function words.

Conclusion

Taken together, the correlations and the exploratory analysis show that there is a relation between the function words that people use and their preferences for books. This relation still holds at the level of part-of-speech tags. This clearly shows that the word usage that helps tell authors apart is to some extent related to artistic preference. A possible explanation would be that the reviewers unconsciously imitate the books they read in their use of function words. That seems unlikely, among other reasons because the effect is also visible when we just look at the reviews in a single genre (second and third column of table 1). The more likely explanation is that function word usage is at least in part determined by artistic preference and related personality characteristics. The 'fingerprint' metaphor that is often used in this context, with its suggestion of an essentially random identifier, unlikely to be related to artistic preference, must therefore be considered as inappropriate.

Literature

Boot, P. (2014). *Dimensions of literary appreciation. Word use and ratings on a book discussion site*. Digital Humanities 2014. Retrieved from http://dharchive.org/paper/DH2014/Paper-825.xml

Boot, P., Zijlstra, H., & Geenen, R. (2017, in press). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6*(1).

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-287.

Cantador, I., Fernández-Tobías, I., Bellogín, A., Kosinski, M., & Stillwell, D. (2013). *Relating Personality Types with User Preferences in Multiple Entertainment Domains.* Proceedings of the 1st Workshop on Emotions and Personality in Personalized Services (EMPIRE 2013), at the 21st Conference on User Modeling, Adaptation and Personalization (UMAP 2013).

Gabrielatos, C., & Marchi, A. (2011). Keyness: Matching metrics to definitions. *Theoretical-methodological challenges in corpus approaches to discourse studies-and some ways of addressing them*.

Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing, 28*(3), 382-387.

Székely, G. J., & Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis, 117*, 193-213.

3. Corpus enrichment for 17th century Dutch: a pilot study

Feike Dietz[1], Marjo van Koppen[2], Irene Kramer[1] and Marijn Schraagen[2]
[1]Institute for Cultural Inquiry, [2]Utrecht Institute of Linguistics OTS
Utrecht University

1 Introduction

The Dutch language in the 17th century was a mixture of fading linguistic properties from the preceding language phase, Middle Dutch, and upcoming new ways to construct words and sentences. Within these language dynamics we observe a type of language variation that has rarely

been addressed before: variation within individual language users (intra-author variation). The aim of the current project is to describe and analyse in detail the linguistic and literary/rhetorical contexts in which intra-author variation occurs. As a prerequisite, the data needs to be annotated linguistically, using part of speech (POS) information and (morpho-) syntactic structure, and sociolinguistically, describing various factors that influence language use.

In a pilot project we restrict our research to the letters of the famous Dutch author and politician P.C. Hooft, written between 1600 and 1638. This collection is relatively large (approximately 800 letters, ~300.000 words) and contains sociolinguistic variation in type of correspondent and type of letter. The corpus can be used, i.a., to study the loss of negative concord in Dutch, which is observed in Hooft's letters from this period (Paardekooper, 2016).

As a starting point for obtaining POS tags, the Adelheid tagger for Middle Dutch (van Halteren and Rem, 2013) is used. Because the tagger is trained on Middle Dutch, the results are not highly accurate for 17th century texts. Therefore, a correction procedure for POS-tags and lemmas is performed by human annotators. Additionally, the annotators provide the necessary sociolinguistic information about letters and correspondents. When annotation is completed, a detailed and systematic analysis of linguistic phenomena will become feasible.

2 Approach

The source data is available in a diplomatic edition (Van Tricht, 1976). We use this edition after separating Hoofts original seventeenth century texts from the metadata (page numbers, foot notes, annotations).

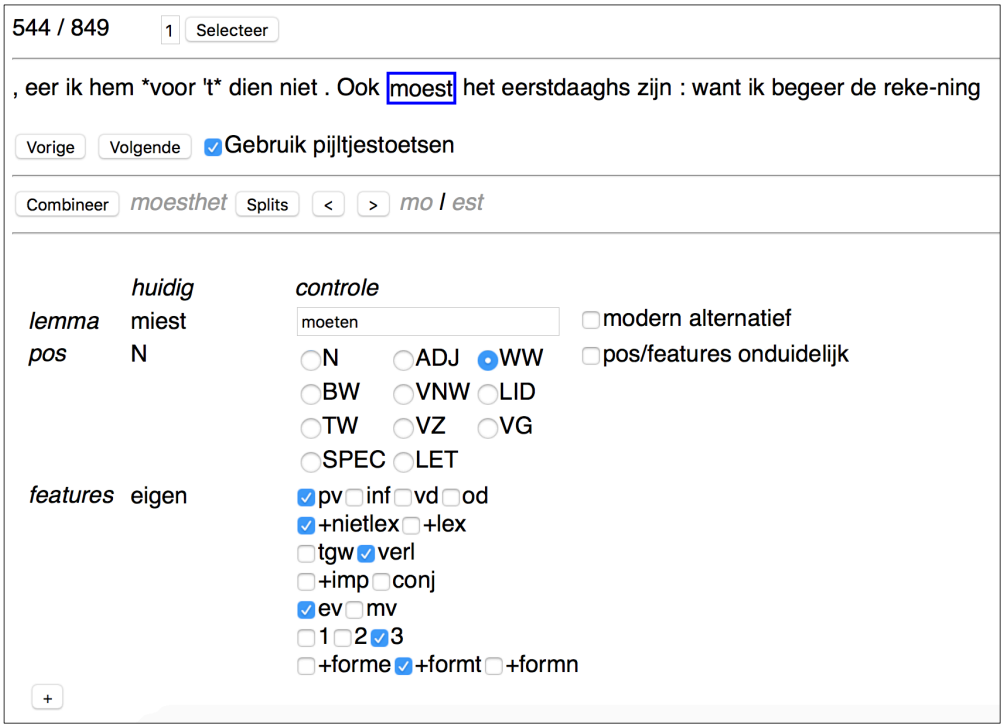


Figure 1: Example of the newly developed annotation tool

2.1 Part-of-Speech tagging

A collaboration with the Nederlab project (Brugman et al., 2016) is established to increase availability of the enriched corpus, by including the POS tagging and sociolinguistic metadata in the Nederlab research infrastructure. The integration necessitates conversion of the CRM tagset used by Adelheid to the CGN tagset used by Nederlab. Additionally, the tagging needs to be represented into the FoLiA

XML format for linguistic annotation (van Gompel and Reynaert, 2013). The CRM tagset is more extensive than CGN, notably in the use of surface form features such as form-e (words ending in -e). Surface form features are related to case marking, which is an important aspect in the study of linguistic variation in 17th century Dutch. Therefore, we decided to keep these features in the mapping to CGN tags (see Figure 1).

2.2 Sociolinguistic tagging

A key hypothesis in intra-author variation is the influence of sociological factors on linguistic choices. To evaluate this hypothesis systematically, all letters are being annotated with the following information:

- Goal: express thanks, ask advice, recommend, invite
- Topic: politics, religion, personal affairs, administration
- For individual correspondents:
  - name, gender, year of birth and death
  - status of correspondent as literary author
  - relation to Hooft: family members, literary friends, politicians, etc.
- For group correspondents:
  - name
  - domain: government, financial or legal institutions, civil associations
- Letter structure: greeting, introduction, narratio, closing formulas

2.3 Annotation process

A tool has been developed (see Figure 1) to perform POS and sociolinguistic annotation in an efficient way. A pool of annotators is available for the task, which will perform partly overlapping annotations to allow for agreement measurements. The annotation process is currently ongoing. A protocol has been developed to guide the post-correction process (see Figure 2 for examples).

Comparative and superlative adjectives are annotated individually. This rule is also applied for irregular adverbs, such as *veel, meer, meest* and *wel/goed, beter, best*. As an example, *minste* in the sentence below (1634, Van Tricht p. 527) receives a separate lemma `minst`:

. . . waer aen het ***minste*** *deel niet en zal hebben, Mê Joffr*^{e}*.*

Nominatives and non-nominatives are differentiated. We chose not to denominate dative, genitive, accusative and ablative. Instead, the surface form, related to case marking, is annotated. An example from 1633 (Van Tricht p. 437):

Veel $\boldsymbol{gelux}_{N(ev,non\text{-}nom,form\text{-}s)}$ *met . . .* $\boldsymbol{den}_{LID(bep,form\text{-}n)}$ $\boldsymbol{jongen}_{N(ev,non\text{-}nom,form\text{-}n)}$ *Arnout, dien god geeve 't lof* $\boldsymbol{des}_{LID(bep,form\text{-}s)}$ *geenen nae te ijvren, daer hij den naem af draeght.*

Figure 2: Annotation guideline examples

3 Analysis

In related work (Kramer, 2016) the use of negation by Hooft has been studied manually. Kramer shows that Hooft uses mostly single negation in different syntactical environments (subclauses, inversion, main clauses, local negation, V1 (verb-initial) sentences). Additionally, the negation particle *niet* can be used as alternative for the noun *nothing*. Furthermore, Hooft uses bipartite negation in almost all syntactical environments as well (all except in V1). In Kramer's research, not

one environment seemed to particularly ask for the use of bipartite negation. This research, however, encompassed only 107 letters. The fully annotated corpus will allow a more quantitative analysis, as well as a larger range and higher level of detail of linguistic phenomena.

Nobels and Rutten (2014) note the influence of gender and social class on negation (p. 41): 'while single negation spread from the north to the south, it also turned into a social variant, as the upper ranks in society and male letter writers seemed to be quicker to pick up on the incoming variant than the lower ranks and female letter writers'. Nobels and Rutten (2014) also note (p. 43) that traditions in letter writing affect linguistic development: 'fixed formulae were memorized as a whole (or copied) by writers from any social background. These fixed formulae occur in certain parts of the letters, mostly in the beginning and the ending'. With the current annotation effort, this type of observations can be studied systematically.

References

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., and van den Bosch, A. (2016). Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In Proceedings of LREC 2016.

van Gompel, M. and Reynaert, M. (2013). Folia: A practical xml format for linguistic annotation-a descriptive and comparative study. Computational Linguistics in the Netherlands Journal, 3:63–81.

van Halteren, H. and Rem, M. (2013). Dealing with orthographic variation in a tagger-lemmatizer for fourteenth century Dutch charters. Language Resources and Evaluation, 47(4):1233–1259.

Kramer, I. (2016). Variatie in negatie, een syntactisch en retorische analyse van het gebruik van enkele en tweeledige negatie in de brieven van P.C. Hooft van 1633 tot 1638 aan Joost Baek en Tesselschade Roemersdochter Visser. BA thesis, Universiteit Utrecht.

Nobels, J. and Rutten, G. (2014). Language norms and language use in seventeenth-century Dutch: negation and the genitive. In Rutten, G., editor, Norms and usage in language history, 1600-1900. A sociolinguistic and comparative perspective., pages 21–48. John Benjamins Publishing Company.

Paardekooper, P. (2016). Bloei en ondergang van onbeperkt ne/en, vooral dat bij niet-woorden. Neerlandistiek.nl.

van Tricht, H. (1976). De briefwisseling van Pieter Corneliszoon Hooft. Tjeenk Willink / Noorduijn.