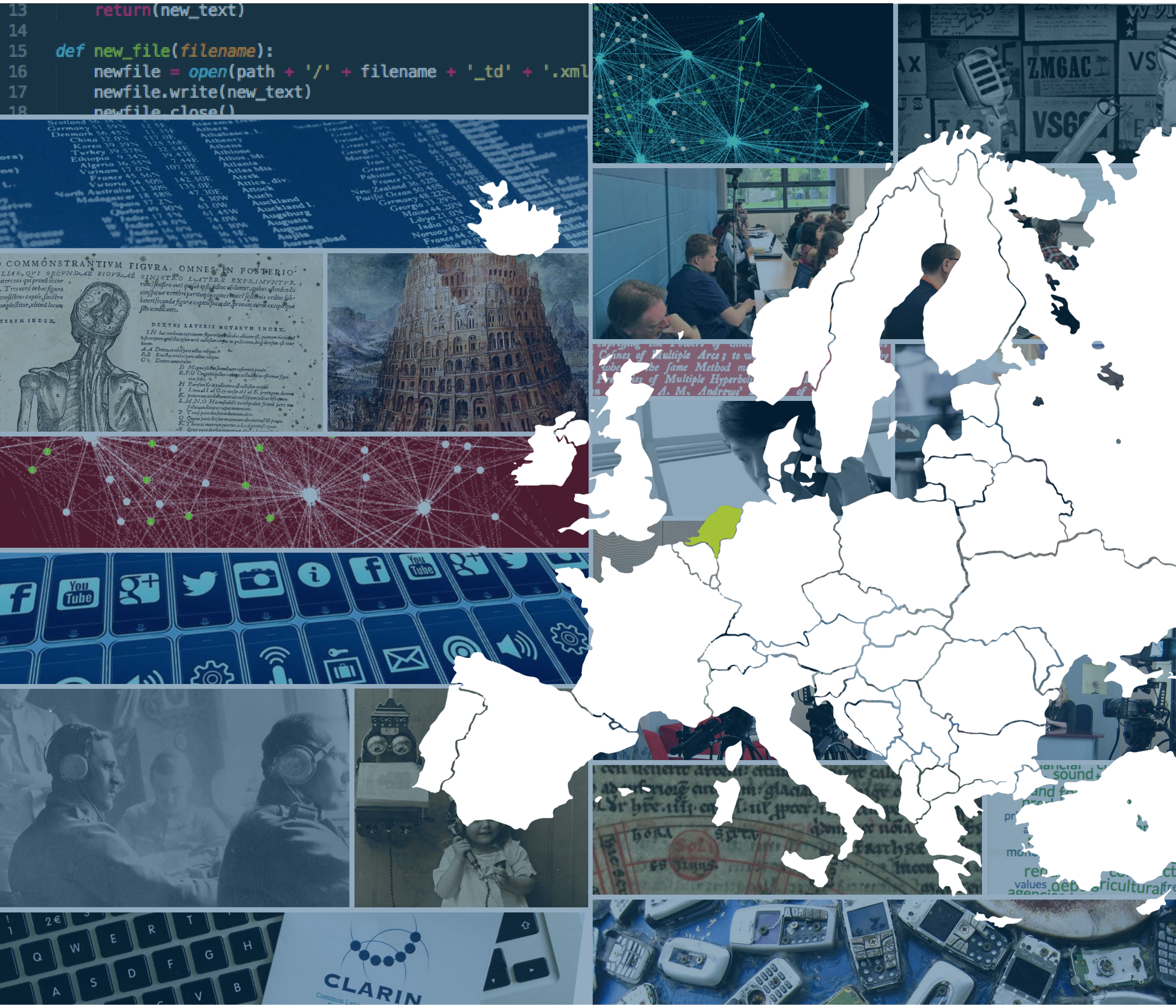
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

The Netherlands



```
13         return(new_text)
14
15     def new_file(filename):
16         newfile = open(path + '/' + filename + '_td' + '.xml
17         newfile.write(new_text)
18         newfile.close()
```

Written and edited by Darja Fišer and Jakob Lenardič

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN national consortia with the aim to increase the visibility of CLARIN consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

This brochure presents the Netherlands and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports on a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

Amsterdam, the Netherlands | photo by Milana Jovanov | Unsplash

The Netherlands

Written by Darja Fišer and Jakob Lenardič

CLARIAH-NL[1] is a project in the Netherlands that is setting up a distributed research infrastructure that provides humanities researchers with access to large collections of digital data and user-friendly processing tools. The Netherlands is a member of both CLARIN ERIC and DARIAH ERIC, so CLARIAH-NL contributes to both CLARIN and DARIAH. CLARIAH-NL not only covers humanities disciplines that work with natural language (the defining characteristics of CLARIN), but also disciplines that work with structured quantitative data. Though CLARIAH aims to cover the humanities as a whole in the long run, it currently focusses on three core disciplines: linguistics, social-economic history, and media studies.

CLARIAH-NL is a collaborative project that involves around 50 partners from universities, knowledge institutions, cultural heritage organisations and several SME companies. Currently, the data and applications of CLARIAH-NL are managed and sustained at eight centres in the Netherlands: Huygens Ing, the Meertens Institute, DANS, the International Institute for Social History, the Max Planck Institute for Psycholinguistics, the Netherlands Institute for Sound and Vision, the National Library of the Netherlands, and Dutch Language Institute. Huygens Ing, the Meertens Institute, the Max Planck Institute for Psycholinguistics, and Dutch Language Institute are Certified CLARIN Type B-centres. The consortium is led by an ten-member board, and its director and national coordinator for CLARIN ERIC is Jan Odijk.

The research, development and outreach activities at CLARIAH-NL are distributed among five work packages: Dissemination and Education and Technology deal with user involvement and the technical design and construction of the infrastructure, respectively, whereas the remaining three work packages focus on three selected research areas: Linguistics, Social and Economic History and Media Studies.

Dissemination and Education work package

In the user involvement-focused Dissemination and Education package, CLARIAH-NL aims to facilitate knowledge sharing among digital humanities and social sciences scholars as well as provide services that cater to the needs of their research. In this respect, CLARIAH-NL has successfully organised a variety of user involvement activities, such as the CLARIAH Linked Data Workshop (described on page 9), which took place in June 2017 and was intended to introduce Linked Data to both novice and advanced researchers.

Linguistics work package

MIMORE

In the Linguistics work package, CLARIAH-NL focusses on developing and improving applications for enriching corpora and searching through them – one such tool is MIMORE, which is described in greater detail on page 6. This enables researchers to investigate morphosyntactic variation in the Dutch dialects by searching three related databases with a common online search engine. The search results can be visualised on geographic maps and exported for statistical analysis. The three databases involved are DynaSAND, DiDDD and GTRP.

SoNaR

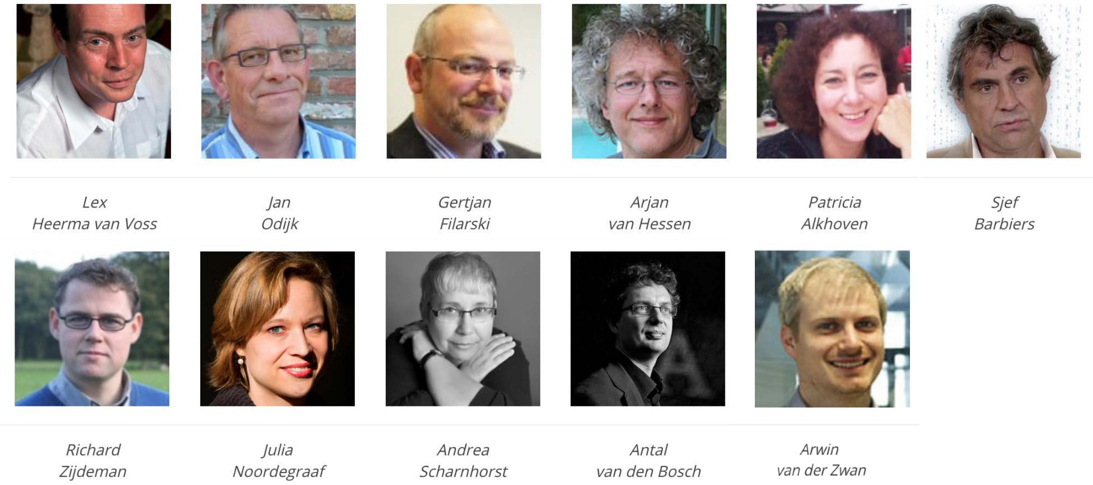
An important data set in this connection is the Dutch reference corpus SoNaR (described in greater detail on page 7), which was created in earlier projects for developing NLP software, but has been opened up for research by humanities scholars through the OpenSoNaR web application. State-of-the-art tools for the enrichment of textual corpora are also developed at the consortium. An example of such software is Frog, which is an NLP suite containing a tokeniser, PoS-tagger, lemmatiser, morphological analyser, and is thus an entity recogniser and dependency parser for Dutch.

Social and Economic History work package

In the Social and Economic History package, structured databases of social-economic history are being integrated into the Linked Data paradigm. The use of a uniform structure and explicit semantics ensures that relations and connections can be searched across different databases, which is of crucial importance for historical analysis and allows researchers to easily test hypotheses that could not be investigated before.

Media Studies work package

Finally, the Media Studies package focuses on providing special tools for viewing, browsing and searching through large collections of audio-visual data, such as films, radio broadcasts, and vlogs. It aims to provide a Media Suite, with access to relevant audio-visual collections by integrating tools developed in earlier projects, such as AVResearcherXL,[2] which is a tool for exploring radio and television programme descriptions, television subtitles and general newspaper articles through a user-friendly graphic interface.



Lex Heerma van Voss | *Jan Odijk* | *Gertjan Filarski* | *Arjan van Hessen* | *Patricia Alkhoven* | *Sjef Barbiers*

Richard Zijdeman | *Julia Noordegraaf* | *Andrea Scharnhorst* | *Antal van den Bosch* | *Arwin van der Zwan*

The board of CLARIAH-NL and National Coordinator Jan Odijk.

[1] http://www.clariah.nl/en/about/international

[2] http://www.clariah.nl/en/about/international

MIMORE

Written by Darja Fišer and Jakob Lenardič

MIMORE[3] is a tool developed at the Meertens Institute by means of which researchers can investigate morphosyntactic variation in Dutch dialects. Using this online environment, three different databases can be queried:

- the Dynamic Syntactic Atlas of the Dutch Dialects, which contains recordings and transcriptions of Dutch spoken in over 300 locations across the Netherlands, Belgium and a part of north-western France. It focuses on the syntactic variation among the dialects in these areas, such as differences in question formation and word order;
- the Diversity in Dutch DP Design, which contains oral and written interviews from about 200 locations in the Dutch language area and focuses on morphosyntactic variation in nominal structures; and
- the Goeman, Taeldeman, van Reenen Project, which focuses on morphological variation (such as the differences in verbal inflection) among roughly 600 locations in the Dutch language area.

MIMORE is very versatile in that it allows researchers to narrow down their search according to parameters relevant for the linguistic study of dialects, such as specific geographic locations or syntactic phenomena in which the dialects might differ from one another. The search engine also allows its users to export the data for statistical analysis, as well as to visualise them on a geographic map of the Dutch language area. It is accompanied by a comprehensive Educational Module, which provides an example of use on how to find sentences where the grammatical subject is realised twice within the same clause, which is a syntactic phenomenon that is limited to the south-western and western-central dialects of Dutch. There is also a case study available that demonstrates the combined use of MIMORE and the treebank search engine GrETEL.

The data that can be accessed through MIMORE point to the fact that many local Dutch dialects have begun to disappear or change in the direction of the standard language, due to increasing mobility and changes in the communication of its speakers. Additionally, the tool is also of great importance for formal linguistics, as it allows researchers to conduct micro-comparative studies of the Dutch language on the basis of data from dialects that show variation in terms of highly specific syntactic phenomena, such as the existence of complementiser agreement in certain Dutch dialects.

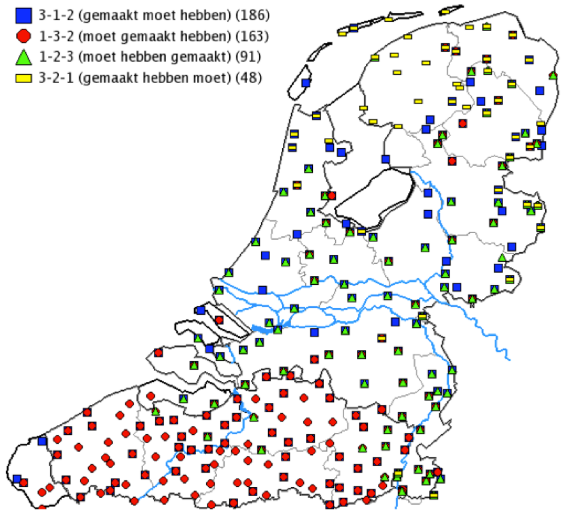

■ 3-1-2 (gemaakt moet hebben) (186)
● 1-3-2 (moet gemaakt hebben) (163)
▲ 1-2-3 (moet hebben gemaakt) (91)
▬ 3-2-1 (gemaakt hebben moet) (48)

Figure 1: The visualisation tool offered by MIMORE which shows how Dutch dialects differ from one another on the basis of different word orders in the verbal phrase.

[3] http://www.meertens.knaw.nl/mimore/

The SoNaR Reference Corpus of Dutch

Written by Darja Fišer and Jakob Lenardič

SoNaR[4] is a reference corpus of standard written Dutch. It comprises contemporary texts ranging from printed media such as books and periodicals to computer-mediated communication such as chats and tweets from the Netherlands and the Dutch-speaking area in Flanders (SoNaR New Media). It is the result of the STEVIN project, which involved major universities in the Netherlands and the Dutch-speaking part of Belgium, Flanders. The aim was to create a corpus of the contemporary written language, originally primarily intended for use by language and speech technology researchers and developers. It was made accessible and usable for humanities researchers in the CLARIN-NL and CLARIAH-NL projects by providing a web application with an interface for humanities researchers.

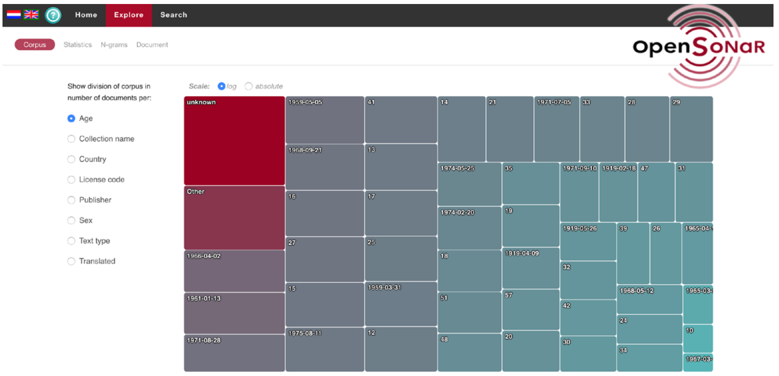
SoNaR consists of two main subcorpora – SoNaR-1 and SoNaR-500. In addition, there is the SoNaR New Media Corpus.

SoNaR-1 contains one million tokens and is very richly annotated, especially in relation to the semantic layers, which consist of named-entity labelling, annotation of co-reference relations, semantic role labelling and annotation of spatial and temporal relations. Additionally, all its annotations have been manually verified. As one of its pivotal subparts, SoNaR-1 includes the Dutch Parallel Corpus, a sentence-aligned parallel corpus of English, Dutch and French. The larger subcorpus, SoNaR-500, contains 500 million tokens of full texts. The texts in SoNaR-500 have been tokenised, tagged for part-of-speech and lemmatised, but without manual verification.

The SoNaR New Media corpus contains approximately 35 million words and consists of tweets, chats and SMS. All texts have been automatically tokenised, tagged for part of speech and lemmatised.

In order to provide easy access to the corpus, CLARIN-NL and CLARIAH-NL have developed the OpenSoNaR search environment. OpenSoNaR, with a frontend called WhiteLab and backend named BlackLab, is a state-of-the-art concordancer which provides two primary interfaces of user-driven functionality that can be used by both laymen and specialist researchers alike. In the Exploration interface (Figure 2), a researcher can investigate the corpus distribution, see the statistical information of the subcorpora and retrieve n-grams. Through the Search interface, four search options are available:

- simple, which limits the search to words only;
- extended, which enables the researcher to query the corpus by either word form or lemma, set the part of speech and choose among semantic metadata filters (Figure 3);
- advanced, which allows users to further specify the lemma or word forms that they're interested in; and
- expert, which provides an input for CQL commands.



Show division of corpus in number of documents per:

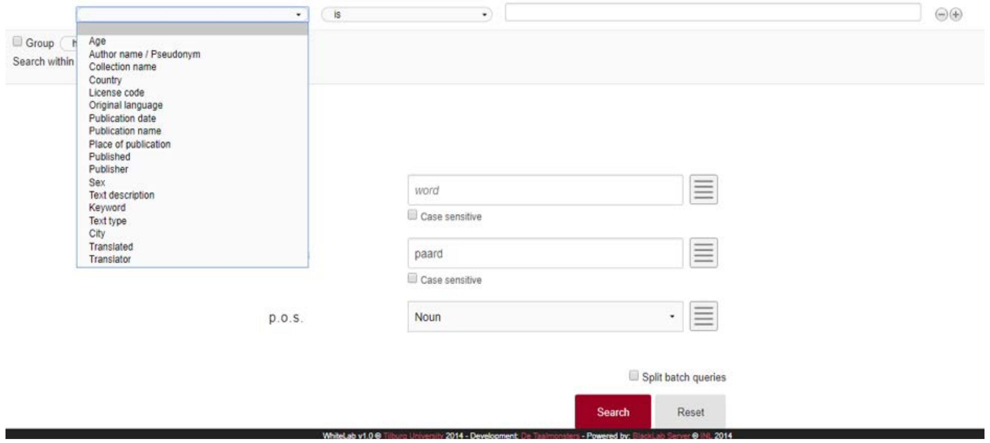
● Age
○ Collection name
○ Country
○ License code
○ Publisher
○ Sex
○ Text type
○ Translated

Figure 2: The Exploration interface showing the distribution of the subcorpora within SoNaR. The highlighted box shows that tweets make up 0.03% of the corpus.

[4] https://dev.clarin.nl/node/4195

The OpenSoNaR environment also stores previous search results, allowing researchers a great degree of flexibility and room for comparison between the temporary subcorpora that they have created during a single search session (Figure 4).

The successor of OpenSoNaR, called OpenSoNaR+, was developed in 2015.



□ Group

Search within

p.o.s. Noun

word

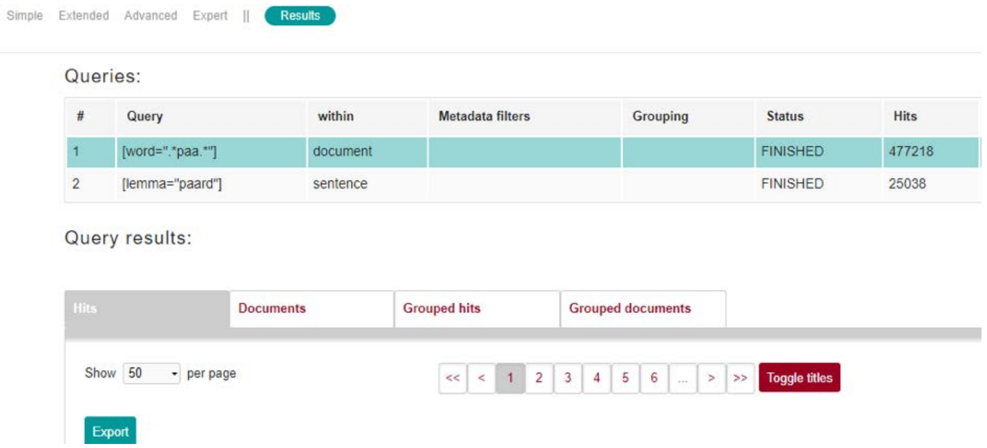
□ Case sensitive

paard

□ Case sensitive

□ Split batch queries

Figure 3: The ”extended” search interface — a search is being performed for the lemma "paard" (“horse”), while the drop menu shows metadata filters.



Queries:

#	Query	within	Metadata filters	Grouping	Status	Hits
1	[word=".*paa.*"]	document			FINISHED	477218
2	[lemma="paard"]	sentence			FINISHED	25038

Query results:

Show 50 per page

Figure 4: The “Results” tab shows that OpenSoNaR stores previous the results of queries.

CLARIAH-NL Workshops on Linked Data

Written by Darja Fišer and Jakob Lenardič

On 12 September 2016, CLARIAH-NL organised the first in a series of workshops on Linked Data,[5] which is a technological initiative that aims to ensure a greater degree of dynamic interoperability between language resources provided by the infrastructure. For instance, when Linked Data is applied to parliamentary corpora, as was the case in the CLARIN project Talk of Europe, the debates are enriched with extra layers of information in the sense that they are linked to their respective speakers, who are in turn annotated with biographic information, such as gender and political affiliation.

The workshop was attended by 40 researchers and primarily served two roles. On the one hand, it was an introduction to Linked Data and how it fares with respect to the CLARIAH-NL infrastructure, both in terms of successful applications and future challenges primarily related to bridging the gap between technological experts and humanities users who are not necessarily technologically savvy. On the other hand, the workshop directly involved arts and humanities researchers so that they could present their experiences of using Linked Data. For instance, Kaspar Beelen and Liliana Malgar from the University of Amsterdam gave a talk on the Digging into Parliamentary Data project, the aim of which was to enrich the parliamentary records of the Netherlands, United Kingdom and Canada with Linked Data so as to give researchers the opportunity to easily investigate the complex socio-historical aspects of politics.

CLARAH-NL Workshop on Linked Data, 12 September 2016. Image: CLARIAH-NL website

On 6 and 7 February 2017, CLARIAH-NL organised a follow-up international workshop that focused primarily on the application of Linked Data to linguistic research. It involved a number of international experts on Linked Data as well as prominent Dutch linguists who presented their current research topics in the fields of lexicology, phonetics and syntax in relation to using Linked Data. Among the speakers was Sjef Barbiers from the University of Leiden, who gave a talk on how Linked Data can benefit comparative syntax by ensuring interoperability between various databases of Dutch dialects that can be accessed through the tool MIMORE (read the presentation of the MIMORE on page 6). Jan Odijk together with Sjef Barbiers concluded the workshop by envisioning that the next step for CLARIAH-NL is to work on further stimulating the use of Linked Data in the fields of linguistic research where it is not yet widely applied, such as general lexica, and thereby reach out to a broader group of linguists.

[5] https://www.clariah.nl/en/new/blogs/clariah-linked-data-workshop

Melvin Wevers

Melvin Wevers is a digital humanities researcher focusing on the study of cultural-historical phenomena with the use of computational means. The following interview took place via Skype on 14 November 2018 and was conducted and transcribed by Jakob Lenardič and edited by Darja Fišer.

1. Can you please briefly describe your research background and tell us how you became a digital humanist who uses computational approaches to studying cultural phenomena?

I have a pretty diverse research background. I started out in the social sciences studying psychology and after I received my degree in 2006 I actually discovered that I didn't want to be a psychologist, but would rather do something that is much more based in the humanities, such as researching culture. Since I've always been interested in American culture, specifically, I applied for a Master's track in American Studies at Utrecht University. What I really liked about this field is its methodological variety in the sense that it combines elements from historical studies, media studies and literature. This kind of multidisciplinary approach made me become very interested in research and I decided to pursue a PhD after receiving my MA in 2009. However, I couldn't find a suitable PhD programme at first so I started studying cultural analysis at the University of Amsterdam, which was based more on quantitative methods, and at the same time began working at a software company that also used text mining, which sparked my interest for language technologies. Then I saw an advertisement for a PhD position at the University of Utrecht for using computational methods to research how American culture was represented in Dutch newspapers throughout the 20th century. I thought that this was a perfect opportunity for me, so I applied and began my career as a digital humanist!

2. How does your research benefit from the CLARIAH-NL infrastructure?

My PhD was funded by the Dutch Science Organisation, and the data that we used were provided by the National Library of the Netherlands. Though the project itself wasn't directly linked to CLARIAH-NL, I met a lot of people affiliated with CLARIAH-NL, like Arjan van Hessen and Franciska de Jong, at various conferences that I attended during the course of my studies. They pointed me to events organised by CLARIN consortia. These tutorials made it much easier for me to learn programming languages like R and Python, and I met a lot of my future colleagues with whom I could discuss my work in relation to source criticism or tool criticism. For instance, through the CLARIAH-NL consortium I learned that the German DARIAH was organising a tutorial on topic modelling. I learned a great deal about specific algorithms related to topic modelling that I would later use for my PhD. I think that CLARIAH-NL serves as an essential network that makes it significantly easier for researchers working in different fields to collaborate, especially since people like Arjan and Franciska put so much effort into helping novice researchers build the much needed connections.

3. In your PhD thesis, "Consuming America", you've applied a quantitative approach to a socio-historical study of how American consumer culture was depicted in the Netherlands throughout the 20th century. What inspired you to start researching this topic? Could you briefly describe your approach as well as the main findings of your research?

My PhD was part of a very large project funded by the Dutch government called "Translantis", which focused on determining how the United States was perceived in Dutch public discourse throughout the 20th century. My role was to focus on consumer goods such as Coca Cola and cigarettes in order to determine how American cultural values were portrayed in Dutch newspapers. This kind of research allowed me to gain a very multifaceted understanding of how Dutch people reacted to notions such as modernisation and globalisation through their perception of consumer goods. Since I've had a lifelong interest in all aspects of American culture—especially its international impact—I felt that this kind of research was a perfect opportunity for me.

In my approach, I combined the close reading of a more traditional historian with data-driven computational methodology. I first looked at a number of specific newspaper articles to get a very general feel of what the Dutch people thought about American culture at different times throughout the 20th century. Then I used quantitative methods like topic modelling on millions of newspaper articles to see whether such perceptions, as reported by these newspapers, constituted broader trends in Dutch history. In American Studies there is a deeply-entrenched idea that the 1950s and 1960s were a turning point during which American influences started becoming pervasive in the Netherlands—in other words, there is an idea of an American cultural invasion after the 1950s. However, by focusing on the depiction of consumer goods in newspapers I was able to show that the Dutch were already very much interested in and directly involved with American culture even before World War I. That is to say, the American influence in the Netherlands was relatively stable throughout the 20th century, so there was no specific mid-century turning point, as the Dutch had started to perceive themselves as modern consumers in the American sense from very early on.

In relation to a specific finding, Coca Cola was one of my case studies, and there I uncovered a very interesting dichotomy. In international advertisements, the Coca Cola company strove to advertise its product as global by omitting references to its American origin; however, in spite of this attempt, Dutch newspapers continued to overwhelmingly associate Coca Cola as something distinctively American. This in turn led me to uncover a major trend in Dutch public discourse, which is that the notion of globalisation became associated with Americanisation.

4. How does such a data-driven approach complement the traditional methods of a historian? Are there any specific advantages to such an approach?

After I finished my psychology degree, I lost interest in the quantitative methods of social sciences like statistics for a time, and instead wanted to solely focus on the traditional methods in the humanities, such as close reading and reading against the grain. However, I soon became critical of the lack of empirical evidence in the humanities, and I again became interested in data-driven methods. Ultimately what I learned is that these two approaches need to be combined, since this greatly increases the breadth of the research questions that a researcher is able to ask. That is, I think that computational methodologies can greatly assist a historian, especially since they make it much easier to adopt a bird's eye view of the periods that are being researched and thereby contextualise them properly as parts of the overarching historical trends.

5. Have historians working in your field generally embraced such quantitative methodology? Are there any changes that you would personally like to see take place within the field?

Unfortunately, in my field—that is, cultural history—using computational approaches is still a very new endeavour, so there are many researchers who outright refuse to use anything other than the traditional non-quantitative methods. This is understandable to an extent, since a senior researcher probably won't find the time to learn how to program late in his or her career. However, I feel that if you want to train a future generation of humanities scholars you should include courses on programming in the curriculum. Of course, this is far easier said than done, since I think this would require a kind of paradigm shift where entire syllabi would have to be revised in order to explicitly define, for instance, how a programming language like Python can be used to tackle research questions in fields where it is not immediately obvious how to apply quantitative methodologies. Because what often happens in practice is that a humanities department has a course on Python, but there are no other related courses that would help students apply their programming knowledge to research problems directly applicable to humanities questions.
In general, my opinion is that there should be a marriage between distant reading and close reading in the humanities, so I would like to see a greater degree of collaboration between scientists from different fields, such as between historians and computational linguists. I've written some papers with people who have a better understanding of mathematics than I do. If I had been left solely to my own devices, I would have had to spend a lot of time learning advanced mathematics, which would in turn probably make me neglect the humanities part of my research question. However, since I know some programming and some mathematics, it is easier for me to communicate with people that are experts in these fields. Such communication has already resulted in some very worthwhile interdisciplinary collaborations.

6. In your opinion, what could CLARIN do to become more widely used by historians? What activities, resources or tools would be needed to achieve this?

In the Netherlands, I think that CLARIN is still associated almost exclusively with computational linguistics even though CLARIAH-NL tries very hard to branch out into other humanities disciplines. So I think that they should continue to organise tutorials and especially focus on showcasing how the various datasets that are already out there are relevant for various disciplines. For instance, there is a plethora of historical sources that have been digitised, but many historians aren't aware of the various exciting ways in which they could direct their research on the basis of the wide availability of these datasets.

7. You've been involved in the development of ShiCo, a tool for the analysis of how words denoting a certain concept change diachronically. Could you briefly describe how this project came about? What are the main advantages of ShiCo?

I have been interested in figuring out how words denoting a certain concept change over time, but found that approaches such as topic modelling were too rigid to do this efficiently. I approached another PhD student, Tom Kenter, who specialises in Natural Language Processing and information retrieval with this problem, and he came up with the idea to use a relatively novel technique to chart how these changes happen. We involved some other researchers working in the history department at the University of Amsterdam so that we could test whether the results of our first prototype were in accordance with their expert knowledge of the domains. Since the prototype was successful, we were encouraged by some of the professors to apply for a grant and turn the prototype into an interactive tool. By working with programmers from the eScience Centre, we eventually managed to turn the tool into ShiCo.[6]

8. Can you highlight any other project that you're currently working on?

After finishing my PhD, I became a post-doc researcher at the National Library of the Netherlands, where I applied techniques from the field of computer vision to gain insights into non-textual trends in the Dutch advertisements landscape. The research produced a dataset of advertisements as well as a tool called SIAMESE to find visually similar images in a large corpus of advertisements. Currently I'm working on applying computational methods to analyse Dutch academic historical journals and thereby determine the trends related to the understanding of history on the part of Dutch historians – for instance, how notions like progress and modernity are discussed and which countries were in focus over different periods of time.

9. What is your vision for the future of CLARIAH-NL and digital humanities in the Netherlands?

I believe that computational methodology should become part and parcel of all kinds of disciplines and that digital humanities should, at a certain point, lose the modifier digital and become the standard way of doing humanities research. I think that CLARIAH-NL can play an important role in bridging the gap between these different fields, especially by offering interactive tutorials and ensuring interoperability between repositories and tools. Like I said previously, one of the problems is that researchers do not know what to do with the available data so CLARIAH-NL could offer these much-needed guidelines and training, as well as educate researchers on concepts like open science so as to ensure that their work is as transparent as possible.

Amsterdam, the Netherlands | photo by Sabina Fratila | Unsplash

[6] https://github.com/NLeSC/ShiCo

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume I (publication number: CLARIN-CE-2018-1341, November 2018).

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

Written by
Darja Fišer and Jakob Lenardič

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2018-1341
November 2018

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands
www.clarin.eu

CLARIN
Common Language Resources and
Technology Infrastructure

CLARIN 2018